Extensions of the Dynamic Programming Framework

by

Morgan Jones

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2021 by the
Graduate Supervisory Committee:

Matthew Peet, Chair
Angelia Nedich
Matthias Kawski
Marc Mignolet
Spring Berman

ARIZONA STATE UNIVERSITY

December 2021

ABSTRACT

Modern life is full of challenging optimization problems that we unknowingly attempt to solve. For instance, a common dilemma often encountered is the decision of picking a parking spot while trying to minimize both the distance to the goal destination and time spent searching for parking; one strategy is to drive as close as possible to the goal destination but risk a penalty cost if no parking spaces can be found. Optimization problems of this class all have underlying time-varying processes that can be altered by a decision/input to minimize some cost. Such optimization problems are commonly solved by a class of methods called Dynamic Programming (DP) that breaks down a complex optimization problem into a simpler family of sub-problems. In the 1950s Richard Bellman introduced a class of DP methods that broke down Multi-Stage Optimization Problems (MSOP) into a nested sequence of "tail problems". Bellman showed that for any MSOP with a cost function that satisfies a condition called additive separability, the solution to the tail problem of the MSOP initialized at time-stage $k > 0$ can be used to solve the tail problem initialized at time-stage $k - 1$. Therefore, by recursively solving each tail problem of the MSOP, a solution to the original MSOP can be found. This dissertation extends Bellman‘s theory to a broader class of MSOPs involving non-additively separable costs by introducing a new state augmentation solution method and generalizing the Bellman Equation. This dissertation also considers the analogous continuous-time counterpart to discrete-time MSOPs, called Optimal Control Problems (OCPs). OCPs can be solved by solving a nonlinear Partial Differential Equation (PDE) called the Hamilton-Jacobi-Bellman (HJB) PDE. Unfortunately, it is rarely possible to obtain an analytical solution to the HJB PDE. This dissertation proposes a method for approximately solving the HJB PDE based on Sum-Of-Squares (SOS) programming. This SOS algorithm can be used to synthesize controllers, hence solving the OCP, and also compute outer bounds of reachable sets of dynamical systems. This methodology is then extended to infinite time horizons, by proposing SOS algorithms that yield Lyapunov functions that can approximate regions of attraction and attractor sets of nonlinear dynamical systems arbitrarily well.

ACKNOWLEDGMENTS

Throughout the writing of this dissertation, I have received a great deal of support and assistance.

I would first like to thank my advisor, Professor Matthew Peet, for the countless meetings, discussions, and encouragements. Your knowledge and guidance were invaluable in writing this dissertation and throughout my Ph.D. You have been an excellent mentor to me; I hope to continue applying the lessons I've learned going forward.

I would also like to thank my committee members Professors Angelia Nedich, Matthias Kawski, Marc Mignolet, and Spring Berman for providing constructive feedback and support along my Ph.D. journey.

I am very grateful to my fellow and former lab colleagues Hesameddin Mohammadi, Brendon Colbert, Sachin Shivakumar, Amritam Das, Reza Kamyar, Declan Jagt, and Aleksandr Talitsky for sharing codes and providing helpful feedback on my research during our many group meetings. In particular, I would like to thank Hesam for the many insightful tips he gave me on how to use Latex and Matlab, as well as for helping me to learn how to drive.

I would like to thank Professors Pedro Peres and Ricardo Oliveira for hosting me at the University of Campinas. It was during this trip that I met my collaborators Lucas Fernandes and Professor Daniel Dotta whom I hope to continue working with in the future. In addition, I would like to thank the many friends I made along my Ph.D. journey. My former roommates Sandesh, Kishen, Kevin, Jag, Kennedy, Zach, and Will made life entertaining. We shared many interesting conversations.

Lastly, I would like to give thanks to my parents and my sister, who were only a phone call away, and my wife, Adriana, and her family for their wise counsel and sympathetic ear. You gave me the motivation to complete my Ph.D.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

> [Gottfried Leibniz] conceives God in the creation of the world like a mathematician who is solving a minimum problem.
>
> Emil du Bois-Reymond

In 2012, it was reported in Solar Energy Power Association (2013), that 95,000 new distributed solar PhotoVoltaic (PV) systems were installed nationally across the USA, a 36% increase from 2011 and yielding a total of approximately 300,000 installations. Further, utility-scale PV generating capacity has increased at an even faster rate, it was reported in Sherwood (2013), that 2012 installations more than doubled that of 2011. Meanwhile, it has been reported in Conti (2014), that partially due to the development of energy-efficient appliances and new insulation materials, US electricity demand has plateaued. As a consequence of these trends, utility companies are faced with the problem that demand *peaks* continue to grow while total US electricity demand remains stagnant. Specifically, as per the US EIA Shear (2014), the ratio of peak demand to average demand has increased dramatically over the last 20 years. The situation in Nevada, California, and Arizona is particularly challenging. Hydro electrical power generators can come online in a matter of minutes to meet sudden unpredictable demand surges; see Kirmani *et al.* (2021). However, due to record low levels in the Lake Mead reservoir, the yearly power supplied to Nevada, California, and Arizona by the Hoover dam has consistently reduced over the years. As reported in Penmetsa (2020), the available energy from Hoover Dam has decreased 2.4% from 2004-2016 with an expected decrease of 3% from 2017-2050.

Fundamentally, the problem faced by utilities is that consumers are typically charged based on total electricity consumption, while utility costs are based on both electricity consumption and maintaining the generating capacity necessary to meet peak demand. Recently, several public and private utilities have moved to address this imbalance by charging residential consumers based on both the total electricity consumed - a cost referred to as Time-of-Use (TOU) charge, and the maximum *rate* ($ per kW) of consumption - a cost referred to as a *demand charge.* Specifically, in Arizona, both major utilities SRP and APS have mandatory demand charges for residential consumers, see SRP (2015).

For consumers, residential electrical power requirements are relatively inflexible, hence the most direct approach to minimizing the effect of demand charges is the use of battery storage devices such as the Tesla Powerwall considered in Farhangi (2010); Mohd *et al.* (2008); Dunn *et al.* (2011). These devices allow consumers to shift electricity consumption away from periods of peak demand, thereby minimizing the effect of demand charges. This naturally leads to a control problem of finding the optimal way to charge and discharge a residential battery in order to minimize the residential electricity cost (involving both the demand charge and TOU charge).

An Optimization Problem for Optimal Battery Scheduling to Minimize Electricity Costs Let us mathematically formulate the problem of designing optimal charge/discharge residential battery schedules to minimize household consumer electricity costs. We consider a billing period given by $\{0, ..., T\}$, where the times denoting the beginning of on-peak and off peak billing hours, t_{on} and t_{off} respectively, are such that $0 \leq t_{\text{on}} < t_{\text{off}} \leq T$. The TOU charge then is given by,

$$J_{\text{TOU}}(\mathbf{u}) = p_{\text{off}} \sum_{k=0}^{t_{\text{on}}-1} (q(k) + u(k)) + p_{\text{on}} \sum_{k=t_{\text{on}}}^{t_{\text{off}}-1} (q(k) + u(k)) + p_{\text{off}} \sum_{k=t_{\text{off}}}^{T} (q(k) + u(k)),$$

where $p_{\text{on}} \in \mathbb{R}$ and $p_{\text{off}} \in \mathbb{R}$ are cost conversion constants for on-peak and off peak hours respectively, $\mathbf{u} = (u(0), ..., u(T-1))$, $u(k)$ is the amount of electricity used by the battery to charge (if $u(k) > 0$) or discharge (if $u(k) < 0$) at time stage k, and $q(k) \in \mathbb{R}$ is the amount of electrical energy used by household appliances at time stage k. Note that the total electrical energy required by the household at time stage k is equal to the appliance energy plus the electrical energy required by the battery, $q(k) + u(k)$.

The demand charge is given by,

$$J_D(\mathbf{u}) = p_d \max_{k \in \{t_{\text{on}}, ..., t_{\text{off}}-1\}} \{q(k) + u(k)\},$$

where $p_d \in \mathbb{R}$ is a electrical energy request to cost conversion constant.

We model the energy stored in the battery by a Markov time process:

$$e(k+1) = \alpha(e(k) + \eta u(k)), \tag{1.1}$$

where $e(k)$ denotes the energy stored in the battery at time step k, $\alpha > 0$ is the bleed rate of the battery and $\eta \in [0, 1]$ is the efficiency of the battery.

Assuming the electrical power required by the residential unit at each time stage, $q(k)$ for $k \in \{0, ..., T\}$, is fixed and known (due to consumer inflexibility), we can formulate the battery scheduling problem as an optimization problem of the following form,

$$\begin{aligned}
&\min_{\mathbf{u}} \{J_{\text{TOU}}(\mathbf{u}) + J_D(\mathbf{u})\} \text{ subject to} \qquad (1.2)\\
&e(k+1) = \alpha(e(k) + \eta u(k)) \text{ for } k = 0, ..., T-1\\
&e(0) = 0\ , e(k) \in [\underline{e}, \bar{e}],\ u(k) \in [\underline{u}, \bar{u}] \text{ for } k = 0, ..., T,\\
&\mathbf{u} = (u(0), ..., u(T-1)),
\end{aligned}$$

where $\underline{u}$ and $\bar{u}$ are bounds on the maximum and minimum rates we can charge/discharge, and $\underline{e}$ and $\bar{e}$ are lower and upper storage capacity bounds on the battery.

Optimization Problem (1.2) is challenging to solve because it has an objective function involving a point-wise maximum term (due to the demand charge) that cannot be separated into costs that only depend on state and inputs at each time stage. The goal of this dissertation is to develop new techniques and methods to solve such challenging optimization problems.

Multi-Stage Optimization Problems and Optimal Control Problems Problems such as the Battery Scheduling Problem (1.2), as well as many other optimization problems commonly encountered throughout Engineering, Economics, and Mathematics, all feature a constraint representing an underlying time-varying process that is controlled over a finite number of time-stages with the goal of minimizing some cost. In this dissertation we develop new methods and techniques to solve and analyze such optimization problems. Specifically, we consider optimization problems in either of the following forms, initialized at some (x_0, t_0):

$$\begin{aligned}
&(\mathbf{u}^*, \mathbf{x}^*) \in \arg\inf_{\mathbf{u}} \left\{ J_{t_0}(\mathbf{u}, \mathbf{x}) \right\} \\
&\text{subject to:} \qquad\qquad\qquad\qquad\qquad\qquad (1.3) \\
&\mathbf{u} = (u(t_0), ..., u(T-1)), \mathbf{x} = (x(t_0), ..., x(T)) \\
&x(t_0) = x_0,\ x(t+1) = f(x(t), u(t)) \\
&\text{for all } t = t_0, .., T-1, \\
&x(t) \in \Omega \subset \mathbb{R}^n,\ u(t) \in U \subset \mathbb{R}^m \\
&\text{for all } t = t_0, .., T.
\end{aligned}$$

$$\begin{aligned}
&(\mathbf{u}^*, x^*) \in \arg\inf_{\mathbf{u}} \Bigg\{ \\
&\qquad \int_{t_0}^{T} c(x(t), \mathbf{u}(t), t) dt + g(x(T)) \Bigg\} \\
&\text{subject to:} \qquad\qquad\qquad\qquad\qquad\qquad (1.4) \\
&x(t_0) = x_0,\ \dot{x}(t) = f(x(t), \mathbf{u}(t)) \\
&\text{for all } t \in [t_0, T], \\
&x(t) \in \Omega \subset \mathbb{R}^n,\ \mathbf{u}(t) \in U \subset \mathbb{R}^m \\
&\text{for all } t \in [t_0, T].
\end{aligned}$$

Throughout this dissertation, we will refer to optimization problems of the Form (1.3) as Multi-Stage Optimization Problems (MSOPs) and optimization problems of the

Form (1.4) as Optimal Control Problems (OCPs). Both MSOPs and OCPs are solved by finding a policy/controller, denoted by $\mathbf{u}$, that minimizes the objective/cost function. This dissertation is split into two parts considering the discrete-time case of MSOPs and the continuous-time case of OCPs separately.

We focus on a particular class of methods used to solve MSOPs and OCPs called Dynamic Programming (DP). DP is a general class of numerical algorithms used to solve optimization problems with the following properties: 1) They break down the optimization problem into a family of simpler sub-problems that are sequentially solved. 2) The solution to each sub-problem is stored in computer memory. 3) Previously stored solutions are used to solve the next sequential sub-problem. 4) A solution to the optimization problem is constructed after solving all members of the family of sub-problems.

An alternative to DP is brute force methods, such methods solve optimization problems by computing and storing the value of the objective function evaluated at every feasible solution, for which there could be an uncountable number, and then outputting the feasible solution that produces the smallest object value when the algorithm terminates. Intuitively, DP methods solve optimization problems more efficiently than brute force methods because DP methods harness computer memory in a way that avoids repeating computations. Rather than repeating the computation, DP methods recall the outcome of such calculations from stored memory.

Aside from computational efficiency concerns, MSOPs (1.3) and OCPs (1.4) are often solved using DP methods due to the fact that they can easily be broken down into a family of simpler sub-problems, their “tail problems”. Specifically, both MSOPs and OCPs have underlying time varying processes, $x(t+1) = f(x(t), u(t))$ and $\dot{x}(t) = f(x(t), u(t))$ respectively. This naturally leads to a family of sub-problems consisting of the “tail problems” indexed by $t \in \{t_0, ..., T\}$ for MSOPs and $t \in [t_0, T]$ for

OCPs. The optimal value of the objective function for each "tail problem" can then be computed and stored in a function called the Value Function (VF).

1.0.1 Part 1: Discrete Time

Part 1 is concerned with MSOPs of Form (1.3). Classically, MSOPs have cost functions of the form $J_{t_0}(\mathbf{u},\mathbf{x}) = \sum_{t=t_0}^{T-1} c_t\left(x\left(t\right), u\left(t\right)\right) + c_T\left(x\left(T\right)\right)$, we call such functions additively separable functions. For MSOPs with additively separable cost functions Richard Bellman derived necessary and sufficient conditions, encapsulated in Bellman's Equation (BE), for an input and state sequence to be optimal, see Bellman (1966). Specifically, Bellman showed that in the additively separable case if we can find a function F that satisfies the BE:

$$\begin{aligned} &F(x,T) = c_T(x) \text{ for all } x \in \Omega \\ &F(x,t) = \inf_{u \in \Gamma_x} \left\{ c_t(x,u) + F(f(x,u),t+1) \right\} \text{ for all } x \in \Omega, t \in \{t_0,..,T-1\}, \end{aligned} \tag{1.5}$$

where $\Gamma_x := \{u \in U : f(x,u) \in \Omega\}$, then a necessary and sufficient condition for a feasible input and state sequence, $\mathbf{u} = (u(t_0),...,u(T-1))$ and $\mathbf{x} = (x(t_0),...,x(T))$, to solve the MSOP (1.3) initialized at x_0 is,

$$\begin{aligned} &u(t) \in \arg \inf_{u \in \Gamma_{x(t)}} \left\{ c_t(x(t),u) + F(f(x(t),u),t+1) \right\} \text{ for all } t \in \{t_0,..,T-1\}, \\ &x(t_0) = x_0 \text{ and } x(t+1) = f(x(t),u(t)) \text{ for all } t \in \{t_0,..,T\}. \end{aligned}$$

Since Bellman first introduced the BE, special cases of MSOPs with additively separable cost functions have been extensively studied. For instance, the Riccatti Equations, important in Linear Quadratic Regulator (LQR) control, can be derived from the BE in the special case when the cost function is quadratic of form $c_t\left(x\left(t\right), u\left(t\right)\right) = x\left(t\right)^T Q x\left(t\right) + u\left(t\right)^T R u\left(t\right)$, where $Q, R > 0$. In more recent years heuristic algorithms have been developed to approximately solve MSOPs with additively separable cost functions, known as reinforcement learning, see Bertsekas (2019). However, little attention has been paid to more general cost functions that do not exhibit additively separable structure. In Chapters 3 and 4 we extend the class of known DP methods to solve MSOPs with non-additively separable cost functions.

A Generalization of Bellman's Equation Let us consider an MSOP (1.3) that is solved by finding the input sequence, $\mathbf{u}$, that drives some time varying process, governed by some map f, to some goal set, $S \subset \mathbb{R}^n$. It follows that the cost function for this MSOP is $J(\mathbf{u},\mathbf{x}) = \min\{T, \{k \in \{0,...,T\} : x(k) \in S\}\}$. It is not immediately clear if it is possible to write J in the additively separable form, $J(\mathbf{u},\mathbf{x}) = \sum_{t=0}^{T-1} c_t\left(x\left(t\right), u\left(t\right)\right) + c_T\left(x\left(T\right)\right)$. Therefore, for this fundamental path planning problem we are unable to immediately solve the MSOP by applying classical DP (ie solving Bellman's Equation (1.5)). In Chapter 3 we propose an extension on the classical DP theory to solve MSOPs whose cost function is not necessarily additively separable (like the cost function associated with path planning).

Specifically, we generalize the class of additively separable cost functions to a class of functions we call Monotonically Backward Separable Functions (MBSF), functions that can be written as a nested composition of maps backwards in time, taking the form:

$$J_{t_0}(\mathbf{u},\mathbf{x}) = \phi_{t_0}(x(t_0), u(t_0), \phi_{t_0+1}(x(t_0+1), u(t_0+1), \ldots \phi_T(x(T)) \ldots)).$$

Analogous to Bellman we derive necessary and sufficient conditions for optimality for problems with cost functions of this form. In order to do this, we provide a generalization of Bellman‘s Equation.

For MSOPs with monotonically backward separable cost functions we show in

Chapter 3 that if we can find a function V that satisfies,

$$\begin{aligned} V(x,T) &= \phi_T(x) \text{ for all } x \in \Omega \\ V(x,t) &= \inf_{u \in \Gamma_{x,t}} \left\{ \phi_t(x,u,V(f(x,u,t),t+1)) \right\} \text{ for all } x \in \Omega, t \in \{t_0,..,T-1\}, \end{aligned} \tag{1.6}$$

where $\Gamma_{x,t} := \{u \in U : f(x,u,t) \in \Omega\}$, then a necessary and sufficient for a feasible input and state sequence, $\mathbf{u} = (u(t_0),...,u(T-1))$ and $\mathbf{x} = (x(t_0),...,x(T))$, to solve the MSOP (1.3) initialized at x_0 is

$$\begin{aligned} & u(t) \in \arg \inf_{u \in \Gamma_{x(t),t}} \left\{ \phi_t \left(x(t), u, V(f(x(t),u,t),t+1) \right) \right\} \text{ for all } t \in \{t_0,..,T-1\} \\ & x(0) = x_0 \text{ and } x(t+1) = f(x(t),u(t)) \text{ for all } t \in \{t_0,..,T\}. \end{aligned}$$

Equation (1.6) can be thought of as a generalization of Bellman's Equation (1.5); since we show that in the case when the cost function is additively separable (a special case of monotonically backward separable functions) Equation (1.6) reduces to Bellman's Equation (1.5).

Formalizing The Principle Of Optimality Given an MSOP, for DP to work efficiently, stored solutions of each of the "tail" problems of the MSOP, should be used to help solve the next sequential sub-problem. In Chapter 3 we show that DP can be used to efficiently solve MSOPs with backward separable cost functions by breaking the problem down into tractable sub-problems based on the "tails" of the MSOP. This method is equivalent to solving the GBE (1.6) backwards in time starting from the terminal time $T \in \mathbb{N}$. However, given an MSOP with an objective function not known to be a backward separable function, how do we know if it is possible to efficiently solve the MSOP by solving its "tail" problems backwards in time? Richard Bellman answered this question by showing MSOPs that satisfy the Principle of Optimality can be efficiently solved using DP methods. In Bellman's own words, an MSOP satisfies the Principle of Optimality if "An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision", see Bellman (1966).

In Chapter 3 we propose a mathematical formulation of Bellman's principle of optimality. We show that any MSOP with a monotonically backward separable cost function must satisfy our definition of the principle of optimality. Therefore, our definition of the principle of optimality provides a necessary condition to be able to solve an MSOP using the GBE (1.6). Furthermore, we conjecture that a necessary and sufficient condition for a function J to be a monotonically backward separable function is that every MSOP (with any vector field f, state and input constraints Ω and U) with cost function J satisfies our definition of the principle of optimality.

Using our proposed definition of the principle of optimality we are able to show not all functions are monotonically backward separable. For instance, we show $J(\mathbf{x},\mathbf{u}) = \max\{\max_{0 \le k \le T-1}\{d_k(u(k),x(k))\}, d_T(x(T))\} + \sum_{s=0}^{T-1} c_s(x(s),u(s)) + c_T(x(T))$ is not monotonically backward separable. Hence, we are unable to solve MSOPs with cost functions with the form of J using the GBE (1.6).

State Augmentation Methods for Dynamic Programming In Chapter 3 we show that

$$J(\mathbf{x},\mathbf{u}) = \max\{\max_{0 \le k \le T-1}\{d_k(u(k),x(k))\}, d_T(x(T))\} + \sum_{s=0}^{T-1} c_s(x(s),u(s)) + c_T(x(T))$$

is not monotonically backward separable. Hence, we cannot use the GBE (1.6) to solve MSOPs with J as a cost function. Unfortunately, the battery scheduling problem, given in Eq. (1.2), has a cost function that is of this form. In Chapter 4 we introduce a new technique, based on state augmentation, that allows us to solve MSOPs with cost functions that may not be monotonically backward separable. In order to do this we consider a class of cost functions we call forward separable functions, functions that can be written as a nested composition of maps forwards in time taking the form:

$$J_{t_0}(\mathbf{u},\mathbf{x}) = \psi_T\left(x\left(T\right), \psi_{T-1}\left(x\left(T-1\right), u\left(T-1\right), \ldots, \psi_{t_0}\left(x\left(t_0\right), u\left(t_0\right)\right) \ldots ..\right)\right).$$

We show in general MSOPs with forward separable cost functions do not satisfy the principle of optimality, and therefore it is not possible to derive necessary and sufficient optimality conditions analogous to the BE without modification or reformulation. To solve such MSOPs, with forward separable cost functions, we show that extending the state space by introducing a new state variable, $z(t+1) = \psi_t(x(t), u(t), z(t))$, an equivalent MSOP with additively separable cost function can be constructed. The equivalent MSOP with additively separable cost function can then be solved using classical DP methods.

Solving The Battery Scheduling Problem for Minimizing Consumer Electricity Costs The battery scheduling problem given in Eq. (1.2) has a cost function that is non-additively separable and also non-monotonically backward separable. However, we show in Chapter 4, that this cost function is forward separable. We then use our proposed state augmentation method combined with classical DP to solve Opt. (1.2). Our derived controller successfully minimizes residential electricity costs by charging the battery during off-peak hours and discharging the battery during peak hours in order to minimize both TOU and demand charges.

1.0.2 Part 2: Continuous Time

Engineered systems are becoming increasingly common and autonomous. Such systems may be designed to operate for months or years without direct human intervention. Thus, in modern societies, there is an increasing need for safety diagnosis tools that can determine whether or not an autonomous system will enter an unsafe region, where the probability of failure is known to be high. In Part 2 of this dissertation, we develop several safety diagnosis tools that can determine the long-term properties of continuous-time systems described by Ordinary Differential Equations (ODEs). Specifically, we develop methods to bound reachable sets, regions of attraction, and attractor sets. We also derive performance bounds for controllers constructed from approximated value functions. In order to do this, we consider OCPs of Form (1.4).

An SOS Algorithm for Approximately Solving the HJB Equation Analogous to the Bellman Equation (1.5) associated with MSOPs (considered in Part 1), necessary and sufficient conditions for optimality of OCPs can be expressed as an equation, called the Hamilton Jacobi Bellman (HJB) Partial Differential Equation (PDE). Specifically, it can be shown that if a function, $V : \mathbb{R}^n \times [0,T] \to \mathbb{R}$, satisfies the following equation, known as the HJB PDE,

$$\begin{aligned} &\nabla_t V(x,t) + \inf_{u \in U} \left\{ c(x,u,t) + \nabla_x V(x,t)^T f(x,u) \right\} = 0 \text{ for all } (x,t) \in \mathbb{R}^n \times (0,T), \\ &V(x,T) = g(x) \quad \text{for all } x \in \mathbb{R}^n, \end{aligned} \tag{1.7}$$

then a solution to the OCP (1.4) can be constructed as follows

$$\begin{aligned} &\mathbf{u}^*(t) = k(x^*(t),t), \text{where } \dot{x}^*(t) = f(x^*(t), k(x^*(t),t)), \\ &\text{and } k(x,t) \in \arg\inf_{u \in U} \left\{ c(x,u,t) + \nabla_x V(x,t)^T f(x,u) \right\}. \end{aligned} \tag{1.8}$$

For a given OCP of Form (1.4) a function, V, that satisfies Eq. (1.7) is referred to as a Value Function (VF) and determines the optimal objective to the OCP (1.4) initialized at $(x_0,t_0) \in \mathbb{R}^n \times [0,T]$.

Eq. (1.8) shows that VFs can be used to solve OCPs. In Chapter 5 we consider the problem of computing an approximate VF. In order to do this, we consider the problem of finding a solution to the HJB PDE (1.7) from a computational perspective as a feasibility (optimization) problem. As the HJB PDE is nonlinear in V, the set of functions that satisfy the HJB PDE is non-convex. Therefore the feasibility problem of finding a solution to the HJB PDE is computationally intractable. To overcome this intractability, we relax the feasibility problem to the problem of finding a function that satisfies linear differential inequalities. We then tighten this problem to a sequence of SOS programming problems, indexed by their degree $d \in \mathbb{N}$, each of which yield a polynomial, P_d. We show the sequence of polynomials $\{P_d\}_{d \in \mathbb{N}}$ converges to the solution to the HJB PDE as $d \to \infty$ with respect to the L^1 norm.

Controller Construction From Approximate Value Functions With Performance Bounds For a given OCP, if V in Eq. (1.8) is not the true VF of the OCP then the resulting controller may not be optimal. In Chapter 5 we proposed a computational method for approximating VFs, however, it is unknown how well a controller constructed from such an approximated VF will perform. In Chapter 6 we consider the problem of bounding the distance from optimality of a controller constructed from some approximate VF (derived from our method proposed in Chapter 5 or any other method). We show that the sub-optimality in performance of a controller constructed using a candidate VF is bounded by the "closeness" of the candidate VF and true VF with respect to some norm. Specifically, we define the sub-optimality in performance of an input, $\mathbf{u}$, as the difference between the objective function in the OCP (1.4) evaluated using $\mathbf{u}$ and the infimum of the objective function. We then show that the sub-optimality in performance of a controller constructed according to Eq. (1.8) using some approximated VF, P, is bounded by $C||V - P||_{W^{1,\infty}}$, where $C > 0$ and V is a VF associated with the OCP (1.4).

Analyzing the Long Term Properties of Autonomous Systems Unlike in Chapters 5 and 6, in Chapters 7 and 8 we assume that the input, $\mathbf{u}$, is fixed and given; hence we assume $\mathbf{u}$ is absorbed into the vector field, f. Therefore, in Chapters 7 and 8 we consider systems defined by nonlinear autonomous ODEs of the form,

$$\dot{x}(t) = f(x(t)) \qquad x(0) = x_0. \tag{1.9}$$

We denote the solution map of the ODE (1.9) by $\phi_f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ which satisfies

$$\frac{d}{dt}\phi_f(x,t) = f(\phi_f(x,t)) \text{ for all } x \in \mathbb{R}^n \text{ and } t \geq 0,$$
$$\phi_f(x,0) = x \text{ for all } x \in \mathbb{R}^n.$$

In both Chapters 7 and 8 we analyze the long term properties of the solution map to the ODE (1.9) as $t \to \infty$.

Specifically, in Chapter 7 will propose an SOS optimization problem that yields an inner approximation of the maximal Region of Attraction (ROA) of the ODE given in Eq. (1.9). For a given equilibrium point, a ROA of a nonlinear Ordinary Differential Equation (ODE) is defined as a set of initial conditions for which the solution map of the ODE tends to that equilibrium point with respect to the euclidean norm. The maximal ROA of an equilibrium point, meanwhile, is defined as the ROA which contains all other ROAs of that equilibrium point. Without loss of generality in Chapter 7 we will assume the equilibrium point of the ODE is at the origin, that is $f(0) = 0$ (note that a linear change of variables allows for any equilibrium point to be transformed to the origin). The maximal ROA is then defined as

$$ROA_f := \{x \in \mathbb{R}^n : \lim_{t \to \infty} ||\phi_f(x,t)||_2 = 0\}.$$

The problem of computing sets which accurately approximate the ROA with respect to some set metric plays a central role in the stability analysis of many engineering applications. An inner approximation of the ROA (a set that is certifiably contained inside of the ROA) provides a set of initial conditions for which solutions to the ODE converge towards some stable equilibrium point and hence can be used to rule out non-steady unsafe solution trajectories. For instance, knowledge of the ROA provides a metric for the susceptibility of the F/A-18 Hornet aircraft experiencing an unsafe out-of-control flight departure phenomenon, called falling leaf mode Chakraborty *et al.* (2011a,b).

In Chapter 8 we will propose an SOS optimization problem that yields an approximation of the minimal attractor set of a given ODE. An attractor set of an ODE (1.9) is a set with the following properties: 1) It is compact and nonempty. 2) It is invariant, that is, any solution map initialized in the set will remain in the set for all time. 3) It is locally attracting, that is, it possesses a neighborhood of initial conditions for which the solution map converges towards the set.

Attractor sets can be thought of as generalized equilibrium points and hence provide a generalized notion of stability of nonlinear ODEs. Thus knowledge of the attractor set of an ODE can be used to certify whether a solution of an ODE will remain in some compact set or become unbounded at $t \to \infty$. Aside from providing a generalized notion of stability, attractor sets are used in secure private communications Cuomo *et al.* (1993); Zhao *et al.* (2018), the computation of Unstable Periodic Orbits (UPOs) Lakshmi *et al.* (2020), and risk quantification of financial systems Gao *et al.* (2018).

Converse Lyapunov Functions for Region of Attraction Approximation Unfortunately, there is no general analytical expression for the ROA of the nonlinear ODE. Rather than trying to solve the ODE (for which no general method exists) and then construct ROA from the solution of the ODE, arguably the most widely used technique for computing ROAs has been to use Lyapunov‘s second method.

Lyapunov‘s second method involves searching for a “generalized energy function”, called a Lyapunov function. A Lyapunov function of an ODE is any function that is positive everywhere, apart from the origin where it is zero, and is strictly decreasing along the solution map of the ODE. Specifically, if we can find a function V such that $V(0) = 0$ and $V(x) > 0$ for all $x \neq 0$, then if $\nabla V(x)^T f(x)$ is negative over the sublevel set $\{x \in \mathbb{R}^n : V(x) \leq a\}$ we have that $\{x \in \mathbb{R}^n : V(x) \leq a\} \subseteq ROA_f$ is a ROA.

The feasibility and accuracy of using Lyapunov‘s second method can be deduced using converse Lyapunov theory. Given an ODE with a stable equilibrium point, converse Lyapunov theory seeks to answer the following questions: Under what conditions does there exist a Lyapunov function to certify the stability of the ODE and furthermore, under what conditions does this Lyapunov function yield the ROA of the ODE? In this dissertation, we are particularly interested in showing the existence of polynomial converse Lyapunov functions that enable us to design SOS algorithms for ROA approximation.

In Vannelli and Vidyasagar (1985) a converse Lyapunov function, called the maximal Lyapunov function, was proposed, taking the form,

$$V(x) = \int_0^\infty \alpha\left(||\phi_f(x,t)||_2\right) dt, \tag{1.10}$$

where $\alpha : \mathbb{R} \to [0, \infty)$. It was shown that for any given asymptotically stable ODE there exists a maximal Lyapunov function whose ∞-sublevel set is equal to the region of attraction of the ODE. However, since by definition any maximal Lyapunov function is unbounded outside of the region of attraction it cannot be approximated arbitrarily well (with respect to any norm) by a polynomial over any compact set that contains points outside of the region of attraction (since polynomials are bounded over compact sets). Thus, it is not possible to design an SOS-based algorithm that can approximate maximal Lyapunov functions arbitrarily well.

In Chapter 7 we propose a new converse Lyapunov function of the form,

$$W_{\lambda,\beta}(x) := \begin{cases} 1 - \exp(-\lambda \int_0^\infty ||\phi_f(x,t)||_2^{2\beta} dt) \text{ when } x \in ROA_f \\ 1 \text{ otherwise,} \end{cases} \tag{1.11}$$

where $\lambda > 0$ and $\beta \in \mathbb{N}$. We show that for any given locally exponentially stable ODE the 1-sublevel set of the function $W_{\lambda,\beta}$ is equal to the ROA of the ODE. Moreover, for sufficiently large $\lambda > 0$ and $\beta \in \mathbb{N}$ we show that $W_{\lambda,\beta}$ is globally Lipschitz continuous. Furthermore, we show that for sufficiently large $\lambda > 0$ and $\beta \in \mathbb{N}$ $W_{\lambda,\beta}$ satisfies the following PDE,

$$\nabla W_{\lambda,\beta}(x)^T f(x) = -\lambda ||x||_2^{2\beta}(1 - W_{\lambda,\beta}(x)) \text{ for almost every } x \in \mathbb{R}^n. \tag{1.12}$$

The problem of computing the ROA of a given ODE is now reduced to the problem of solving the PDE (1.12).

Previously, in Chapter 5 we consider the problem of finding a solution to the HJB PDE (1.7) from a computational perspective as a feasibility (optimization) problem. Similarly, in Chapter 7 we view the problem of approximating the ROA from a computational perspective as a feasibility (optimization) problem that is solved by solving the PDE (1.12). We approximately solve the PDE (1.12) by relaxing the PDE to a Partial Differential Inequality (PDI) while minimizing the L^1 norm between solutions to the PDI and $W_{\lambda,\beta}$. We then tighten this optimization problem to a family of d-degree SOS optimization problems. We show that the sequence of solutions to each of our d-degree SOS optimization problems yields a sequence of 1-sublevel sets that tends to the ROA as $d \to \infty$ in the volume metric.

Of course, there is no coincidence between the similarity of the methods proposed in Chapter 5 and Chapter 7. Converse Lyapunov functions can be thought of as value functions associated with infinite time horizon OCPs with positive cost functions that are zero at the origin. Therefore, each converse Lyapunov function should satisfy an HJB-"type" PDE which can be solved using a similar method to the SOS-based method proposed in Chapter 5.

A Lyapunov Type Condition For Minimal Attractor Set Approximation In Lin *et al.* (1996) it was shown that $A \subset \mathbb{R}^n$ is an attractor set of an ODE defined by f if and only if there exists $V \in C^\infty(\mathbb{R}^n, [0,\infty))$ such that

$$\kappa_1 \left(\inf_{y \in A} ||x-y||_2 \right) \le V(x) \le \kappa_2 \left(\inf_{y \in A} ||x-y||_2 \right) \text{ for all } x \in \mathbb{R}^n, \tag{1.13}$$

$$\nabla V(x)^T f(x) \le -\kappa_3 \left(\inf_{y \in A} ||x-y||_2 \right) \text{ for all } x \in \mathbb{R}^n/A, \tag{1.14}$$

where κ_1 and κ_2 are class K_∞ functions (where a function k is of class K_∞ if $k : \mathbb{R} \to \mathbb{R}$ is monotonically increasing, $k(0) = 0$ and $\lim_{r \to \infty} k(r) = \infty$) and $\kappa_3 : \mathbb{R} \to (0,\infty)$ is a continuous positive definite function.

We can show $A \subset \mathbb{R}^n$ is an attractor set of an ODE by solving an optimization (feasibility) problem by finding V satisfying the inequality constraints given in Eqs. (1.13) and (1.14). However, given an ODE, in general the attractor sets of the ODE are unknown. In the case when A and V are unknown, the optimization problem with inequality constraints given in Eqs. (1.13) and (1.14) is nonlinear (in terms of the decision variable $A \subset \mathbb{R}^n$). To overcome this challenge, in Chapter 8, we propose a new Lyapunov type condition that only involves a single decision variable V and which can yield an attractor set of a given ODE. Specifically, we show that if $V \in C^1(\mathbb{R}^n, \mathbb{R})$ satisfies

$$\nabla V(x)^T f(x) \le -(V(x) - 1) \text{ for all } x \in \Omega, \tag{1.15}$$

$$\{x \in \Omega : V(x) \le 1\} \subseteq \Omega^\circ, \tag{1.16}$$

$$\{x \in \Omega : V(x) \le 1\} \neq \emptyset, \tag{1.17}$$

where $\Omega \subset \mathbb{R}^n$ is a compact set, then $\{x \in \Omega : V(x) \leq 1\}$ is an attractor set of the ODE defined by f.

Unlike the conditions in Eqs. (1.13) and (1.14) the conditions in Eqs. (1.15), (1.16), and (1.17) only involve one unknown variable V. In Chapter 8 we propose a sequence of d-degree SOS optimization problems, each being solved by an SOS polynomial, V, with minimal 1-sublevel set volume while satisfying Eqs. (1.15), (1.16), and (1.17). We show, given an ODE defined by f, as $d \to \infty$ the sequence of 1-sublevel sets constructed from the solutions to our SOS optimization problem converge to the minimal attractor set of the ODE in the volume metric.

Unfortunately, given an SOS polynomial $V(x) = z_d(x)^T P z_d(x)$ where $P > 0$, there is no known convex analytic expression for the volume of the 1-sublevel set of an SOS polynomial, $\{x \in \mathbb{R}^n : V(x) \leq 1\}$. Therefore, for implementation, we propose an alternative sequence of SOS optimization problems with the objective of maximizing $\log\det(P)$, a function known to be convex. Heuristically, maximizing the determinate of a positive matrix maximizes the eigenvalues of the matrix. Thus, $V(x) = z_d(x)^T P z_d(x)$ is maximized for each $x \in \mathbb{R}^n$, implying that the volume of 1-sublevel set of V is minimized. We show in several numerical examples, including the Lorenz attractor and Van-der-Poll oscillator, that this heuristic implementation is able to produce tight approximations of minimal attractor sets associated with nonlinear ODEs.

Summary of Contribution In this dissertation, we consider both discrete-time and continuous-time problems. In discrete-time we extend classical DP methods to handle MSOPs with non-additively separable cost functions, namely, monotonically backward separable functions and forward separable functions. Our proposed methods are able to solve MSOPs of practical importance, such as those associated with battery scheduling or path planning.

In continuous-time we propose an SOS-based algorithm that approximately solves the HJB PDE, yielding a VF that can be used to construct a solution to the OCP. For approximate VFs we derive performance bounds, showing that controllers constructed from approximated VFs are bounded in sub-optimality by the Sobolev distance between the approximated VF and the true VF. Applying the intuition gained by solving HJB PDEs associated with finite-time horizon OCPs we then propose an SOS-based algorithm for approximating ROAs of nonlinear ODEs. In order for us to design such an SOS algorithm, we propose a new converse Lyapunov function that can be considered as a special case of a VF associated with an infinite-time horizon OCP. Finally, we propose a new Lyapunov condition that we show is well suited to the problem of computing optimal outer approximations of minimal attractors.

We note that the work presented in this dissertation is based on the following papers:

- Chapter 3 is based on"A Generalization of Bellman's Equation with Application to Path Planning, Obstacle Avoidance and Invariant Set Estimation," Jones and Peet (2021d).

- Chapter 4 is based on "Extensions of the Dynamic Programming Framework: Battery Scheduling, Demand Charges, and Renewable Integration," Jones and Peet (2021c).

- Chapters 5 and 6 are based on "Polynomial Approximation of Value Functions and Nonlinear Controller Design with Performance Bounds," Jones and Peet (2021e).

- Chapter 7 is based on "Converse Lyapunov Functions and Converging Inner

Approximations to Maximal Regions of Attraction of Nonlinear Systems," Jones and Peet (2021a).

- Chapter 8 is based on "A Converse Sum-of-Squares Lyapunov Function for Outer Approximation of Minimal Attractor Sets of Nonlinear Systems," Jones and Peet (2021b).

Chapter 2

NOTATION

> We could, of course, use any notation we want; do not laugh at notations; invent them, they are powerful. In fact,mathematics is, to a large extent, invention of better notations.
>
> Richard Feynman

MSOP and OCP Notation: $\mathcal{M}^{Discrete}_{Addative}$ is the class of additively separable MSOPs (Defn. 3.4). $\mathcal{M}^{Discrete}_{Backward}$ is the class of naturally backward separable MSOPs (Defn. 3.4). $\mathcal{M}^{Discrete}_{Finite}$ is the class of naturally backward separable with finite cardinality state and input spaces (Defn. 3.7). $\mathcal{M}^{Discrete}_{Path}$ is the class of MSOPs associated with the path planning problem appearing in Section 3.5.1. $\mathcal{M}^{Discrete}_{Forward}$ is the class of naturally forward separable MSOPs (Defn. 4.3). $\mathcal{M}^{Continuous}_{Lip}$ is the class of Lipschitz continuous OCPs (Defn. 5.2). $\mathcal{M}^{Continuous}_{Poly}$ is the class of polynomial OCPs (Defn. 5.10).

Set Notation: We denote the power set of $\mathbb{R}^n$, the set of all subsets of $\mathbb{R}^n$, as $P(\mathbb{R}^n) = \{X : X \subset \mathbb{R}^n\}$. For two sets $A, B \in \mathbb{R}^n$ we denote $A/B = \{x \in A : x \notin B\}$. For $x \in \mathbb{R}^n$ and $p \in \mathbb{N}$ we denote $||x||_p = (\sum_{i=1}^n x_i^p)^{\frac{1}{p}}$. We denote the minimal distance between a point, $x \in \mathbb{R}^n$, and a set, $A \subset \mathbb{R}^n$, by $D(x, A) := \inf_{y \in A}\{||x - y||_2\}$. For $\eta > 0$ and a point $y \in \mathbb{R}^n$ we denote the set $B_\eta(y) = \{x \in \mathbb{R}^n : ||x - y||_2 < \eta\}$. For $\eta > 0$ and a set $A \subset \mathbb{R}^n$ we denote the set $B_\eta(A) = \cup_{x \in A} B_\eta(x)$. For a set $X \subset \mathbb{R}^n$ we say $x \in X$ is an interior point of X if there exists $\varepsilon > 0$ such that $\{y \in \mathbb{R}^n : ||x-y|| < \varepsilon\} \subset X$. We denote the set of all interior points of X by X°. The point $x \in X$ is a limit point of X if for all $\varepsilon > 0$ there exists $y \in \{y \in \mathbb{R}^n/\{x\} : ||x-y|| < \varepsilon\}$ such that $y \in X$; we denote the set of all limit points of X, called the closure of X, as $(X)^{cl}$. Moreover, we denote the boundary of X by $\partial X = (X)^{cl}/X^{\circ}$. For $A \subset \mathbb{R}^n$ we denote the indicator function by $\mathbb{1}_A : \mathbb{R}^n \to \mathbb{R}$ that is defined as $\mathbb{1}_A(x) = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ otherwise.} \end{cases}$ For $B \subseteq \mathbb{R}^n$, $\mu(B) := \int_{\mathbb{R}^n} \mathbb{1}_B(x)dx$ is the Lebesgue measure of B. For sets $A, B \subset \mathbb{R}^n$, we denote the volume metric as $D_V(A,B)$, where $D_V(A,B) := \mu((A/B) \cup (B/A))$. We note that D_V is a metric (Defn. A.1), as shown in Lem. A.1. Let us denote bounded subsets of $\mathbb{R}^n$ by $\mathcal{B} := \{B \subset \mathbb{R}^n : \mu(B) < \infty\}$. If M is a subspace of a vector space X we denote equivalence relation $\sim_M$ for $x, y \in X$ by $x \sim_M y$ if $x - y \in M$. We denote quotient space by $X \pmod{M} := \{\{y \in X : y \sim_M x\} : x \in X\}$. For an open set $\Omega \subset \mathbb{R}^n$ and $\sigma > 0$ we denote $< \Omega >_\sigma := \{x \in \Omega : B_\sigma(x) \subset \Omega\}$. We denote the set of $n \times n$ matrices with strictly positive eigenvalues as S_n^{++}.

Function and Continuity Notation: For a function $f : X \to Y$ we denote the image set of the function as $Image\{f\} := \{y \in Y : \text{ there exists } x \in X \text{ such that } f(x) = y\}$. Let $C(\Omega, \Theta)$ be the set of continuous functions with domain $\Omega \subset \mathbb{R}^n$ and image $\Theta \subset \mathbb{R}^m$. We denote the set of locally and uniformly Lipschitz continuous functions on Θ_1 and Θ_2, Defn. 5.1, by $LocLip(\Theta_1, \Theta_2)$ and $Lip(\Theta_1, \Theta_2)$ respectively. For $\alpha \in \mathbb{N}^n$ we denote the partial derivative $D^\alpha f(x) := \Pi_{i=1}^n \frac{\partial^{\alpha_i} f}{\partial x_i^{\alpha_i}}(x)$ where by convention if $\alpha = [0,..,0]^T$ we denote $D^\alpha f(x) := f(x)$. We denote the set of i continuously differentiable functions by $C^i(\Omega, \Theta) := \{f \in C(\Omega, \Theta) : D^\alpha f \in C(\Omega, \Theta) \text{ for all } \alpha \in \mathbb{N}^n \text{ such that } \sum_{j=1}^n \alpha_j \leq i\}$. For $V \in C^1(\mathbb{R}^n, \mathbb{R})$ we denote $\nabla V := (\frac{\partial V}{\partial x_1}, ..., \frac{\partial V}{\partial x_n})^T$ and for $V \in C^1(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ we denote $\nabla_x V := (\frac{\partial V}{\partial x_1}, ..., \frac{\partial V}{\partial x_n})^T$ and $\nabla_t V = \frac{\partial V}{\partial x_{n+1}}$. We denote the essential supremum by $\operatorname{ess\,sup}_{x \in X} f(x) := \inf\{a \in \mathbb{R} : \mu(\{x \in X : f(x) > a\}) = 0\}$. We say $k : \mathbb{R} \to \mathbb{R}$ is a class K_∞ function (used in (8.10)), denoted $k \in K_\infty$, if $k : \mathbb{R} \to \mathbb{R}$ is monotonically increasing, $k(0) = 0$, and $\lim_{r \to \infty} k(r) = \infty$.

Sobolev Space Notation: For an open set $\Omega \subset \mathbb{R}^n$ and $p \in [1, \infty)$ we denote the set of p-integrable functions by $L^p(\Omega, \mathbb{R}) := \{f : \Omega \to \mathbb{R} \text{ measurable } : \int_\Omega |f|^p < \infty\}$, in the case $p = \infty$ we denote $L^\infty(\Omega, \mathbb{R}) := \{f : \Omega \to \mathbb{R} \text{ measurable } : \operatorname{ess\,sup}_{x \in \Omega} |f(x)| < \infty\}$. For $k \in \mathbb{N}$ and $1 \leq p \leq \infty$ we denote the Sobolev space of functions with weak derivatives (Defn. B.1) by $W^{k,p}(\Omega, \mathbb{R}) := \{u \in L^p(\Omega, \mathbb{R}) : D^\alpha u \in L^p(\Omega, \mathbb{R}) \text{ for all } |\alpha| \leq k\}$. For $u \in W^{k,p}(\Omega, \mathbb{R})$ we denote the Sobolev norm

$$||u||_{W^{k,p}(\Omega,\mathbb{R})} := \begin{cases} \left(\sum_{|\alpha| \leq k} \int_\Omega (D^\alpha u(x))^p dx \right)^{\frac{1}{p}} \text{ if } 1 \leq p < \infty \\ \sum_{|\alpha| \leq k} \operatorname{ess\,sup}_{x \in \Omega} \{|D^\alpha u(x)|\} \text{ if } p = \infty. \end{cases}$$

In the case $k = 0$ we have $W^{0,p}(\Omega, \mathbb{R}) = L^p(\Omega, \mathbb{R})$ and thus we use the notation $|| \cdot ||_{L^p(\Omega,\mathbb{R})} := || \cdot ||_{W^{0,p}(\Omega,\mathbb{R})}$. The σ-mollification of a function $V \in L^1(\Omega, \mathbb{R})$ is denoted by $[V]_\sigma :< \Omega >_\sigma \to \mathbb{R}$ and defined in Eq. (B.2).

Polynomial Notation: We denote the space of polynomials $p : \Omega \to \Theta$ by $\mathcal{P}(\Omega, \Theta)$ and polynomials with degree at most $d \in \mathbb{N}$ by $\mathcal{P}_d(\Omega, \Theta)$. We say $p \in \mathcal{P}_{2d}(\mathbb{R}^n, \mathbb{R})$ is Sum-of-Squares (SOS) if for $k \in \{1, ...k\} \subset \mathbb{N}$ there exists $p_i \in \mathcal{P}_d(\mathbb{R}^n, \mathbb{R})$ such that $p(x) = \sum_{i=1}^k (p_i(x))^2$. We denote $\sum_{SOS}^d$ to be the set of SOS polynomials of at most degree $d \in \mathbb{N}$ and the set of all SOS polynomials as $\sum_{SOS}$. We denote $Z_d : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^{\mathcal{N}_d}$ as the vector of monomials of degree $d \in \mathbb{N}$ or less, where $\mathcal{N}_d := \binom{d+n}{d}$.

Part 1

DISCRETE TIME

Chapter 3

MULTI-STAGE OPTIMIZATION PROBLEMS WITH BACKWARD SEPARABLE COSTS

> In my next life I want to live my life backwards.
> You start out dead and get that out of the way.
> Then you wake up in an old people's home
> feeling better every day.
>
> Woody Allen

3.1 Background and Motivation

Consider Multi-Stage Optimization Problems (MSOPs) of the following form,

$$\begin{aligned}&\inf\left\{ J(u(0),...,u(T-1),x(0),...,x(T)) \right\}\\ &\quad x(t+1)=f(x(t),u(t),t) \text{ for } t=0,..,T-1, \text{ and } x(0)=x_0,\\ &\quad x(t)\in X_t\subset \mathbb{R}^n,\ u(t)\in U\subset\mathbb{R}^m \text{ for } t=0,..,T.\end{aligned}$$

Such problems consist of 1) a cost function $J:\mathbb{R}^{m\times T}\times\mathbb{R}^{n\times(T+1)}\to\mathbb{R}$, 2) an underlying discrete-time dynamical system governed by the plant equation $f:\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{N}\to\mathbb{R}^n$, 3) a state space $X_t\subset\mathbb{R}^n$, 4) an admissible input space $U\subset\mathbb{R}^m$, and 5) a terminal time $T>0$. Examples of such optimization problems include: optimal battery scheduling to minimize consumer electricity bills considered in Jones and Peet (2017); energy-optimal speed planning for road vehicles considered in Zeng and Wang (2018); optimal maintenance of manufacturing systems considered in Liu *et al.* (2019); etc.

MSOPs are members of the class of constrained nonlinear optimization problems. Such optimization problems over small time horizons can be solved using nonlinear solvers such as SNOPT, found in Gill *et al.* (2005). However, arguably the most commonly used class of methods for solving MSOPs is Dynamic Programming (DP), see Bertsekas (1995). DP methods exploit the structure of MSOPs to decompose the optimization problem into lower dimensional sub-problems that can be solved recursively to give the solution to the original higher dimensional MSOP. Typically, DP is used to solve problems with cost functions of the form $J(\mathbf{u}, \mathbf{x}) = \sum_{t=0}^{T-1} c_t(x(t), u(t)) + c_T(x(T))$. These functions, defined in Definition 3.2, are called additively separable functions, as they can be additively separated into sub-functions, each of which only depend on a single time-stage, $t \in \{0, ..., T\}$. In the additively separable case it was shown in Bellman (1966) that if we can find a function F that satisfies Bellman's Equation,

$$\begin{aligned} F(x, T) &= c_T(x) \quad \text{ for all } x \in X_T \\ F(x, t) &= \inf_{u \in \Gamma_{x,t}} \left\{ c_t(x, u) + F(f(x, u, t), t + 1) \right\} \text{ for all } x \in X_t, t \in \{0, .., T - 1\}, \end{aligned}$$

where $\Gamma_{x,t} := \{u \in U : f(x, u, t) \in X_t\}$, then a necessary and sufficient condition for a feasible input and state sequence, $\mathbf{u} = (u(0), ..., u(T - 1))$ and $\mathbf{x} = (x(0), ..., x(T))$, to be optimal is

$$u(t) \in \arg \inf_{u \in \Gamma_{x(t),t}} \left\{ c_t(x(t), u) + F(f(x(t), u, t), t + 1) \right\} \text{ for all } t \in \{0, .., T - 1\}.$$

We consider MSOPs with cost functions of the more general form $J(\mathbf{u}, \mathbf{x}) = \phi_0(x(0), u(0), \phi_1(x(1), u(1), \dots \phi_T(x(T)) \dots))$, where maps $\phi_t : X_t \times U \times \mathbb{R} \to \mathbb{R}$ are monotonic in their third argument for $t = 0, \cdots T - 1$. Such functions are called monotonically backward separable, defined in Definition 3.3, and shown to contain the class of additively separable functions in Lemma 3.1. For MSOPs with monotonically backward separable cost functions we show in Theorem 3.2 that if we can find a function V that satisfies

$$\begin{aligned} V(x,T) &= \phi_T(x) \text{ for all } x \in X_T \\ V(x,t) &= \inf_{u \in \Gamma_{x,t}} \left\{ \phi_t(x,u,V(f(x,u,t),t+1)) \right\} \text{ for all } x \in X_t, t \in \{0,..,T-1\}, \end{aligned} \tag{3.1}$$

where $\Gamma_{x,t} := \{u \in U : f(x,u,t) \in X_t\}$, then a necessary and sufficient for a feasible input and state sequence, $\mathbf{u} = (u(0),...,u(T-1))$ and $\mathbf{x} = (x(0),...,x(T))$, to be optimal is

$$u(t) \in \arg \inf_{u \in \Gamma_{x(t),t}} \left\{ \phi_t\left(x(t),u,V(f(x(t),u,t),t+1)\right) \right\} \text{ for all } t \in \{0,..,T-1\}.$$

Equation (3.1) can be thought of as a generalization of Bellman's Equation; as it is shown in Corollary 3.2 that in the special case when the cost function is additively separable Equation (3.1) reduces to Bellman's Equation. We therefore refer to Equation (3.1) as the Generalized Bellman's Equation (GBE). Through several examples we show a solution, V, to the GBE can be obtained numerically by recursively solving the GBE backwards in time for each element of X_t, the same way Bellman's Equation is solved, thereby extending traditional DP methods to solve a larger class of MSOPs with non-additively separable cost functions.

By recursively solving the GBE it is possible to synthesize optimal input sequences for many important practical problems. In this chapter we consider two such problems; path planning with obstacle avoidance and maximal invariant sets. First, we define the path planning problem as the search for a sequence of inputs that drives a dynamical system to a target set in minimum time while avoiding obstacles defined by subsets of the state-space. In Section 3.5 we show that such problems can be formulated as an MSOP with monotonically backward separable objective, of form $J(\mathbf{u},\mathbf{x}) = \min\{\inf\{t \in [0,T] : x(t) \in S\}, T\}$, implying that the solution to the path planning problem can be found using the solution to the GBE.

Similarly, in Section 3.6 we show that computation of maximal invariant sets can be formulated as an MSOP with monotonically backward separable objective of form $J(\mathbf{u}, \mathbf{x}) = \max\{\max_{0 \le k \le T-1}\{c_k(u(k), x(k))\}, c_T(x(T))\}$.

Path planning with obstacle avoidance has been extensively studied (see the surveys given in Dreyfus (1969), and Gallo and Pallottino (1988)) and has many applications; including UAV surveillance considered in Xie *et al.* (2019). In Rippel *et al.* (2005) the path planning problem is separated into two separate problems: the "geometric problem", in which the shortest curve, $\tilde{x}(t)$, between the initial set and target set is calculated, and the "tracking problem", in which a controller, $u(t)$, is synthesized so that $\sum_{t=0}^{T} ||x(t) - \tilde{x}(t)||_2^2$ is minimized, where $x(t+1) = f(x(t), u(t), t)$ and $|| \cdot ||_2$ is the Euclidean norm. Separating the path planning problem allows for the use of efficient algorithms such as A^*-search to solve the "geometric problem" and LQR control to solve the "tracking problem", however, there is no guaranteed that this method will produce the true solution to the original path planning problem. The same approach is used in Cowlagi and Tsiotras (2011), where it is shown through numerical examples that a controller closer to optimality can be derived when the state space is augmented with historic trajectory information. Our approach of using the GBE to solve the path planning does not separate the problem into the "geometric or "tracking" problem and thus does not require any state augmentation. For systems described in continuous time (rather than the discrete systems considered in this chapter) with obstacles that satisfy certain boundary curvature assumptions, assumptions not made in this chapter, it has been shown in Savkin and Hoy (2013) that a path planning sliding mode controller can be efficiently computed. Furthermore, this sliding mode controller can be used for effective path planning in unknown environments, a case not considered in this chapter.

The GBE can also be used in the application of computing the Finite Time Horizon

Maximal Invariant Set (FTHMIS), defined as the largest set of initial conditions for a discrete time process such that there exists a feasible input sequence for which the state of the system never violates a time-varying constraint. Knowledge of this set can be used to design controllers that ensure the system never violates given safety constraints. We show that FTHMISs are equivalent to the sublevel set of solutions to the GBE. To the best of the authors knowledge the problem of computing FTHMISs has not previously been addressed in the literature. However, a proposed methodology for computing maximal invariant sets over infinite time horizons can be found in Xue and Zhan (2018); Esterhuizen *et al.* (2019); Wang *et al.* (2019). Similar continuous-time formulations of this problem can be found in Jones and Peet (2019c,b).

Substantial work on generalizations of Bellman's Equation for both infinite and finite time MSOPs can be found in Bertsekas (2018). Our work differs from Bertsekas (2018) as rather than attempting to generalize the "Bellman's operator", as Bertsekas (2018) does, we consider a wider class of cost functions associated with MSOPs, introducing monotonically backward separable cost functions, leading to a derivation of the GBE (3.1). Unlike in Bertsekas (2018), we formalize the link between the cost function of an MSOP and the GBE (3.1). Other examples in the literature of MSOPs with non-additively separable cost functions can be found in the pioneering work of Li and Haimes (1991, 1990b,a); Li (1990). Li considered MSOPs with k-separable cost functions; functions of the form $J(\mathbf{u}, \mathbf{x}) = H(J_1(\mathbf{u}, \mathbf{x}), ..., J_k(\mathbf{u}, \mathbf{x}))$, where $H : \mathbb{R}^k \to \mathbb{R}$ is strictly increasing and differentiable, and each of the functions, J_i, are differentiable monotonically backward separable functions. Li showed that for problems in this class of MSOP, an equivalent multi-objective optimization problem with k-separable cost functions can be constructed. The multi-objective optimization problem can then be analytically solved, using methods relying of the differentiability of the cost function, to find the optimal input sequence for the MSOP. We do not assume, as in Li, that the cost function is differentiable or k-separable and our solution does not require the solution of a multi-objective optimization problem.

In related work, coherent risk measures, from Shapiro and Ugurlu (2016); Shapiro (2009); Ruszczyński (2010), result in MSOPs with non-additively separable cost functions of the form $J(\mathbf{u}, \mathbf{x}) = c_0(x(0), u(0)) + \rho_1(c_1(x(1), u(1)) + \rho_2(c_2(x(2), u(2)) + ... + \rho_T(c_T(x(T)))....))$. Such MSOPs are solved recursively using a modified Bellman's Equation. Coherent risk measure functions are a special case of monotonically backward separable functions; in this case our GBE reduces to the previously proposed modified Bellman's equation.

3.2 Monotonically Backward Separable Functions

In this section we formally define the general class of optimization problems called Multi-Stage Optimization Problems (MSOPs) we are concerned with. We show this class contains the class of problems that classical Dynamic Programming (DP) theory is able to solve; MSOPs with additively separable cost functions Eq. (3.3). We then propose a more general class of cost functions called monotonically backward separable functions, Eq. (3.4), that contains the class of additively separable functions. Using this framework we are then able to derive necessary and sufficient conditions for an input sequence to solve an MSOP with monotonically backward separable cost function. Such conditions are shown to reduce to the classical conditions proposed by Bellman (1966) in the special case when the cost function is additively separable.

Definition 3.1. *For a given initial condition $x_0 \in \mathbb{R}^n$, for every tuple of the form $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\}$, where $J : \mathbb{R}^{m \times T} \times \mathbb{R}^{n \times (T+1)} \to \mathbb{R}$, $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{N} \to \mathbb{R}^n$,*

$X_t \subset \mathbb{R}^n$, $U \subset \mathbb{R}^m$, and $T \in \mathbb{N}$, we associate a MSOP of the following form

$$\begin{aligned}
&(\mathbf{u}^*, \mathbf{x}^*) \in \arg \inf_{\mathbf{u},\mathbf{x}} J(\mathbf{u}, \mathbf{x}) \textit{ subject to:} \qquad (3.2)\\
&x(t+1) = f(x(t), u(t), t) \textit{ for } t = 0, .., T-1\\
&x(0) = x_0, \; x(t) \in X_t \subset \mathbb{R}^n \textit{ for } t = 0, .., T\\
&u(t) \in U \subset \mathbb{R}^m \textit{ for } t = 0, .., T-1\\
&\mathbf{u} = (u(0), ..., u(T-1)) \textit{ and } \mathbf{x} = (x(0), ..., x(T))
\end{aligned}$$

For a given tuple $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\}$, the function J represents the cost function, f represents the plant dynamics, X_t represents the set of admissible states at time step $t \in \{0, ..., T\}$, and U represents the set of admissible inputs.

Classical DP theory is concerned with the special case when the cost function, $J : \mathbb{R}^{m \times T} \times \mathbb{R}^{n \times (T+1)} \to \mathbb{R}$, has an additively separable structure defined as follows.

Definition 3.2. *The function $J : U^T \times \Pi_{t=0}^T X_t \to \mathbb{R}$ is said to be **additively separable** if there exists functions, $c_T(x) : X_T \to \mathbb{R}$, and $c_t : X_t \times U \to \mathbb{R}$ for $t = 0, \cdots T-1$ such that,*

$$J(\mathbf{u}, \mathbf{x}) = \sum_{t=0}^{T-1} c_t(x(t), u(t)) + c_T(x(T)), \qquad (3.3)$$

where $\mathbf{u} = (u(0), ..., u(T-1))$ and $\mathbf{x} = (x(0), ..., x(T))$.

We consider the class of "monotonic backward separable" cost functions defined next. The definition of this class of functions uses the image set of a function. Specifically, for a function $f : X \to Y$ we denote the image set of the function as $Image\{f\} := \{y \in Y : \text{ there exists } x \in X \text{ such that } f(x) = y\}$.

Definition 3.3. *The function $J : U^T \times \Pi_{t=0}^T X_t \to \mathbb{R}$, where $U \subset \mathbb{R}^m$ and $X_t \subset \mathbb{R}^n$ is said to be **monotonically backward separable** if there exists representation maps,*

$\phi_T : X_T \to \mathbb{R}$, and $\phi_t : X_t \times U \times Image\{\phi_{t+1}\} \to \mathbb{R}$ for $t = 0, \cdots T-1$ such that the following holds:

1. *The function J can be expressed as the composition of representation maps, $\{\phi_t\}_{t=0}^T$, ordered backwards in time. That is J satisfies*

$$J(\mathbf{u}, \mathbf{x}) = \phi_0(x(0), u(0), \phi_1(x(1), u(1), \dots \phi_T(x(T)) \dots)), \tag{3.4}$$

   where $\mathbf{u} = (u(0), ..., u(T-1))$ and $\mathbf{x} = (x(0), ..., x(T))$.

2. *Each representation map, ϕ_t, is monotonic in its third argument. That is if $z, w \in Image\{\phi_{t+1}\}$ are such that $z \geq w$ then*

$$\phi_t(x, u, z) \geq \phi_t(x, u, w) \text{ for all } (x, u) \in X_t \times U \tag{3.5}$$

*Moreover if J also satisfies the following properties than we say J is **naturally monotonically backward separable**:*

1. *Each representation map, ϕ_t, is upper semi-continuous in its third argument. That is for any $t \in \{0, .., T-1\}$, $x \in X_t$, $u \in U$ and any monotonically decreasing sequence $\{z_n\}_{n \in \mathbb{N}} \subset Image\{\phi_{t+1}\}$, such that $z_{n+1} \leq z_n$ for all $n \in \mathbb{N}$, then*

$$\lim_{n \to \infty} \phi_t(x, u, z_n) = \phi_t(x, u, \lim_{n \to \infty} z_n). \tag{3.6}$$

2. *Each representation map, ϕ_t, satisfies the following boundedness property. For any $t \in \{0, ..., T-1\}$ and $(x, u, z) \in X_t \times U \times Image\{\phi_{t+1}\}$ we have $|\phi_t(x, u, z)| < \infty$ and for all $x \in X_T$ we have $|\phi_T(x)| < \infty$; that is for each $t \in \{0, ..., T\}$ there exists $R > 0$ such that*

$$Image\{\phi_t\} \subset \{x \in \mathbb{R} : |x| < R\}. \tag{3.7}$$

We show in Sec. 3.3 that monotonically backward separable functions share a deep connection with Bellman's Principle of Optimality (Defn. 3.6). However, we also consider naturally monotonically backward separable functions as the added semi-continuity and boundedness properties are used in the derivation of necessary and sufficient conditions for an input sequence to solve an MSOP with naturally monotonically backward separable cost function (Theorem 3.2). These necessary and sufficient conditions are later used in Section 3.4 to design efficient numerical algorithms for solving MSOPs with naturally monotonically backward separable cost functions.

Before proceeding we next introduce notation for the class of MSOPs with naturally monotonically backward separable cost functions and additively separable cost functions. We will later in Lemma 3.1 show that the class MSOPs with naturally monotonically backward separable cost functions contains the class of MSOPs with additively separable cost functions.

Definition 3.4. *We say the five tuple $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\}$ is a* ***naturally backward separable MSOP*** *or $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$ if J is a naturally monotonically backward separable function, $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{N} \to \mathbb{R}^n$, $X_t \subset \mathbb{R}^n$, $U \subset \mathbb{R}^m$, and $T \in \mathbb{N}$. Moreover, if instead J is an additively separable function with associated cost functions, $\{c_t\}_{t=0}^T$, that are bounded over their domains, we say that the five tuple $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\}$ is an* ***additively separable MSOP*** *or $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Additive}$.*

We next show the class of backward separable MSOPs contains the class of additively separable MSOPs.

Lemma 3.1. $\mathcal{M}^{Discrete}_{Additive} \subseteq \mathcal{M}^{Discrete}_{Backward}$.

Proof. To prove Lemma 3.1 we must show every additively separable function with bounded cost functions is a naturally monotonically backward separable function.

Given an additively separable function, J, we know there exists functions $\{c_t\}_{0\leq t\leq T}$ such that Eq. (3.3) holds. To prove J is monotonically backward separable we construct representation maps $\{\phi_t\}_{t=0}^{T}$ such that Eqs. (3.4) and (3.5) hold. We define these representation maps as follows:

$$\begin{aligned} &\phi_i(x,u,z) = c_i(x,u) + z \quad \text{for } i = 0, \cdots, T-1 \\ &\phi_T(x,w) = c_T(x). \end{aligned} \tag{3.8}$$

Now, $\frac{\partial \phi_t(x,y,z)}{\partial z} = 1 > 0$ for all $t \in \{0, ..., T-1\}$, $x \in X_t$ and $u \in U$, implying the monotonicity property in Eq. (3.5).

Now assuming the functions $\{c_t\}_{t=0}^{T}$ are bounded over $X_t \times U$ it follows trivially that the representation maps $\{\phi_t\}_{t=0}^{T}$, given in Eq. (3.8), satisfy the semi-continuity and boundedness properties given in Eqs. (3.6) and (3.7). Thus J is naturally monotonically backward separable function. □

Lemma 3.1 showed $\mathcal{M}_{Additive}^{Discrete} \subseteq \mathcal{M}_{Backward}^{Discrete}$. We next strengthen this result by showing $\mathcal{M}_{Additive}^{Discrete}$ is a strict subset of $\mathcal{M}_{Backward}^{Discrete}$, that is $\mathcal{M}_{Additive}^{Discrete} \subset \mathcal{M}_{Backward}^{Discrete}$.

Corollary 3.1. $\mathcal{M}_{Additive}^{Discrete} \subset \mathcal{M}_{Backward}^{Discrete}$.

Proof. By Lemma 3.1 we have that $\mathcal{M}_{Additive}^{Discrete} \subseteq \mathcal{M}_{Backward}^{Discrete}$. Therefore, to show $\mathcal{M}_{Additive}^{Discrete} \subset \mathcal{M}_{Backward}^{Discrete}$ we must show there exists a naturally monotonically backward separable function (Defn. 3.3) that is not additively separable. Consider the function $J : [0,1]^2 \to \mathbb{R}$ defined by $J(x(0), x(1)) = x(0)x(1)$. Then J is clearly a naturally monotonically backward separable function since it can be written in the Form (3.4) using the representation maps,

$$\phi_0(x,z) = xz \text{ and } \phi_1(x) = x, \tag{3.9}$$

it satisfies Eq. (3.6) since ϕ_0 and ϕ_1 are both clearly continuous, and it satisfies Eq. (3.7) since $Image\{\phi_0\} \subseteq \{x \in \mathbb{R} : |x| \leq 1\}$.

Now, for contradiction suppose J is an additively separable function (Defn. 3.2). Then there exists $c_0 : [0,1] \to \mathbb{R}$ and $c_1 : [0,1] \to \mathbb{R}$ such that $J(x(1), x(2)) = c_0(x(0)) + c_1(x(1))$ and therefore

$$x(1)x(2) = c_0(x(0)) + c_1(x(1)) \text{ for all } x(1), x(2) \in [0,1]^2. \tag{3.10}$$

Hence, by Eq. (3.10) it now follows

$$0 = c_0(0) + c_1(0) \text{ by subsituting } x(0) = 0 \text{ and } x(1) = 0. \tag{3.11}$$

$$0 = c_0(x) + c_1(0) \text{ by subsituting } x(0) = x \in [0,1] \text{ and } x(1) = 0.$$

By Eq. (3.11) we get that $c_0(x) = c_0(0)$ for all $x \in [0,1]$. By a similar argument we get $c_1(x) = c_1(0)$ for all $x \in [0,1]$. Now, by substituting $(x(1), x(2)) = (0.5, 0.5)$ into Eq. (3.10) we get $0.25 = c_0(0) + c_1(0)$. Alternatively, substituting $(x(1), x(2)) = (1,1)$ into Eq. (3.10) we get $1 = c_0(0) + c_1(0)$. Thus it follows $0.25 = c_0(0) + c_1(0) = 1$, providing a contradiction. We conclude that J is monotonically backward separable but not additively separable. □

Later, in Section 3.2.3, we will provide several more examples of monotonically backward separable functions (Defn. 3.3), different from $J(x(1), x(2)) = x(1)x(2)$ in Corollary 3.1, that are not necessarily additively separable (Defn. 3.2).

3.2.1 Interchanging the Order of Composition and Infimum in Monotonically Backward Separable Functions

As we will show in Lemma 3.2, naturally monotonically backward separable functions have the special property that the order of an infimum and composition of representation maps can be interchanged. To show this we must use the monotonic convergence theorem.

Theorem 3.1 (Monotone Convergence Theorem)**.** *Suppose* $\{z_n\}_{n\in\mathbb{N}} \subset \mathbb{R}$ *is a bounded sequence that is monotonically decreasing,* $z_{n+1} \le z_n$ *for all* $n \in \mathbb{N}$*. Then* $\lim_{n\to\infty} z_n = \inf_{n\in\mathbb{N}} z_n$.

Before proving in Lemma 3.2 we introduce notation for the set of feasible controls. Given a tuple $\{J, f, \{X_t\}_{0\le t\le T}, U, T\}$ for $x \in X_t$ and $s \in \{0, ..., T-1\}$ we denote

$$\Gamma_{x,s} := \{u \in U : f(x,u,s) \in X_{s+1}\}.$$

Moreover we say

$$(u(s), ..., u(T-1)) \in \Gamma_{x_0,[s,T-1]} \tag{3.12}$$

if $u(t) \in \Gamma_{x(t),t}$ for all $t \in \{s, ..., T-1\}$, where $x(s) = x_0$ and $x(k+1) = f(x(k), u(k), k)$ for $k \in \{s, ..., T-1\}$.

Lemma 3.2 (Interchanging the order of composition and infimum)**.** *Consider an MSOP of Form* (3.2) *associated with* $\{J, f, \{X_t\}_{0\le t\le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$*. Suppose* $\Gamma_{x,t} \ne \emptyset$ *for all* $(x,t) \in X_t \times \{0, ..., T-1\}$*. Then for* $k \in \{0, ..., T-1\}$ *and any* $x \in X_k$ *we have*

$$\begin{aligned}
&\inf_{u(k)\in\Gamma_{x,k}} \Bigg\{\phi_k\Bigg(x(k), u(k), \inf_{(u(k+1),...,u(T-1))\in\Gamma_{x(k+1),[k+1,T-1]}} \Bigg\{\phi_{k+1}(\\
&\qquad\qquad x(k+1), u(k+1), \phi_{k+2}(x(k+2), u(k+2), ...\phi_T(x(T))...))\Bigg\}\Bigg)\Bigg\} \\
&= \inf_{(u(k),...,u(T-1))\in\Gamma_{x,[k,T-1]}} \Bigg\{\phi_k(x(k), u(k), \phi_{k+1}(x(k+1), u(k+1), ...\phi_T(x(T))...))\Bigg\},
\end{aligned} \tag{3.13}$$

where $\{\phi_t\}_{t=0}^T$ *are the representation maps of* $J : U^T \times \Pi_{t=0}^T X_t \to \mathbb{R}$*, and* $x(t+1) = f(x(t), u(t), t)$ *for* $t \in \{k, ..., T-1\}$ *and* $x(k) = x$.

Proof. To show Eq. (3.13) we will split the proof into two parts. In Part 1 we will show the left hand side of Eq. (3.13) is less than or equal to the right hand side of

Eq. (3.13). In Part 2 we will show the right hand side of Eq. (3.13) is less than or equal to the left hand side of Eq. (3.13).

Part 1 of proof: By the definition of the infimum it follows for all $y \in X_{k+1}$ that

$$\inf_{(u(k+1),...,u(T-1)) \in \Gamma_{y,[k+1,T-1]}} \phi_{k+1}(x(k+1), u(k+1), ...\phi_T(x(T))...) \tag{3.14}$$
$$\leq \phi_{k+1}(\tilde{x}(k+1), \tilde{u}(k+1), ...\phi_T(\tilde{x}(T))...),$$

for any $(\tilde{u}(k+1), ..., \tilde{u}(T-1)) \in \Gamma_{x(k+1),[k+1,T-1]}$, where $\tilde{x}(t+1) = f(\tilde{x}(t), \tilde{u}(t), t)$, $x(t+1) = f(x(t), u(t), t)$ for $t \in \{k+1, ...T-1\}$, and $x(k+1) = \tilde{x}(k+1) = y$.

Since ϕ_k is monotonic in its third argument (Eq. (3.5)) it follows from Eq. (3.14) that for any $(x, u) \in X_k \times \Gamma_{x,k}$ that

$$\phi_k(x(k), u(k), \inf_{(u(k+1),...,u(T-1)) \in \Gamma_{x(k+1),[k+1,T-1]}} \{\phi_{k+1}(x(k+1), u(k+1), ...\phi_T(x(T))...)\}$$
$$\leq \phi_k(x(k), u(k), \phi_{k+1}(\tilde{x}(k+1), \tilde{u}(k+1), ...\phi_T(\tilde{x}(T))...)), \tag{3.15}$$

for any $(\tilde{u}(k+1), ..., \tilde{u}(T-1)) \in \Gamma_{x(k+1),[k+1,T-1]}$, where $\tilde{x}(t+1) = f(\tilde{x}(t), \tilde{u}(t), t)$, $x(t+1) = f(x(t), u(t), t)$ for $t \in \{k, ...T-1\}$, $x(k) = \tilde{x}(k) = x$, and $u(k) = u$.

Now, since Eq. (3.15) holds for any $u \in \Gamma_{x,k}$ and $(\tilde{u}(k+1), ..., \tilde{u}(T-1)) \in \Gamma_{x(k+1),[k+1,T-1]}$ we are able to take the infimum over these in Eq. (3.15), deducing the left hand side of Eq. (3.13) is less or than or equal to its right hand side.

Part 2 of proof: Let us fix $(x, u) \in X_k \times \Gamma_{x(k),k}$. Since $\Gamma_{x,t} \neq \emptyset$ for all $(x, t) \in X_t \times \{0, ..., T-1\}$ it follows from the definition of the infimum for all $n \in \mathbb{N}$ there exists $(u_n(k+1), ..., u_n(T-1)) \in \Gamma_{x(k+1),[k+1,T-1]}$ such that

$$\inf_{(u(k+1),...,u(T-1)) \in \Gamma_{x(k+1),[k+1,T-1]}} \phi_{k+1}(x(k+1), u(k+1), ...\phi_T(x(T))...)$$
$$\leq \phi_{k+1}(x_n(k+1), u_n(k+1), ...\phi_T(x_n(T))...) \tag{3.16}$$
$$\leq \inf_{(u(k+1),...,u(T-1)) \in \Gamma_{x(k+1),[k+1,T-1]}} \phi_{k+1}(x(k+1), u(k+1), ...\phi_T(x(T))...) + \frac{1}{n},$$

where $x_n(t+1) = f(x_n(t), u_n(t), t)$ for $t \in \{k+1, ..., T-1\}$, and $x_n(k+1) = x(k+1) = f(x, u, k)$.

Now, let us denote $a_n := \phi_{k+1}(x_n(k+1), u_n(k+1), ...\phi_T(x_n(T))...)$. It follows from Eq. (3.16) that,

$$\lim_{n \to \infty} a_n = \inf_{(u(k+1),...,u(T-1)) \in \Gamma_{x(k+1),[k+1,T-1]}} \phi_{k+1}(x(k+1), u(k+1), ...\phi_T(x(T))...),$$

and

$$a_n \geq \inf_{(u(k+1),...,u(T-1)) \in \Gamma_{x(k+1),[k+1,T-1]}} \phi_{k+1}(x(k+1), u(k+1), ...\phi_T(x(T))...) \text{ for all } n \in \mathbb{N}.$$

Since $\{a_n\}_{n \in \mathbb{N}}$ converges to some limit from above there exists a monotonically decreasing subsequence $\{b_n\}_{n \in \mathbb{N}} \subseteq \{a_n\}_{n \in \mathbb{N}}$ such that $b_{n+1} \leq b_n$ for $n \in \mathbb{N}$. Using $\{b_n\}_{n \in \mathbb{N}}$ we now define

$$z_n := \phi_k(x, u, b_n).$$

Since ϕ_k is monotonic in its third argument (Eq. (3.5)) and $b_{n+1} \leq b_n$ it follows $z_{n+1} = \phi_k(x, u, b_{n+1}) \leq \phi_k(x, u, b_n) \leq z_n$. Hence $\{z_n\}_{n \in \mathbb{N}}$ is a monotonically decreasing sequence. Moreover, since ϕ_k has the property that it is a bounded over $X_k \times U \times Image\{\phi_{k+1}\}$ (Eq. (3.7)) it follows that $\{z_n\}_{n \in \mathbb{N}}$ is a bounded sequence. Now by the monotone convergence theorem (Theorem 3.1) we have that $\inf_{n \in \mathbb{N}} z_n = \lim_{n \to \infty} z_n$.

It now follows since ϕ_k is upper semi-continuous (Eq. (3.6)) in its third argument that

$$
\begin{aligned}
&\inf_{(u(k+1),...,u(T-1))\in\Gamma_{x(k+1),[k,T-1]}} \{\phi_k(x,u,\phi_{k+1}(x(k+1),u(k+1),...\phi_T(x(T))...))\} \quad (3.17)\\
&\leq \inf_{n\in\mathbb{N}} \phi_k(x,u,\phi_{k+1}(x_n(k+1),u_n(k+1),...\phi_T(x_n(T))...))\\
&\leq \inf_{n\in\mathbb{N}} z_n = \lim_{n\to\infty} z_n\\
&= \lim_{n\to\infty} \phi_k(x,u,b_n) = \phi_k(x,u,\lim_{n\to\infty} b_n) = \phi_k(x,u,\lim_{n\to\infty} a_n)\\
&= \phi_k(x,u, \inf_{(u(k+1),...,u(T-1))\in\Gamma_{x(k+1),[k+1,T-1]}} \{\phi_{k+1}(x(k+1),u(k+1),...\phi_T(x(T))...)\}.
\end{aligned}
$$

Since Eq. (3.17) holds for any arbitrarily selected $(x,u) \in X_k \times \Gamma_{x,k}$ we are able to take the infimum with respect to $u \in \Gamma_{x,k}$, showing the right hand side of Eq. (3.13) is less than or equal to its left hand side.

In Part 1 of the proof we have shown that the left hand side of Eq. (3.13) is less than or equal to the right hand side of Eq. (3.13). In Part 2 of the proof we have shown that the right hand side of Eq. (3.13) is less than or equal to the left hand side of Eq. (3.13). Putting these two parts together we deduce the left hand side must equal the right hand side, therefore completing the proof and showing Eq. (3.13) holds. □

3.2.2 A Generalization of Bellman's Equation

Additively separable MSOPs, $\{J,f,\{X_t\}_{0\leq t\leq T},U,T\} \in \mathcal{M}^{Discrete}_{Additive}$, can be solved recursively using Bellman's Equation, shown in Bellman (1966). In this section we show that a similar approach can be used to solve backward separable MSOPs, $\{J,f,\{X_t\}_{0\leq t\leq T},U,T\} \in \mathcal{M}^{Discrete}_{Backward}$.

We next define conditions under which a function, V, is said to be a *value function* for an associated MSOP.

Definition 3.5. *Consider the following backward separable MSOP,*

$\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$, *where* $J : \mathbb{R}^{m \times T} \times \mathbb{R}^{n \times (T+1)} \to \mathbb{R}$ *has representation functions* $\{\phi_t\}_{0 \le t \le T}$. *We say the function* $V : \mathbb{R}^n \times [0, T] \to \mathbb{R}$ *is a* ***value function*** *of the MSOP* $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$ *if for all* $x \in X_T$

$$V(x, T) = \phi_T(x), \tag{3.18}$$

and for all $x \in X_t$ *and* $t \in \{0, ..., T-1\}$

$$\begin{aligned} V(x,t) = \inf_{u(t) \in \Gamma_{x,t}, \ldots, u(T-1) \in \Gamma_{x(T-1), T-1}} \Bigg\{ & \qquad (3.19) \\ & \phi_t(x(t), u(t), \phi_{t+1}(x(t+1), u(t+1), \ldots \phi_T(x(T)) \ldots)) \Bigg\}, \end{aligned}$$

where $x(t) = x$ *and* $x(k+1) = f(x(k), u(k), k)$ *for* $k \in \{t, ..., T-1\}$.

We note that the value function has the special property that $V(x_0, 0) = J^*$, where J^* is the minimum value of the cost function of the MSOP (3.2). In the special case when J is an additively separable function the value function reduces to the optimal cost-to-go function.

Proposition 3.1 (Generalized Bellman's Equation (GBE))**.** *Consider an MSOP of Form* (3.2) *associated with* $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$. *Suppose* $\{\phi_t\}_{t=0}^T$ *are the representation maps of* $J : U^T \times \Pi_{t=0}^T X_t \to \mathbb{R}$ *(Defn. 3.3) and* $\Gamma_{x,t} \neq \emptyset$ *for all* $(x, t) \in X_t \times \{0, ..., T-1\}$. *Then if* $F : \mathbb{R}^n \times [0, T] \to \mathbb{R}$ *satisfies*

$$\begin{aligned} & F(x,T) = \phi_T(x) \text{ for all } x \in X_T \qquad \text{and} \qquad (3.20) \\ & F(x,t) = \inf_{u \in \Gamma_{x,t}} \Big\{ \phi_t(x, u, F(f(x,u,t), t+1)) \Big\} \text{ for all } x \in X_t, t \in \{0, .., T-1\}, \end{aligned}$$

then F *is a value function (Defn. 3.5) of the backward separable MSOP* $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$.

Proof. Suppose F satisfies Eq. (3.20). To show F is a value function of the backward separable MSOP $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$ we must show it satisfies Equations (3.18) and (3.19). We prove this using backward induction in the time variable of F. Clearly F satisfies Eq. (3.18) for $k = T$. Now, for our induction hypothesis, let us assume for some $k \in \{0, ..., T-1\}$ that F satisfies Eq. (3.19) at time-stage $k+1$ for all $x \in X_{k+1}$. We will now show that the induction hypothesis implies F must also satisfy Eq. (3.19) at time-stage k for all $x \in X_k$. Letting $x \in X_k$ we have

$$\begin{aligned}
F(x,k) &= \inf_{u \in \Gamma_{x,k}} \left\{ \phi_k(x, u, F(f(x,u,k), k+1)) \right\} \\
&= \inf_{u \in \Gamma_{x,k}} \left\{ \phi_k \left(x, u, \inf_{u(k+1) \in \Gamma_{x(k+1),k+1}, ..., u(T-1) \in \Gamma_{x(T-1),T-1}} \left\{ \phi_{k+1}(\right. \right. \right. \\
&\qquad \left. \left. \left. x(k+1), u(k+1), \phi_{k+2}(x(k+2), u(k+2), ...\phi_T(x(T))...)) \right\} \right) \right\} \\
&= \inf_{u(k) \in \Gamma_{x,k}, ..., u(T-1) \in \Gamma_{x(T-1),T-1}} \left\{ \phi_k(x(k), u(k), \phi_{k+1}(x(k+1), u(k+1), ...\phi_T(x(T))...)) \right\},
\end{aligned}$$

where $x(k) = x$ and $x(t+1) = f(x(t), u(t), t)$ for $t \in \{k, ..., T-1\}$. The first equality follows as F satisfies Eq. (3.20); the second equality follows from the induction hypothesis; the third equality follows by Lemma 3.2.

Therefore, by backward induction, we conclude F satisfies Eqs. (3.18) and (3.19) and hence is a value function for the MSOP associated with the tuple $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$. □

We next propose sufficient conditions showing an input sequence is optimal if it recursively minimizes the right hand side of the GBE (3.20). Later in Theorem 3.2 we propose necessary and sufficient conditions involving the GBE (3.20).

Proposition 3.2 (Sufficient conditions for optimality). *Consider an MSOP of Form* (3.2) *associated with* $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$. *Suppose* $\{\phi_t\}_{t=0}^T$ *are the representation maps of* $J : U^T \times \Pi_{t=0}^T X_t \to \mathbb{R}$ *(Defn. 3.3),* $\Gamma_{x,t} \neq \emptyset$ *for all* $(x,t) \in X_t \times \{0, ..., T-1\}$, $V : \mathbb{R}^n \times [0,T] \to \mathbb{R}$ *satisfies the GBE* (3.20), *and the state sequence* $\mathbf{x}^* = (x^*(0), ..., x^*(T))$ *and input sequence* $\mathbf{u}^* = (u^*(0), ..., u^*(T-1))$

satisfy

$$u^*(k) \in \arg \inf_{u \in \Gamma_{x^*(k),k}} \left\{ \phi_t(x^*(k), u, V(f(x^*(k), u, k), k+1)) \right\} \textit{ for } k \in \{0, ..., T-1\}. \tag{3.21}$$

$$x^*(0) = x_0, \quad x^*(k+1) = f(x^*(k), u^*(k), k) \textit{ for } k \in \{0, ..., T-1\}. \tag{3.22}$$

Then $(\mathbf{u}^*, \mathbf{x}^*)$ *solve the MSOP given in Eq.* (3.2)*, associated with the tuple* $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$.

Proof. Suppose $(\mathbf{u}^*, \mathbf{x}^*)$ satisfy Eqs. (3.21) and (3.22). It follows the pair $(\mathbf{u}^*,\mathbf{x}^*)$ is a feasible solution for the MSOP $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$ since Eq. (3.21) implies $u^*(k) \in \Gamma_{x^*(k),k}$, thus $u^*(k) \in U$ and, using Eq. (3.22), $x^*(k+1) = f(x^*(k), u^*(k), k) \in X_{k+1}$ for all $k \in \{0, ..., T-1\}$.

By Eq. (3.21) it follows that

$$\begin{aligned} &\inf_{u \in \Gamma_{x^*(k),k}} \left\{ \phi_k(x^*(k), u, V(f(x^*(k), u, k), k+1)) \right\} \\ &\qquad = \phi_k(x^*(k), u^*(k), V(f(x^*(k), u^*(k), k), k+1)) \text{ for all } k \in \{0, ..., T-1\}. \end{aligned} \tag{3.23}$$

We will now show Eq. (3.23) implies $(\mathbf{u}^*, \mathbf{x}^*)$ solve the MSOP.

$$\begin{aligned}
&\inf_{\mathbf{u}\in\Gamma_{x_0,[0,T-1]}} J(\mathbf{u},\mathbf{x}) = V(x_0,0)\\
&= \inf_{u\in\Gamma_{x^*(0),0}} \Big\{\phi_0(x^*(0),u,V(f(x^*(0),u,0),1))\Big\}\\
&= \phi_0(x^*(0),u^*(0),V(x^*(1),1))\\
&= \phi_0\bigg(x^*(0),u^*(0),\inf_{u\in\Gamma_{x^*(1),1}} \Big\{\phi_1(x^*(1),u,V(f(x^*(1),u,1),2))\Big\}\bigg)\\
&\qquad\qquad\vdots\\
&= \phi_0\bigg(x^*(0),u^*(0),...,\phi_k\bigg(x^*(k),u^*(k),\\
&\qquad\qquad \phi_{k+1}\bigg(x^*(k+1),u^*(k+1),....\phi_T(x^*(T))\bigg)...\bigg)...\bigg)\\
&= J(\mathbf{u}^*,\mathbf{x}^*),
\end{aligned}$$

where the first equality follows as it was shown in Prop. 3.1 that V is a value function of the MSOP, the second equality follows since V satisfies the GBE (3.20) and using $x^*(0) = x_0$, the third equality follows by Eq. (3.23), the fourth inequality follows again using the GBE (3.20), and the fifth inequality follows by recursively using the GBE (3.20) together with Eq. (3.23). Thus, if $(\mathbf{u}^*, \mathbf{x}^*)$ satisfies Eqs. (3.22) and (3.21) then $(\mathbf{u}^*, \mathbf{x}^*)$ solves the MSOP $\{J, f, \{X_t\}_{0\le t\le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$. □

Consider an MSOP associated with $\{J, f, \{X_t\}_{0\le t\le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$. As we will show next, if the representation maps $\{\phi_t\}_{t=0}^T$, associated with J are strictly monotonic (Eq. (3.24)) then Eqs. (3.21) and (3.22) of Prop. 3.2 become sufficient and necessary for optimality. In Sec. 3.2.3 we will give several examples of naturally monotonically backward functions with associated strictly monotonic representation maps.

Theorem 3.2 (Necessary and sufficient conditions for optimality)**.** *Consider an MSOP of Form* (3.2) *associated with* $\{J, f, \{X_t\}_{0\le t\le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$. *Suppose*

$\{\phi_t\}_{t=0}^T$ *are the representation maps of* $J : U^T \times \Pi_{t=0}^T X_t \to \mathbb{R}$ *(Defn. 3.3), and* $\Gamma_{x,t} \neq \emptyset$ *for all* $(x,t) \in X_t \times \{0,...,T-1\}$*. Furthermore, suppose the representation maps are strictly monotonic in their third argument. That is if* $z, w \in Image\{\phi_{t+1}\}$ *are such that* $z > w$ *then*

$$\phi_t(x,u,z) > \phi_t(x,u,w) \textit{ for all } (x,u) \in X_t \times U. \tag{3.24}$$

Then $(\mathbf{u}^*, \mathbf{x}^*)$ *solves the MSOP if and only if* $(\mathbf{u}^*, \mathbf{x}^*)$ *satisfies Eqs.* (3.21) *and* (3.22)*.*

Proof. If $(\mathbf{u}^*, \mathbf{x}^*)$ satisfies Eqs. (3.21) and (3.22) then Prop. 3.2 shows $(\mathbf{u}^*, \mathbf{x}^*)$ solves the MSOP associated with $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}_{Backward}^{Discrete}$.

Now assume the representation maps $\{\phi_t\}_{t=0}^T$ are strictly monotonic in their third argument (Eq. (3.24)) and $(\mathbf{u}^*, \mathbf{x}^*)$ solves the MSOP associated with the tuple $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}_{Backward}^{Discrete}$. As we have assumed $(\mathbf{u}^*, \mathbf{x}^*)$ is a solution to the MSOP, it follows that $(\mathbf{u}^*, \mathbf{x}^*)$ is feasible and thus Eq. (3.22) is trivially satisfied. To prove Eq. (3.21) is also satisfied let us suppose for contradiction the negation of Eq. (3.21), that there exists $k \in \{0,...,T-1\}$ such that

$$u^*(k) \notin \arg \inf_{u \in \Gamma_{x^*(k),k}} \left\{ \phi_t(x^*(k), u, V(f(x^*(k), u, k), k+1)) \right\},$$

where $V : \mathbb{R}^n \times [0,T] \to \mathbb{R}$ satisfies the GBE (3.20), and hence it follows

$$\begin{aligned} \inf_{u \in \Gamma_{x,t}} & \left\{ \phi_k(x^*(k), u, V(f(x^*(k), u, k), k+1)) \right\} \\ & < \phi_k(x^*(k), u^*(k), V(f(x^*(k), u^*(k), k), k+1)). \end{aligned} \tag{3.25}$$

Using Eq. (3.25) we have that,

$$\begin{aligned}
&J(\mathbf{u}^*, \mathbf{x}^*) = \inf_{\mathbf{u} \in \Gamma_{x_0,[0,T-1]}} J((u(0), .., u(T-1)), (x(0), ..., x(T))) \qquad (3.26)\\
&\leq \inf_{\mathbf{w} \in \Gamma_{x^*(k),[k,T-1]}} J((u^*(0), .., u^*(k-1), w(k), .., w(T-1)),\\
&\qquad\qquad\qquad\qquad (x^*(0), ..., x^*(k), z(k+1), ..., z(T)))\\
&= \phi_0\Bigg(x^*(0), u^*(0), ..., \inf_{w(k) \in \Gamma_{x^*(k),k}} \Bigg\{ \phi_k(x^*(k), w(k),\\
&\qquad \inf_{\mathbf{w} \in \Gamma_{f(x^*(k),w(k),k),[k+1,T-1]}} \phi_{k+1}(z(k+1), w(k+1), ...\phi_T(z(T))...) \Bigg\} ... \Bigg)\\
&= \phi_0\Bigg(x^*(0), u^*(0), ..., \inf_{w(k) \in \Gamma_{x^*(k),k}} \Bigg\{\phi_k(x^*(k), w(k), V(f(x^*(k), w(k), k), k+1))\Bigg\}, ...\Bigg)\\
&< \phi_0(x^*(0), u^*(0), ..., \phi_k(x^*(k), u^*(k), V(f(x^*(k), u^*(k), k), k+1)), ..,)\\
&= \phi_0\Bigg(x^*(0), u^*(0), ..., \phi_k\Bigg(x^*(k), u^*(k), \inf_{\mathbf{w} \in \Gamma_{f(x^*(k),w(k),k),[k+1,T-1]}}\\
&\qquad \Bigg\{ \phi_{k+1}(z(k+1), w(k+1), ...\phi_T(z(T)))...) \Bigg\} \Bigg) ... \Bigg)\\
&\leq \phi_0\Bigg(x^*(0), u^*(0), ..., \phi_k\Bigg(x^*(k), u^*(k),\\
&\qquad \phi_{k+1}\Bigg(x^*(k+1), u^*(k+1), ...\phi_T(x^*(T))\Bigg) ... \Bigg) ... \Bigg)\\
&= J(\mathbf{u}^*, \mathbf{x}^*).
\end{aligned}$$

Where the first equality in Eq. (3.26) follows as the pair $(\mathbf{u}^*, \mathbf{x}^*)$ is assumed to solve the MSOP. The first inequality in Eq. (3.26) follows by taking the infimum only over the input and state sequences from time stage $k+1$ onwards and fixing the first k input and state sequences as $(u^*(0), .., u^*(k-1))$ and $(x^*(0), ..., x^*(k))$ (which are known to be feasible as the pair $(\mathbf{u}^*, \mathbf{x}^*)$ is assumed to solve the MSOP). The second equality in Eq. (3.26) follows by Lemma 3.2. The third equality follows by Prop. 3.1 that shows V is the value function. The second inequality in Eq. (3.26) follows from Eq. (3.25) and using the assumed strict monotonic property of the representation maps (Eq. (3.24)). The fourth equality in Eq. (3.26) follows using Prop. 3.1, that shows V is the value function. The third inequality in Eq. (3.26) follows by fixing the decision variables of the infimum to $(u^*(k), ..., u^*(T-1))$ and $(x^*(k+1), ..., x^*(T))$ (which are known to be feasible as the pair $(\mathbf{u}^*, \mathbf{x}^*)$ is assumed to solve the MSOP) and using monotonic property of the representation maps (Eq. (3.5)).

Eq. (3.26) therefore provides a contradiction, that $J(\mathbf{u}^*, \mathbf{x}^*) < J(\mathbf{u}^*, \mathbf{x}^*)$; showing if $(\mathbf{u}^*, \mathbf{x}^*)$ solves the MSOP then Eqs. (3.22) and (3.21) must hold. □

In the next corollary we show that when the cost function, J, is additively separable, the GBE (3.20) reduces to Bellman's Equation (3.27); thus showing Bellman's Equation is an implication of the GBE. Therefore we have generalized the necessary and sufficient conditions for optimality, encapsulated in Bellman's Equation, to the GBE that provides such optimality conditions for a larger class of MSOPs with monotonically backward separable cost functions; that no longer need be additively separable.

Corollary 3.2 (Bellman's Equation)**.** *Consider an MSOP of Form* (3.2) *associated with* $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Addative}$. *Suppose the cost functions,* $\{c_t\}_{t=0}^T$, *associated with* $J : U^T \times \Pi_{t=0}^T X_t \to \mathbb{R}$ *(Defn. 3.2), are bounded over* $X_t \times U$. *Then if* $F : \mathbb{R}^n \times [0, T] \to \mathbb{R}$ *satisfies*

$$F(x,T) = c_T(x) \quad \text{for all } x \in X_T, \tag{3.27}$$
$$F(x,t) = \inf_{u \in \Gamma_{x,t}} \left\{ c_t(x,u) + F(f(x,u,t), t+1) \right\} \text{ for all } x \in X_t,\ t \in \{0,..,T-1\},$$

then F is a value function (Defn. 3.5) for the MSOP associated with the tuple $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Addative}$.

Moreover, if $\Gamma_{x,t} \neq \emptyset$ *for all* $(x,t) \in X_t \times \{0, ..., T\}$ *then* $\mathbf{x}^* = (x^*(0), ..., x^*(T))$

and $\mathbf{u}^* = (u^*(0), ..., u^*(T-1))$ *solve the MSOP if and only if the following is satisfied*

$$u^*(k) \in \arg \inf_{u \in \Gamma_{x^*(k),k}} \{c_k(x^*(k), u) + F(f(x^*(k), u, k), k+1)\}, \tag{3.28}$$

$$x^*(0) = x_0, \quad x^*(k+1) = f(x^*(k), u^*(k), k) \text{ for } k \in \{0, ..., T-1\}. \tag{3.29}$$

Proof. By Lemma 3.1 it follows J is naturally monotonically backward separable and can be written in Form (3.4) using the representation maps given in Eq. (3.8). Substituting the representation maps in Eq. (3.8) into the GBE (3.20), we obtain Bellman's Equation (3.27). Prop. 3.1 then shows F is a value function for the MSOP, associated with the tuple $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Addative}$.

Moreover as the representation maps in Eq. (3.8) are clearly strictly monotonic in their third argument (Eq. (3.24)) it follows by Theorem 3.2 that $(\mathbf{x}^*, \mathbf{u}^*)$ solve the MSOP if and only if $(\mathbf{x}^*, \mathbf{u}^*)$ satisfy Eqs. (3.28) and (3.29). □

3.2.3 Examples: Backward Separable Functions

In Subsection 3.2.2, we have shown that MSOPs with cost functions that are naturally monotonically backward separable (Defn. 3.3) can be solved efficiently using the GBE in Eq. (3.20). We now give examples of non-additively separable, yet monotonically backward separable cost functions which may be of significant interest. We note that this is not a complete list of all monotonically backward separable functions. Currently little is known about size and structure of the set of all monotonically backward separable functions.

The first function we consider is the point-wise maximum function. This function occurs in MSOPs when demand charges are present, see Chapter 4 and Jones and Peet (2021c), and in maximal invariant set estimation, see Xue and Zhan (2018).

Example 3.1 (Point wise maximum function)**.** *Suppose* $J : U^T \times \Pi_{i=0}^T X_t \to \mathbb{R}$ *is* *of the form*

$$J(\mathbf{u},\mathbf{x}) = \max\left\{ \max_{0\le k\le T-1}\{d_k(x(k),u(k))\}, d_T(x(T))\right\},$$

where $\mathbf{u} = (u(0),...,u(T-1))$, $\mathbf{x} = (x(0),...,x(T))$, $U \subset \mathbb{R}^m$ *and* $X_t \subset \mathbb{R}^n$ *are compact sets,* $d_k : X_k \times U \to \mathbb{R}$ *for* $0 \le k \le T-1$ *and* $d_T : X_T \to \mathbb{R}$. *Then* J *is a monotonically backward separable function. Moreover, if* $\{d_t\}_{t=0}^T$ *are bounded functions, then* J *is naturally monotonically backward separable.*

Proof. We can write J in Form (3.4) using the representation functions

$$\phi_T(x) = d_T(x), \quad \phi_i(x,u,z) = \max\{d_i(x,u), z\} \text{ for all } i \in \{0,..,T-1\}. \tag{3.30}$$

The monotonicity property in (3.5) follows since if $y \ge z$ then for all $i \in \{0,..,T-1\}$ we have that

$$\phi_i(x,u,y) = \max\{d_i(x,u), y\} \ge \max\{d_i(x,u), z\} = \phi_i(x,u,z),$$

where the above inequality follows by separately considering the cases $d_i(x,u) \ge y$ and $d_i(x,u) < y$.

Assuming $\{d_t\}_{t=0}^T$ are bounded functions the boundedness property, given in Eq. (3.7), is clearly satisfied by the representation maps given in Eq. (3.30) by induction on $i \in \{0,...,T-1\}$. The semi-continuity property (Eq. (3.6)) follow since the point-wise max function, ie $f(x) = \max_{1\le i\le n}\{x_i\}$, is Lipschitz continuous and hence upper semi-continuous. □

In the next example we consider multiplicative costs. A special case of this cost function, of the form $J(\mathbf{u},\mathbf{x}) = \mathbb{E}_{\mathbf{w}}[\exp(\sum_{t=0}^{T-1} c_t(x(t),u(t),w(t)) + c_T(x(T),w(t)))] := \int \exp(\sum_{t=0}^{T-1} c_t(x(t),u(t),w(t)) + c_T(x(T),w(t)))p(\mathbf{w})d\mathbf{w}$, where $p(\mathbf{w})$ is the probability density function of $\mathbf{w} = (w(0),...,w(T))$, has previously appeared in Jacobson (1973) and Glover and Doyle (1988).

Example 3.2 (Stochastic multiplicative functions)**.** *Suppose* $J : U^T \times \Pi_{i=0}^T X_t \to \mathbb{R}$ *is of the form*

$$J(\mathbf{u},\mathbf{x}) = \mathbb{E}_{\mathbf{w}}[c_T(x(T),w(T))\Pi_{t=0}^{T-1}c_t(x(t),u(t),w(t))]$$
$$:= \int_{I_0\times..I_T} c_T(x(T),w(T))\Pi_{t=0}^{T-1}c_t(x(t),u(t),w(t))$$
$$p_T(x(T),w(T))\Pi_{t=0}^{T-1}p_t(x(t),u(t),w(t))dw(0)...dw(T),$$

where $\mathbf{u} = (u(0),...,u(T-1))$*,* $\mathbf{x} = (x(0),...,x(T))$*,* $\mathbf{w} = (w(0),...,w(T))$*,* $U \subset \mathbb{R}^m$ *and* $X_t \subset \mathbb{R}^n$ *are compact sets,* $I_t \subset \mathbb{R}^k$*,* $c_t : X_t \times U \times I_t \to \mathbb{R}^+$ *for* $0 \le t \le T-1$*,* $c_T : X_T \times I_T \to \mathbb{R}$*, and* $p_t : X_t \times U \times I_t \to \mathbb{R}^+$*,* $p_T : X_T \times I_T \to \mathbb{R}$ *satisfy* $\int_{I_t} p_t(x,u,w)dw = 1$ *and* $\int_{I_T} p_T(x,w)dw = 1$ *for* $0 \le t \le T-1$*. Then* J *is a monotonically backward separable function. Moreover, if* $\{c_t\}_{t=0}^T$ *and* $\{p_t\}_{t=0}^T$ *are bounded functions, and sets* $\{I_t\}_{t=0}^T$ *have finite measure, then* J *is naturally monotonically backward separable. Furthermore, if* $\int_{I_i} p_i(x,u,w)c_i(x,u,w)dw \neq 0$ *for all* $(x,u,i) \in X_i \times U \times \{0,...,T-1\}$ *then the associated representation maps are strictly monotonic (Eq.* (3.24)*).*

Proof. We can write J in Form (3.4) using the representation functions

$$\phi_T(x) = \int_{I_T} c_T(x,w)p_T(x,w)dw, \tag{3.31}$$
$$\phi_i(x,u,z) = \int_{I_i} zp_i(x,u,w)c_i(x,u,w)dw \quad \text{for all } i \in \{0,..,T-1\}.$$

The monotonicity property (3.5) follows as $c_i(x,u,w) \ge 0$ and $p_i(x,u,w) \ge 0$ for all $(x,u,w) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^k$ and $i \in \{0,...,T-1\}$. Furthermore, if $\int_{I_i} p_i(x,u,w)c_i(x,u,w)dw \neq 0$ for all $(x,u,i) \in X_i \times U \times \{0,...,T-1\}$, then clearly the representation maps are strictly monotonic (Eq. (3.24)).

Assuming $\{c_t\}_{t=0}^T$ and $\{p_t\}_{t=0}^T$ are bounded functions, and sets $\{I_t\}_{t=0}^T$ have finite measure, the representation maps in Eq. (3.31) clearly satisfy the boundedness property (Eq. (3.7)) by induction on $i \in \{0,...,T-1\}$. For fixed $i \in \{0,..,T-1\}$ and

$(x,u) \in X_i \times U$ it follows that $\phi_i(x,u,z) = \lambda z$, where $\lambda \geq 0$ is some constant that depends on (x,u,i), is clearly upper semi continuous (as in Eq. (3.6)). □

In the next example we consider a function that can be interpreted as the expectation of cumulative additive costs, where at each time stage, $t \in \{0,...,T-1\}$, a cost $c_t(x(t),u(t))$ is added and there is an independent probability, $p_t(x(t),u(t)) \in [0,1]$, of stopping, incurring no further future costs. For a state and input trajectory, $(\mathbf{u},\mathbf{x}) \in \mathbb{R}^{m\times T} \times \mathbb{R}^{n\times(T+1)}$, let us denote the stopping time by $T(\mathbf{u},\mathbf{x})$; it then follows the distribution of this random variable is given as

$$\begin{aligned} &\mathbb{P}(T(\mathbf{u},\mathbf{x}) = T) = p_T(x(T))\Pi_{i=1}^{T-1}(1 - p_i(x(i),u(i))), \\ &\text{and } \mathbb{P}(T(\mathbf{u},\mathbf{x}) = t) = p_t(x(t),u(t))\Pi_{i=1}^{t-1}(1 - p_i(x(i),u(i))) \text{ for all } t \in \mathbb{N}, \end{aligned} \tag{3.32}$$

where we slightly abuse notation to write $\Pi_{i=1}^{-1}(1 - p_i(x(i),u(i))) = 1$ so $\mathbb{P}(T(\mathbf{u},\mathbf{x}) = 0) = p_0(x(0),u(0))$.

The stopped additive function is then given as

$$\begin{aligned} J(\mathbf{u},\mathbf{x}) = \mathbb{E}_{T(\mathbf{u},\mathbf{x})}\Bigg[&\sum_{t=0}^{\min\{T(\mathbf{u},\mathbf{x}),T-1\}} c_t(x(t),u(t)) \\ &+ \mathbb{1}_{\{(\mathbf{u},\mathbf{x}) \in U^T \times \Pi_{t=0}^T X_t : T(\mathbf{u},\mathbf{x}) = T\}}(\mathbf{u},\mathbf{x}) c_T(x(T))\Bigg]. \end{aligned} \tag{3.33}$$

To show the function in Eq. (3.33) is monotonically backward separable we will assume that the probability of the stopping time occurring inside the finite time horizon $\{0,...,T\}$ is one; this gives us the following “law of total probability“ equation: $\sum_{t=0}^T \mathbb{P}(T(\mathbf{u},\mathbf{x}) = t) = 1$ for all $(\mathbf{u},\mathbf{x}) \in \mathbb{R}^{m\times T} \times \mathbb{R}^{n\times(T+1)}$, which can be rewritten in terms of its probability density functions as,

$$\sum_{t=0}^{T-1} p_t(x(t),u(t))\Pi_{i=1}^{t-1}(1 - p_i(x(i),u(i))) + p_T(x(T))\Pi_{i=1}^{T-1}(1 - p_i(x(i),u(i))) \equiv 1. \tag{3.34}$$

Note, if $p_T(x(T)) \equiv 1$ then trivially (3.34) holds for any functions $p_i : \mathbb{R}^n \times \mathbb{R}^m \to [0,1]$.

Assuming Eq. (3.34) holds and using the law of total expectation, conditioning on the probability of each stopping time, it follows

$$\begin{aligned}
&J(\mathbf{u},\mathbf{x}) \qquad (3.35)\\
&= \mathbb{E}_{T(\mathbf{u},\mathbf{x})}\Bigg[\sum_{t=0}^{\min\{T(\mathbf{u},\mathbf{x}),T-1\}} c_t(x(t),u(t)) + \mathbb{1}_{\{(\mathbf{u},\mathbf{x})\in U^T\times\Pi_{t=0}^T X_t : T(\mathbf{u},\mathbf{x})=T\}}(\mathbf{u},\mathbf{x}) c_T(x(T))\Bigg]\\
&= \sum_{t=0}^{T-1}\Bigg(\sum_{s=0}^{t} c_s(x(s),u(s))\Bigg)\mathbb{P}(T(\mathbf{u},\mathbf{x})=t)\\
&\qquad\qquad + \Bigg(\sum_{s=0}^{T} c_s(x(s),u(s)) + c_T(x(T))\Bigg)\mathbb{P}(T(\mathbf{u},\mathbf{x})=T)\\
&= \sum_{t=0}^{T-1}\Bigg(\sum_{s=0}^{t} c_s(x(s)u(s))\Bigg)p_t(x(t),u(t))\Pi_{i=0}^{t-1}(1-p_i(x(i),u(i)))\\
&\qquad + \Bigg(\sum_{t=0}^{T-1} c_t(x(t),u(t)) + c_T(x(T))\Bigg)p_T(x(T))\Pi_{i=0}^{T-1}(1-p_i(x(i),u(i))).
\end{aligned}$$

We next state and prove that the function J given in Eq. (3.35) is monotonically backward separable.

Example 3.3 (Stochastically stopped additive cost)**.** *Suppose* $J : U^T \times \Pi_{i=0}^T X_t \to \mathbb{R}$ *is of the form*

$$\begin{aligned}
J(\mathbf{u},\mathbf{x}) = &\sum_{t=1}^{T-1}\Bigg(\sum_{s=0}^{t} c_s(x(s)u(s))\Bigg)p_t(x(t),u(t))\Pi_{i=0}^{t-1}(1-p_i(x(i),u(i))) \qquad (3.36)\\
&+ \Bigg(\sum_{t=0}^{T-1} c_t(x(t),u(t)) + c_T(x(T))\Bigg)p_T(x(T))\Pi_{i=0}^{T-1}(1-p_i(x(i),u(i))),
\end{aligned}$$

where $p_k : X_k \times U \to [0,1]$ *and* $p_T : X_T \to [0,1]$ *satisfy Eq.* (3.34)*,* $\mathbf{u} = (u(0),...,u(T-1))$*,* $\mathbf{x} = (x(0),...,x(T))$*,* $U \subset \mathbb{R}^m$ *and* $X_t \subset \mathbb{R}^n$*,* $c_k : X_k \times U \to \mathbb{R}$ *and* $c_T : X_T \to \mathbb{R}$*. Then* J *is a monotonically backward separable function. Moreover, if* $\{c_t\}_{t=0}^T$ *are bounded functions, then* J *is naturally monotonically backward separable. Furthermore, if* $p_i(x,u) \neq 1$ *for all* $(x,u,i) \in X_i \times U \times \{0,...,T-1\}$ *then the associated representation maps are strictly monotonic (Eq.* (3.24)*).*

Proof. Before writing J in the backward separable form, given in Equation (3.4), we first simplify J by switching the order of the double summation in Eq. (3.36). Let $T(\mathbf{u},\mathbf{x})$ be a random variable with distribution given in Eq. (3.32). As it is assumed $\{p_t\}_{0\leq t\leq T}$ satisfy Eq. (3.34) and each time-stage has independent probability of stopping it follows $\sum_{t=s}^{T}\mathbb{P}(T(\mathbf{u},\mathbf{x})=t)=\mathbb{P}(T(\mathbf{u},\mathbf{x})\geq s)=\mathbb{P}(\cap_{i=0}^{s-1}T(\mathbf{u},\mathbf{x})\neq s)=\Pi_{i=0}^{s-1}\mathbb{P}(T(\mathbf{u},\mathbf{x})\neq s)$. Now,

$$\begin{aligned}J(\mathbf{u},\mathbf{x}) &= \sum_{t=0}^{T-1}\Bigg(\sum_{s=0}^{T}c_s(x(s),u(s))\Bigg)\mathbb{P}(T(\mathbf{u},\mathbf{x})=t)\\ &\qquad\qquad+\Bigg(\sum_{s=0}^{t}c_s(x(s),u(s))+c_T(x(T))\Bigg)\mathbb{P}(T(\mathbf{u},\mathbf{x})=T)\\ &=\sum_{s=0}^{T-1}c_s(x(s),u(s))\mathbb{P}(T(\mathbf{u},\mathbf{x})\geq s)+c_T(x(T))\mathbb{P}(T(\mathbf{u},\mathbf{x})\geq T)\\ &=\sum_{s=0}^{T-1}c_s(x(s),u(s))\Pi_{i=0}^{s-1}\mathbb{P}(T(\mathbf{u},\mathbf{x})\neq i)+c_T(x(T))\Pi_{i=0}^{T-1}\mathbb{P}(T(\mathbf{u},\mathbf{x})\neq i)\\ &=\sum_{s=0}^{T-1}c_s(x(s),u(s))\Pi_{i=0}^{s-1}(1-p_i(x(i),u(i)))+c_T(x(T))p_T(x(T))\Pi_{i=0}^{T-1}(1-p_i(x(i),u(i))).\end{aligned}$$

It then follows J satisfies Eq. (3.4) using the representation maps

$$\begin{aligned}\phi_i(x,u,z)&=c_i(x,u)+z(1-p_i(x,u))\text{ for all } i\in\{0,..,T-1\},\\ \phi_T(x)&=c_T(x)p_T(x).\end{aligned}\tag{3.37}$$

The monotonicity property in Eq. (3.5) follows as $(1-p_i(x,u))\geq 0$ for all $(x,u)\in X_i\times U$ and $i\in\{0,...,T-1\}$. Strict monotonicity (Eq. (3.24)) trivially follows when $p_i(x,u)\neq 1$ for all $(x,u,i)\in X_i\times U\times\{0,...,T-1\}$.

Assuming $\{c_t\}_{t=0}^{T}$ are bounded functions the representation maps, given in Eq. (3.37), clearly satisfy the boundedness property (Eq. (3.7)) by induction on $i\in\{0,...,T-1\}$. For fixed $i\in\{0,..,T-1\}$ and $(x,u)\in X_i\times U$ it follows $\phi_i(x,u,z)=c_0+c_1z$, where $c_0,c_1\in\mathbb{R}$ are constants that depends on (x,u,i), clearly satisfies the upper semi continuity property (Eq. (3.6)). □

In the next example we introduce a function representing the number of time-steps a trajectory spends outside some target set. Later, in Section 3.5, we will use this function as the cost function for path planning problems.

Example 3.4 (Minimum time set entry function)**.** *Suppose* $J : U^T \times \Pi_{t=0}^T X_t \to \mathbb{R}$ *is of the form*

$$J(\mathbf{u},\mathbf{x}) = \min \left\{ \inf \left\{ t \in [0,T] : x(t) \in S \right\}, T \right\}, \tag{3.38}$$

where $\mathbf{u} = (u(0), ..., u(T-1))$*,* $u(t) \in \mathbb{R}^m$*,* $\mathbf{x} = (x(0), ..., x(T))$*,* $x(t) \in \mathbb{R}^n$*,* $U \subset \mathbb{R}^m$ *and* $X_t \subset \mathbb{R}^n$*, and* $S \subset \mathbb{R}^n$*. If the set* $\{t \in [0,T] : x(t) \in S\}$ *is empty, we define the infimum to be infinity. Then* J *is a naturally monotonically backward separable function.*

Proof. The function given in Eq. (3.38) is actually a special case of the function given in Eq. (3.36) with

$$p_T(x) \equiv 1, \quad p_t(x,u) = \mathbb{1}_S(x) \text{ for } t \in \{0, ..., T-1\},$$
$$c_T(x) \equiv T, \quad c_t(x,u) \equiv t.$$

Note, the functions $\{p_k\}_{0 \le k \le T}$ trivially satisfy Eq. (3.34) as $p_T(x) \equiv 1$. Moreover clearly $\{c_t\}_{t=0}^T$ are bounded functions. Therefore J is naturally monotonically backward separable by Example 3.3. □

3.3 The Principle of Optimality: A Necessary Condition for Monotonic Backward Separability

Given a function, $J : \mathbb{R}^{m \times T} \times \mathbb{R}^{n \times (T+1)} \to \mathbb{R}$, there is no obvious way to determine whether J is monotonically backward separable. Instead, in this section we will recall a necessary condition proposed in Bellman (1966), called the Principle of Optimality

(Defn. 3.6), that we show in Prop. 3.3 that all MSOPs with monotonically backward separable cost functions satisfy. Before recalling the definition of the Principle of Optimality it is useful to consider a **family of MSOPs** with different initializations, rather than a single MSOP (as we have been doing so in previous sections). Specifically, let us consider a family of MSOPs, associated with the sequence of tuples $\{J_{t_0}, f, \{X_t\}_{t_0 \le t \le T}, U, T\}_{t_0=0}^{T}$, each initialized at $(x_0, t_0) \in \mathbb{R}^n \times \{0, ..., T\}$, and of the form:

$$\begin{aligned}
(\mathbf{u}^*, \mathbf{x}^*) \in & \arg\min_{\mathbf{u},\mathbf{x}} J_{t_0}(\mathbf{u}, \mathbf{x}) \text{ subject to:} \qquad (3.39)\\
& x(t+1) = f(x(t), u(t), t) \text{ for } t = t_0, .., T-1\\
& x(t_0) = x_0, \;\; x(t) \in X_t \subset \mathbb{R}^n \text{ for } t = t_0, .., T\\
& u(t) \in U \subset \mathbb{R}^m \text{ for } t = t_0, .., T-1\\
& \mathbf{u} = (u(t_0), ..., u(T-1)) \text{ and } \mathbf{x} = (x(t_0), ..., x(T))
\end{aligned}$$

Definition 3.6. *We say the family of MSOPs of Form* (3.39)*, associated with the sequence of tuples* $\{J_{t_0}, f, \{X_t\}_{t_0 \le t \le T}, U, T\}_{t_0=0}^{T}$*, satisfies* ***the Principle of Optimality*** *at* $x_0 \in X_0$ *if the following holds. For any* $t \in \{0, ..., T\}$ *if* $\mathbf{u} = (u(0), ..., u(T-1))$ *and* $\mathbf{x} = (x(0), ..., x(T))$ *solve the MSOP initialized at* $(x_0, 0)$ *then* $\mathbf{v} = (u(t), ..., u(T-1))$ *and* $\mathbf{h} = (x(t), ..., x(T))$ *solve the MSOP initialized at* $(x(t), t)$*.*

Proposition 3.3. *Consider a family of MSOPs of Form* (3.39) *associated with the sequence of tuples* $\{J_t, f, \{X_t\}_{t \le s \le T}, U, T\}_{t=0}^{T}$*. Suppose either of the following are satisfied:*

(A) The MSOP initialized at $(x_0, 0)$ *has a unique solution and* J_t *is monotonically backward separable (Defn. 3.3) for* $t \in \{0, ..., T\}$*.*

(B) The function J_t *is monotonically backward separable (Defn. 3.3) for* $t \in \{0, ..., T\}$

with representation maps that are strictly monotonic, that is the representation maps satisfy Eq. (3.24)*.*

Then the family of MSOPs satisfies the Principle of Optimality at $x_0 \in X_0$ *(Defn. 3.6).*

Proof. First, let us deal with Case (A) by assuming the MSOP initialized at $(x_0, 0)$ has a unique solution and J_t is monotonically backward separable (Defn. 3.3) for $t \in \{0, ..., T\}$.

Since J_t is monotonically backward separable (Defn. 3.3) for $t \in \{0, ..., T\}$ there exists representation maps $\{\phi_t\}_{0 \le t \le T}$ such that

$$J_t(\mathbf{u}, \mathbf{x}) = \phi_t(x(t), u(t), \phi_{t+1}(x(t+1), u(t+1), \dots \phi_T(x(T)) \dots)).$$

Now, suppose $\mathbf{u}^* = (u(0), ..., u(T-1))$ and $\mathbf{x}^* = (x(0), ..., x(T))$ solve the MSOP initialized at $(x_0, 0)$ of Form (3.39) associated with the sequence of tuples $\{J_t, f, \{X_t\}_{t \le s \le T}, U, T\}_{t=0}^{T}$. Furthermore, suppose for contradiction that there exists some $t \ge 0$ such that $0 \le t < T$ and $\mathbf{v} = (u(t), ..., u(T-1))$ and $\mathbf{h} = (x(t), ..., x(T))$ do not solve MSOP initialized at $(x(t), t)$. We will show that this implies that the MSOP initialized at $(x_0, 0)$ does not have a unique solution, thus providing a contradiction and verifying the conditions of the Principle of Optimality. If $(\mathbf{v}, \mathbf{h})$ do not solve MSOP initialized at $(x(t), t)$, then there exist feasible $\mathbf{w} = (w(t), ..., w(T-1))$ and $\mathbf{z} = (z(t), ..., z(T))$ such that $J_t(\mathbf{w}, \mathbf{z}) < J_t(\mathbf{v}, \mathbf{h})$. i.e.

$$\begin{aligned} J_t(\mathbf{w}, \mathbf{z}) &= \phi_t(z(t), w(t), \phi_{t+1}(z(t+1), w(t+1), \dots \phi_T(z(T)) \dots)) \qquad (3.40) \\ &< \phi_t(x(t), u(t), \phi_{t+1}(x(t+1), u(t+1), \dots \phi_T(x(T)) \dots)) \\ &= J_t(\mathbf{v}, \mathbf{h}). \end{aligned}$$

Now, consider the proposed feasible sequences $\hat{\mathbf{u}} = (u(0), ..., u(t-1), w(t), ..., w(T-1))$ and $\hat{\mathbf{x}} = (x(0), ..., x(t-1), z(t), ..., z(T-1))$. It follows using the monotonicity property (Eq. (3.5)) of monotonically backward separable functions and Inequality (3.40), that

$$\begin{aligned}
J_0(\hat{\mathbf{u}},\hat{\mathbf{x}}) &= \phi_0(x(0),u(0),\phi_1(x(1),u(1),\dots\phi_t(z(t),w(t)\dots\phi_T(z(T))\dots))\dots) \quad (3.41)\\
&= \phi_0(x(0),u(0),\dots\phi_{t-1}(x(t-1),u(t-1),J_t(\mathbf{w},\mathbf{z}))\dots)\\
&\leq \phi_0(x(0),u(0),\dots\phi_{t-1}(x(t-1),u(t-1),J_t(\mathbf{v},\mathbf{h}))\dots)\\
&= J_0(\mathbf{u}^*,\mathbf{x}^*),
\end{aligned}$$

which shows $(\hat{\mathbf{u}},\hat{\mathbf{x}})$ is also an optimal solution, contradicting that $(\mathbf{u}^*,\mathbf{x}^*)$ is the unique solution of the MSOP at $(x_0,0)$.

We next deal with Case (B) by assuming J_t is monotonically backward separable (Defn. 3.3) for $t \in \{0,...,T\}$ with representation maps that are strictly monotonic (Eq. (3.24)).

Suppose $\mathbf{u}^* = (u(0),...,u(T-1))$ and $\mathbf{x}^* = (x(0),...,x(T))$ solve the MSOP initialized at $(x_0,0)$ of Form (3.39). By following the same argument as in Case (A), considering the same $(\hat{\mathbf{u}},\hat{\mathbf{x}})$, we get that Eq. (3.41) holds. However, since the representation maps $\{\phi_t\}_{t=0}^T$ are assumed to be strictly monotonic (Eq. (3.24)) we have that the inequality in Eq. (3.41) holds strictly. That is $J_0(\hat{\mathbf{u}},\hat{\mathbf{x}}) < J_0(\mathbf{u}^*,\mathbf{x}^*)$, contradicting the fact $(\mathbf{u}^*,\mathbf{x}^*)$ is the optimal solution.

□

Prop. 3.3 shows the Principle of Optimality (Defn. 3.6) is a necessary condition that all families of MSOPs with unique solutions and monotonically backward separable cost functions must satisfy. We now conjecture a necessary and sufficient condition. The following notation is used in this conjecture. Given J_t, $\{X_t\}_{0 \le t \le T}$ and U let us denote the set $\mathcal{F}$, where

$(f,x_0) \in \mathcal{F}$ if $x_0 \in X_0$ and the MSOP associated with$\{J_0, f, \{X_t\}_{0 \le t \le T}, U, T\}$

initialized at $(x_0,0)$ has a unique solution. (3.42)

Conjecture 3.1. *Consider* $\{X_t\}_{0 \le t \le T} \subset \mathbb{R}^{n \times T}$, $U \subset \mathbb{R}^m$ *and* $J_t : U^{T-t} \times \Pi_{s=t}^T X_s \to \mathbb{R}$. *Then, for any* $(f, x_0) \in \mathcal{F}$ *(where* $\mathcal{F}$ *is as in Eq.* (3.42)*) the family of MSOPs associated with the sequence of tuples* $\{J_t, f, \{X_t\}_{t \le s \le T}, U, T\}_{t=0}^T$ *satisfies the Principle of Optimality (Defn. 3.6) at* $x_0 \in X_0$ *if and only if* J_t *is monotonically backward separable (Defn. 3.3).*

Regardless of whether Conjecture 3.1 is true, Prop. 3.3 is useful. Prop. 3.3 provides a way of proving a function $J_t : U^{T-t} \times \Pi_{s=t}^T X_s \to \mathbb{R}$ is not monotonically backward separable. Rather than showing J_t does not satisfy Defn. 3.3 for every family of representation maps $\{\phi_s\}_{s=t}^T$, for which there are an uncountably many, we find any f for which the family of MSOP's associated with $\{J_t, f, \{X_s\}_{t \le s \le T}, U, T\}_{t=0}^T$ has a unique solution for some initialization $(x_0, 0)$ and does not satisfy the Principle of Optimality. Then Prop. 3.3 shows J_t is not monotonically backward separable. We demonstrate this proof strategy in the following lemma.

Lemma 3.3. *The function* $J_t : \mathbb{R}^{m \times (T-t)} \times \mathbb{R}^{n \times (T+1-t)} \to \mathbb{R}$, *defined as*

$$J_t(\mathbf{u}, \mathbf{x}) := \max_{t \le s \le T} d(x(s)) + \sum_{s=t}^{T-1} c_s(x(s), u(s)), \tag{3.43}$$

is not monotonically backward separable (Defn. 3.3) for all functions $c_k : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ *and* $d_k : \mathbb{R}^n \to \mathbb{R}$.

Proof. Let $T = 3$, $n = 1$ and $m = 1$. Consider the cost functions $c_0(x, u) = -u$, $c_1(x, u) = u$, $c_2(x, u) = -u/2$, and $d(x) = x$. Consider the dynamics $f(x, u, t) = x + u$ and constraints $X_t = [0, h]$ and $U = \{-h, 0, h\}$, where $h > 0$. Let us consider the MSOP of Form (3.39) associated with $\{J_0, f, \{X_t\}_{0 \le t \le 3}, U, 3\}$ and initialized at

Table 3.1: This table shows the corresponding cost of each feasible input sequence of the MSOP given in Eq. (3.44) found in Lemma 3.3.

feasible $\mathbf{u}$	objective value	feasible $\mathbf{u}$	objective value
$(0, 0, 0)$	0	$(h, 0, -h)$	h/2
$(0, 0, h)$	h/2	$(h, 0, 0)$	0
$(0, h, 0)$	2h	$(h, -h, 0)$	-h
$(0, h, -h)$	(5/2)h	$(h, -h, h)$	-(3/2)h

$(x_0, t_0) = (0, 0)$:

$$\min_{\mathbf{u},\mathbf{x}} \quad \max_{t_0 \le k \le 3} x(k) + \sum_{t=\min\{t_0,2\}}^{2} c_t(x(t), u(t)) \tag{3.44}$$

$$\text{subject to: } x(t+1) = x(t) + u(t) \quad \text{for all } t \in \{t_0, ..., 3\},$$

$$x(t_0) = x_0,\ 0 \le x(t) \le h,\ u(t) \in \{-h, 0, h\},$$

It can be shown there are $3^3 = 27$ input sequences in $\{-h, 0, h\}^3$, only 8 of which are feasible to the MSOP in Eq. (3.44) initialized at $(x_0, t_0) = (0, 0)$. Table 3.1 presents each feasible input sequence with associated cost. We deduce the unique optimal input sequence is $\mathbf{u} = (h, -h, h)$, yielding a unique optimal trajectory of $\mathbf{x} = (0, h, 0, h)$. Following the input sequence to $t = 2$ we examine the MSOP (3.44) initialized at $(x_0, t_0) = (0, 2)$. For the MSOP initialized at $(x_0, t_0) = (0, 2)$ there are only two feasible inputs: $u(2) = 0$ or $u(2) = h$. Of these, the first is optimal (cost of 0 vs $h/2$). Thus although $\mathbf{u} = (h, -h, h)$ and $\mathbf{x} = (0, h, 0, h)$ solve the MSOP initialized at $(x_0, t_0) = (0, 0)$, $\mathbf{v} = (h)$ and $\mathbf{h} = (0, h)$ do not solve the MSOP initialized at $(x_0, t_0) = (0, 2)$. We conclude the family of MSOPs associated with $\{J_t, f, \{X_s\}_{t \le s \le 3}, U, 3\}_{t=0}^3$ does not satisfy the Principle of Optimality at $x_0 = 0$, although the MSOP initialized at $(x_0, t_0) = (0, 0)$ does have a unique solution. Therefore by Prop. 3.3 the function J_t is not monotonically backward separable. $\square$

Remark 3.1. *The function given in Eq.* (3.43) *can clearly be expressed as the addition of two monotonically backward separable functions,* $J_1(\mathbf{u},\mathbf{x}) = \sum_{s=t}^{T-1} c_s(u(s))$ *(Lemma 3.1) and* $J_2(\mathbf{u},\mathbf{x}) = \max_{t\leq s\leq T} d(x(s))$ *(Example 3.1). Therefore, Lemma 3.3 shows that the property of monotonically backward separability is not preserved under addition.*

3.4 Numerically Solving MSOPs with Backward Separable Costs

Before proceeding with applications (path planning in Section 3.5 and invariant set estimation in Section 3.6) we address the numerical implementation of approximately solving an MSOP with monotonically backward separable cost function, using the necessary and sufficient conditions for optimality given in Theorem 3.2.

For implementation, we use a "discretization"/"look up tables" approximation scheme that maps our class of MSOPs to a much simpler class of MSOPs with finite state and control spaces. For MSOPs with finite input and state spaces the GBE (3.20) can be solved by enumeration. Similar numerical schemes with convergence proofs can be found in Jones and Peet (2018) and Dufour and Prieto-Rumeau (2012).

Unfortunately, it is well known that discretization techniques that solve MSOPs suffer from the "curse of dimensionality"; where the number of grid points required to uniformly sample a set grows exponentially with respect to the dimension of the set. Therefore, the techniques presented in this section may not scale well with respect to the input and state space dimension. We do note however, there is scope to improve the scalability of discretization methods since such discretization schemes are known to be parallelizable, see Maidens *et al.* (2016). Alternatively, we show in Section 4.5 that rather than solving the GBE (3.20) at each grid point, as is the case with discretization schemes, it is possible to use an Approximate Dynamic

Programming (ADP)/Reinforcement Learning (RL) algorithm to heuristically solve the MSOP. Such numerical schemes are shown to have lower computational times when compared to methods that solve the GBE (3.20) exactly at each grid points. This demonstrates that MSOPs with monotonically backward separable cost functions can be heuristically solved using the same methods developed in the ADP literature with the aid of the GBE (3.20). We do note however that ADP/RL algorithms typically do not have theoretical performance bounds, that is given an MSOP there is no guarantee that the feasible solution obtained using an ADP/RL algorithm is "close" to the optimal solution. For this reason we prefer to consider discretization based methods for solving MSOPs that have state and input spaces with a relatively small dimension.

3.4.1 Discretization: A Map onto MSOPs with Finite Input and State Spaces

Consider an MSOP of Form (3.2) associated with the tuple $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backwards}$ where the state space and input space are of the form $X_t = [\underline{x}, \bar{x}]^n$ and $U = [\underline{u}, \bar{u}]^m$ respectively. For MSOPs of this form it is not generally possible to analytically solve GBE (3.20). We thus need to consider a sequence of "close" MSOPs with finite state and control spaces, for which the GBE (3.20) can be solved by enumeration. We define such MSOPs with finite state and control spaces next.

Definition 3.7. *Given $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backwards}$ we say $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Finite}$ if the cardinality of the sets $\{X_t\}_{0 \le t \le T}$ and U are finite.*

Given an MSOP, associated with the tuple $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backwards}$, and some $k \in \mathbb{N}$ we now consider a "discretized"

MSOP, associated with a tuple in $\mathcal{M}_{Finite}^{Discrete}$, initialized at $x_0 \in X_0$,

$$\begin{aligned}
&\min_{\mathbf{u},\mathbf{x}} \phi_0(x(0), u(0), \phi_1(x(1), u(1), \ldots \phi_T(x(T)) \ldots)) \qquad (3.45)\\
&\text{subject to:}\\
&x(t+1) = \arg\min_{y \in X_k} \{||y - f(x(t), u(t), t)||_2\},\\
&x(t_0) = x_0,\ x(t) \in X_k \subset \mathbb{R}^n,\ u(t) \in U_k \subset \mathbb{R}^m \text{ for } t = 0, .., T,\\
&\mathbf{u} = (u(0), ..., u(T-1)) \text{ and } \mathbf{x} = (x(0), ..., x(T)),
\end{aligned}$$

where $X_k = \{x_1, ..., x_k\}^n$ such that $\underline{x} = x_1 < x_2 < ... < x_k = \bar{x}$ and $||x_{i+1} - x_i||_2 = \frac{\bar{x} - \underline{x}}{k}$ for $1 \le i \le k-1$, and $U_k = \{u_1, ..., u_k\}^m$ such that $\underline{u} = u_1 < u_2 < ... < u_k = \bar{u}$ and $||u_{i+1} - u_i||_2 = \frac{\bar{u} - \underline{u}}{k}$ for $1 \le i \le k-1$.

We now consider the map that sends MSOPs with tuples in $\mathcal{M}_{Backwards}^{Discrete}$ to MSOPs of Form (3.45) with tuples in $\mathcal{M}_{Finite}^{Discrete}$. Specifically, we denote the approximation map $\chi : \mathcal{M}_{Backwards}^{Discrete} \times \mathbb{N} \to \mathcal{M}_{Finite}^{Discrete}$ which is defined for $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}_{Backwards}^{Discrete}$, where $X_t = [\underline{x}, \bar{x}]^n$ and $U = [\underline{u}, \bar{u}]^m$, by

$$\chi(J, f, \{X_t\}_{0 \le t \le T}, U, T\}, k) := \{J, \tilde{f}, \{\tilde{X}\}_{0 \le t \le T}, \tilde{U}, T\}, \qquad (3.46)$$

where $\tilde{X} = \{x_1, ..., x_k\}^n$ such that $\underline{x} = x_1 < x_2 < ... < x_k = \bar{x}$ and $||x_{i+1} - x_i||_2 = \frac{\bar{x} - \underline{x}}{k}$ for $1 \le i \le k-1$, $\tilde{U}_k = \{u_1, ..., u_k\}^m$ such that $\underline{u} = u_1 < u_2 < ... < u_k = \bar{u}$ and $||u_{i+1} - u_i||_2 = \frac{\bar{u} - \underline{u}}{k}$ for $1 \le i \le k-1$, and $\tilde{f}(x, u, t) = \arg\min_{y \in X_k} \{||y - f(x, u, t)||_2\}$.

Now given an MSOP in $\mathcal{M}_{Backwards}^{Discrete}$ and a discretization level $k \in \mathbb{N}$, we use the approximation map, χ, to map this MSOP to a "discretized" MSOP of class $\mathcal{M}_{Finite}^{Discrete}$. For MSOPs of class $\mathcal{M}_{Finite}^{Discrete}$ the GBE (3.20) can be solved by enumeration yielding an analytical solution to the "discretized" MSOP. In the following sub-sections we show how a feasible input sequence to the original MSOP can then be constructed from the solution to the "discretized" MSOP.

3.4.2 Constructing a Feasible Solution to an MSOP using Discretization

Consider an MSOP associated with $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backwards}$. For $k \in \mathbb{N}$ consider the discretized MSOP $\mathcal{X}(J, f, \{X_t\}_{0 \le t \le T}, U, T\}, k) :=$ $\{J, \tilde{f}, \{\tilde{X}\}_{0 \le t \le T}, \tilde{U}, T\}$. By iteratively solving the GBE (3.20) we can find an optimal solution to the "discretized" MSOP associated with $\{J, \tilde{f}, \{\tilde{X}\}_{0 \le t \le T}, \tilde{U}, T\}$, which we denote as $(\mathbf{v}_k, \mathbf{h}_k)$. Note that $(\mathbf{v}_k, \mathbf{h}_k)$ does not necessary solve, or is necessary feasible to the original MSOP, $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backwards}$. However, $(\mathbf{v}_k, \mathbf{h}_k)$ can be used to construct a feasible solution for MSOP associated with $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backwards}$ initialized at $x_0 \in X_0$ in the following way,

$$u_k(t) := \arg \min_{u \in \Gamma_{t,x}} ||v_k(t) - u||_2 \text{ for all } t \in \{0, ..., T-1\}, \tag{3.47}$$

$$x_k(t+1) := f(x_k(t), u_k(t)) \text{ for all } t \in \{0, ..., T-1\} \text{ and } x_k(0) = x_0,$$

where we recall $\Gamma_{t,x}$ is the set of feasible inputs such that if $u \in \Gamma_{t,x}$ then $u \in U$ and $f(x, u, t) \in X_t$. Then, $(\mathbf{u}_k, \mathbf{x}_k)$, where $\mathbf{u}_k = (u_k(0), ..., u_k(T-1))$ and $\mathbf{x}_k = (x_k(0), ..., x_k(T))$, is feasible for the MSOP associated with $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backwards}$.

3.4.3 Convergence of Feasible Solutions Constructed using Discretization

Consider an MSOP associated with $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backwards}$. Suppose $(\mathbf{u}_k, \mathbf{x}_k)$ is as in Eq. (3.47). It follows that $(\mathbf{u}_k, \mathbf{x}_k)$ is a feasible solution to the MSOP associated with $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backwards}$. In the case J is additively separable (Defn. 3.2) it was shown in Theorem 2 from Jones and Peet (2018) that if $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backwards}$ satisfies certain continuity assumptions then

$$\lim_{k \to \infty} |J(\mathbf{u}_k, \mathbf{x}_k) - J^*| = 0, \tag{3.48}$$

where J^* is the optimal value of the objective function of the MSOP associated with $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backwards}$. Furthermore, it can be shown that Eq. (3.48) holds in the more general case when J is naturally monotonically backward separable (Defn. 3.3) using a similar argument to Jones and Peet (2018).

3.5 Application: Path Planning and Obstacle Avoidance

In this section we design a full state feedback controller (Markov Policy) for a discrete time dynamical system with the objective of reaching a target set in minimum time while avoiding moving obstacles. In order to do this we solve the GBE (3.20) using discretization schemes outlined in Section 3.4.

3.5.1 Path Planning MSOPs

We say $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\}$ is a path planning MSOP, or $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Path}$, if

- $J(\mathbf{u}, \mathbf{x}) = \min \Big\{ \inf \Big\{ t \in [0,T] : x(t) \in S \Big\}, T \Big\}$.
- $S = \{x \in \mathbb{R}^n : g(x) < 0\}$, where $g : \mathbb{R}^n \to \mathbb{R}$.
- $X_t = \mathbb{R}^n / (\cup_{i=1}^N O_{t,i})$, where $O_{t,i} = \{x \in \mathbb{R}^n : h_{t,i}(x) < 0\}$ and $h_{t,i} : \mathbb{R}^n \to \mathbb{R}$.
- There exits a feasible solution, $(\mathbf{u}, \mathbf{x})$, to the MSOP (3.2) associated with the tuple $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\}$ such that $x(k) \in S$ for some $k \in \{0, ..., T\}$.

Clearly, solving an the MSOP (3.2) associated with a path planning problem tuple, $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Path}$, is equivalent to finding the input sequence that drives a discrete time system, governed by the vector field f, to a target set S in minimum time while avoiding the moving obstacles, represented as sets $O_{t,i} \subset \mathbb{R}^n$. Moreover, as shown in Example 3.4, the function $J(\mathbf{u}, \mathbf{x}) = \min \Big\{ \inf \Big\{ t \in [0,T] :$

$x(t) \in S \Big\}, T \Big\}$ is a naturally monotonically backward separable function (Defn. 3.3), and hence $\mathcal{M}_{Path}^{Discrete} \subset \mathcal{M}_{Backward}^{Discrete}$.

3.5.2 Path Planning for Dubin's Car

We now numerically solve a path planning MSOP $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}_{Path}^{Discrete}$ with dynamics as defined in Maidens *et al.* (2018); also known as the Dubin's car dynamics, which is given as

$$f\left(x, u, t\right) = \Big[x_1 + v\cos(x_3), x_2 + v\sin(x_3), x_3 + \frac{v}{L}\tan(u)\Big]^T, \tag{3.49}$$

where $(x_1, x_2) \in \mathbb{R}^2$ is the position of the car, $x_3 \in \mathbb{R}$ denotes the angle the car is pointing, $u \in \mathbb{R}$ is the steering angle input, $v \in \mathbb{R}$ is the fixed speed of the car, and L is a parameter that determines the turning radius of the car.

We solve the path planning MSOP using the discretization scheme from Section 3.4. The target set, obstacles, state space, and input constraint sets are given by

$$\begin{aligned}
S &= \{(x_1, x_2) \in \mathbb{R}^2 : -0.25 < x_1 - 0.75 < 0.25, -0.25 < x_2 + 0.75 < 0.25\},\\
O_{t,i} &= \{(x_1, x_2) \in \mathbb{R}^2 : (x_1 - X_i)^2 + (x_2 - Y_i)^2 - R_i^2 < 0\}\\
&\qquad\qquad\qquad\qquad \text{for all } i \in \{1, ..., 15\} \text{ and } t \in \{0, ..., T\},\\
X_t &= [-1, 1]^2 \times \mathbb{R} \quad \text{ for all } t \in \{0, ..., T\}, \quad U = [-1, 1],
\end{aligned}$$

where $X, Y, R \in \mathbb{R}^{15}$ are randomly generated vectors. The parameters of the system are set to $v = 0.1$ and $L = 1/6$.

Figure 3.1 shows three approximately optimal state sequences starting from different initial conditions. These state sequences are found by numerically solving the GBE, Equation (3.20), where $\{\phi_t\}_{t=0}^T$ are as in Example 3.4. To numerically solve the GBE the state space, $X_t \subset \mathbb{R}^3$, is discretized as a $60 \times 60 \times 60$-grid

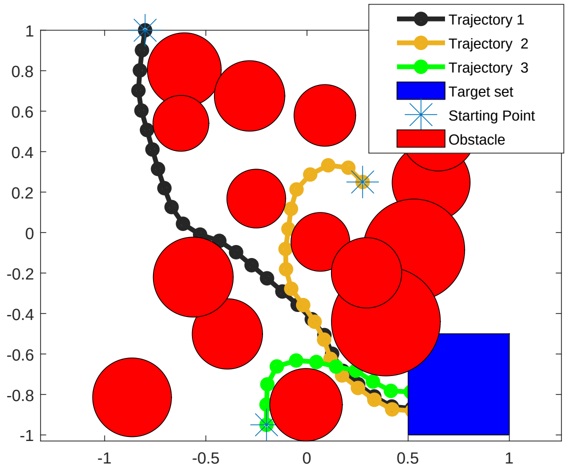

Figure 3.1: Graph showing approximate optimal trajectories, shown as the gold, black and green curves, with dynamics given in Eq. (3.49) and the goal of reaching the target set, shown as the blue square, while avoiding obstacles, shown as red circles.

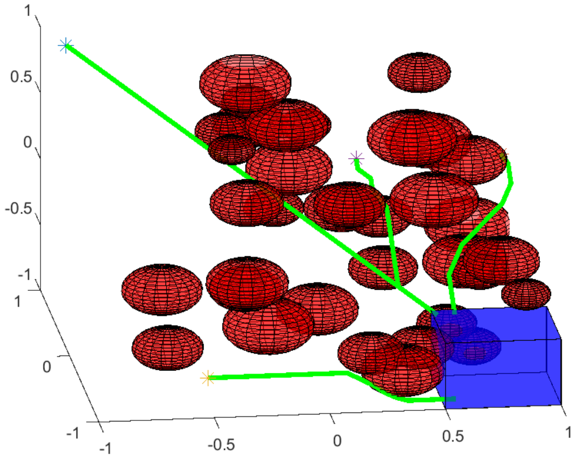

Figure 3.2: Graph showing approximate optimal trajectories, shown as the green curves, with dynamics given in Eq. (3.50) and the goal of reaching the target set, shown as the blue cube, while avoiding obstacles, shown as red spheres.

between $[-1,1]^2 \times [0, 2\pi]$ and the input space, $U \subset \mathbb{R}$, is discretized as 100 grid points within $[-1,1]$. The first state sequence was chosen to have initial condition $[-0.8, 1, -0.55\pi]^T \in \mathbb{R}^3$ (the furthest of the three trajectories from the target) and took 25 steps to reach its goal. The second state sequence was chosen to have initial condition $[0.275, 0.25, 0.75\pi]^T \in \mathbb{R}^3$; in this case as $x_3(0) = 0.75\pi$ Dunbin's car initially is directed towards the top left corner. The input sequence successfully turns the car downwards between two obstacles and into the target set, taking 18 steps. The third trajectory was chosen to have initial condition $[-0.2, 0.95, 0.5\pi]^T \in \mathbb{R}^3$-starting very closely to an obstacle facing upwards. This trajectory had to use the full turning radius of the car to navigate around the obstacle towards the target set and took 10 steps.

3.5.3 Path Planning in 3D

We now numerically solve a path planning MSOP $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}_{Path}^{Discrete}$ problem with dynamics given by

$$f(x,u,t) = [x_1 + u_1, x_2 + u_2, x_3 + u_3]^T. \tag{3.50}$$

The target set, obstacles, state space and input constraint set were respectively are given by

$$\begin{aligned} S = \{&(x_1, x_2, x_3) \in \mathbb{R}^2 : -0.25 < x_1 - 0.75 < 0.25, \\ &\qquad - 0.25 < x_2 + 0.75 < 0.25, -0.25 < x_2 + 0.75 < 0.25\} \\ O_{t,i} = \{&(x_1, x_2, x_3) \in \mathbb{R}^3 : (x_1 - A_i - \alpha_i t)^2 + (x_2 - B_i - \beta_i t)^2 \\ &\qquad + (x_2 - C_i - \gamma_i t)^2 - R_i^2 < 0\} \text{ for all } i \in \{1, ..., 35\}, t \in \{0, ..., T\} \\ X_t - &[-1,1]^3 \text{ for all } t \in \{0, ..., T\}, \qquad U = [-0.05, 0.05]^3, \end{aligned}$$

where $A, B, C, \alpha, \beta, \gamma, R \in \mathbb{R}^{35}$ are randomly generated vectors. Note, when α, β, γ are non-zero the center of the spherical obstacles moves with time. For presentation purposes in this chapter we consider stationary obstacles, selecting $\alpha = \beta = \gamma = 0$, however, a downloadable .gif file showing the numerical solution for moving obstacles can be found at Jones and Peet (2020).

using the discretization scheme from Section 3.4 by computing the solution at each grid point to

This path planning MSOP can be numerically solved by computing the solution to the GBE (3.20), where $\{\phi_t\}_{t=0}^T$ are as in Example 3.4. To numerically solve the GBE we use the discretization scheme from Section 3.4, discretizing the state and input space, $X_t \subset \mathbb{R}$ and $U \subset \mathbb{R}^3$, as a $40 \times 40 \times 40$ uniform grid on $[-1, 1]^3$ and a $5 \times 5 \times 5$ uniform grid on $[-0.05, 0.05]^3$ respectively. Figure 3.2 shows four optimal state sequences, shown as green lines, starting from various initial conditions. All trajectories successfully avoid the obstacles, represented as red spheres, and reach the target set, shown as a blue cube.

GPU Implementation All DP methods involving discretization fall prey to the curse of dimensionality, where the number of points required to sample a space increases exponentially with respect to the dimension of the space. For this reason solving MSOP's in dimensions greater than three can be computationally challenging. Fortunately, our discretization approach to solving the GBE (Equation (3.20)), can be parallelized at each time-step. To improve the scalability of the proposed approach, we have therefore constructed in Matlab a GPU accelerated DP algorithm for solving the 3D path planning problem. This code is available for download at Code Ocean, see Jones and Peet (2019a).

3.6 Application: Maximal Invariant Sets

The Finite Time Horizon Maximal Invariant Set (FTHMIS) is the largest set of initial conditions such that there exists an input sequence that produces a feasible state sequence over a finite time period. Computation of the maximal robust invariant sets over infinite time horizons was considered in Xue and Zhan (2018). Before we define the FTHMIS we introduce some notation.

For $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{N} \to \mathbb{R}^n$ we say the map $\rho_f : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^{m \times (T-1)} \to \mathbb{R}^n$ is the solution map associated with f if for any $T > 0$ the following holds for all $t \in \{0, ..., T\}$

$$\rho_f(x_0, t, \mathbf{u}) = x(t), \tag{3.51}$$

where $\mathbf{u} = (u(0), ..., u(T-1))$, $x(k+1) = f(x(k), u(k), k)$ for all $k \in \{0, .., k-1\}$, and $x(0) = x_0$.

Definition 3.8. *For $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{N} \to \mathbb{R}^n$, $X_t \subseteq \mathbb{R}^n$, $U \subset \mathbb{R}^m$, $T \in \mathbb{N}$, and $\mathcal{A}_t \subseteq \mathbb{R}^n$ we define the Finite Time Horizon Maximal Invariant Set (FTHMIS), denoted by $\mathcal{R}$, by*

$$\mathcal{R} := \Bigg\{ x_0 \in \mathbb{R}^n : \text{ there exists } \mathbf{u} \in \Gamma_{x_0,[0,T-1]} \text{ such that } \rho_f(x_0, t, \mathbf{u}) \in \mathcal{A}_t \\ \text{for all } t \in \{0, ..., T\} \Bigg\},$$

where the notation $\Gamma_{x_0,[0,T-1]}$ is as in Eq. (3.12).

We next show that the sublevel set of the value function (Defn. 3.5) associated with a certain MSOP, $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$, can completely characterize the FTHMIS (Defn. 3.8).

Theorem 3.3. *Consider the sets $\mathcal{A}_t = \{x \in \mathbb{R}^n : g_t(x) < 0\}$, where $g_t : \mathbb{R}^n \to \mathbb{R}$. Suppose V is a value function (Defn. 3.5) associated with the MSOP, defined by the* *tuple* $\{J, f, \{X_t\}_{0\le t\le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$*, where* $J(\mathbf{u}, \mathbf{x}) = \max_{0\le k\le T} g_k(x(k))$*. Then*

$$\mathcal{R} = \{x \in \mathbb{R}^n : V(x, 0) < 0\}, \tag{3.52}$$

where the set $\mathcal{R} \subset \mathbb{R}^n$ *is the FTHMIS as in Definition 3.8.*

Proof. The function $J(\mathbf{u}, \mathbf{x}) = \max_{0\le k\le T} g_k(x(k))$ is monotonically backward separable as shown in Example 3.1 using representation maps given by

$$\begin{aligned} \phi_i(x, u, z) &= \max\{g_i(x), z\} \text{ for all } i \in \{0, .., T-1\} \\ \phi_T(x) &= g_T(x). \end{aligned}$$

Therefore by Definition 3.5 any value function, $V : \mathbb{R}^n \to \mathbb{R}$, associated with $\{J, f, \{X_t\}_{0\le t\le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$ satisfies

$$V(x, T) = g_T(x) \text{ for all } x \in X_T, \tag{3.53}$$

and for all $t \in \{0, 1, .., T-1\}$ and $x \in X_t$

$$V(x, t) = \inf_{\mathbf{u}\in\Gamma_{x,[0,T-1]}} \max_{t\le k\le T} g_k(\rho_f(x, k, \mathbf{u})). \tag{3.54}$$

We will first show that $\mathcal{R} \subseteq \{x \in \mathbb{R}^n : V(x, 0) < 0\}$. Let $x_0 \in \mathcal{R}$ then by Definition 3.8 there exists $\mathbf{u}_0 \in \Gamma_{x_0,[0,T-1]}$ such that

$$\rho_f(x_0, t, \mathbf{u}_0) \in \mathcal{A}_t \text{ for all } t \in \{0, ..., T\}.$$

Since $\mathcal{A}_t = \{x \in \mathbb{R}^n : g_t(x) < 0\}$ we deduce from the above equation that

$$g_t(\rho_f(x_0, t, \mathbf{u}_0)) < 0 \text{ for all } t \in \{0, ..., T\}. \tag{3.55}$$

Therefore,

$$V(x_0, 0) = \inf_{\mathbf{u}\in\Gamma_{x_0,[0,T-1]}} \max_{0\le k\le T} g_k(\rho_f(x_0, k, \mathbf{u})) \le \max_{0\le k\le T} g_k(\rho_f(x_0, k, \mathbf{u}_0)) < 0,$$

where the second inequality follows by (3.55). We therefore deduce $x_0 \in \{x \in \mathbb{R}^n : V(x,0) < 0\}$ and hence $\mathcal{R} \subseteq \{x \in \mathbb{R}^n : V(x,0) < 0\}$.

We next show $\{x \in \mathbb{R}^n : V(x,0) < 0\} \subseteq \mathcal{R}$. Let $x_0 \in \{x \in \mathbb{R}^n : V(x,0) < 0\}$ then,

$$\inf_{\mathbf{u} \in \Gamma_{x_0,[0,T-1]}} \max_{0 \le k \le T} g_k(\rho_f(x_0,k,\mathbf{u})) = V(x_0,0) < 0.$$

Therefore as the above inequality is strict, there exists some $\varepsilon > 0$ such that

$$\inf_{\mathbf{u} \in \Gamma_{x_0}} \max_{0 \le k \le T} g_k(\rho_f(x_0,k,\mathbf{u})) = V(x_0,0) < -\varepsilon. \tag{3.56}$$

By the definition of the infimum for any $\delta > 0$ there exits $\mathbf{w} \in \Gamma_{x_0,[0,T-1]}$ such that

$$\max_{0 \le k \le T} g_k(\rho_f(x_0,k,\mathbf{w})) < \inf_{\mathbf{u} \in \Gamma_{x_0,[0,T-1]}} \max_{0 \le k \le T} g_k(\rho_f(x_0,k,\mathbf{u})) + \delta. \tag{3.57}$$

Hence by letting $0 < \delta < \varepsilon$ we get

$$\max_{0 \le k \le T} g_k(\rho_f(x_0,k,\mathbf{w})) < \inf_{\mathbf{u} \in \Gamma_{x_0,[0,T-1]}} \max_{0 \le k \le T} g_k(\rho_f(x_0,k,\mathbf{u})) + \delta < -\varepsilon + \delta < 0, \tag{3.58}$$

where the first inequality follows by (3.57), the second inequality follows from (3.56), and the third inequality follows from selecting $\delta < \varepsilon$.

Therefore by (3.58) there exists $\mathbf{w} \in \Gamma_{x_0,[0,T-1]}$ such that $\max_{0 \le k \le T} g_k(\rho_f(x_0,k,\mathbf{w})) < 0$. We now deduce that for any $t \in \{0,...,T\}$

$$g_t(\rho_f(x_0,t,\mathbf{w})) \le \max_{0 \le k \le T} g_k(\rho_f(x_0,k,\mathbf{w})) < 0.$$

Thus $\rho_f(x_0,t,\mathbf{u}_0) \in \mathcal{A}_t$, implying $x_0 \in \mathcal{R}$. Therefore $\{x \in \mathbb{R}^n : V(x,0) < 0\} \subseteq \mathcal{R}$. □

3.6.1 Numerical Example: Maximal Invariant Sets

Value functions can characterize FTHMISs, as shown by Theorem 3.3. We now approximate a FTHMIS by computing a value function associated with an MSOP $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Backward}$, where $J(\mathbf{u},\mathbf{x}) = \max_{0 \le k \le T} g_k(x(k))$. To compute the value function we use solve the GBE (3.20) for representations maps $\{\phi_t\}_{t=0}^{T}$

as given in Example 3.1 using the discretization scheme outlined in Section 3.4. Let us consider a discrete time switching system, whose Robust Maximal Invariant Set (RMIS) was previously computed in Xue and Zhan (2018):

$$f(x,u,t) = \begin{cases} \begin{bmatrix} x_1 \\ (0.5+u)x_1 - 0.1x_2 \end{bmatrix} & \text{if } 1-(x_1-1)^2 - x_2^2 \le 0 \\ \begin{bmatrix} x_2 \\ 0.2x_1 - (0.1+u)x_2 + x_2^2 \end{bmatrix} & \text{otherwise.} \end{cases} \tag{3.59}$$

We now compute the FTHMIS, denoted by $\mathcal{R}$, associated with

$$\begin{aligned} \mathcal{A}_t &= \{x \in \mathbb{R}^2 : g_t(x) \le 0\} \text{ for all } t \in \{0,..,T\}, \\ g_t(x) &= \left(x_1 - \frac{(t-1)}{4}\right)^2 + \left(x_2 - \frac{(t+1)}{4}\right)^2 - 1.5, \\ X_t &= [-1,1]^2 \text{ for all } t \in \{0,..,T\}, \\ U &= \{u \in \mathbb{R} : u^2 - 0.01 \le 0\}, \quad T = 4. \end{aligned}$$

Figure 3.3 shows the FTHMIS, $\mathcal{R}$, found by using the discretization scheme outlined in Section 3.4 to solve the GBE (3.20) for 5×5 state grid points in $[-1,1]^2$. To represent $\mathcal{R}$ in $\mathbb{R}^2$, once the value function, V, is found at each grid point a polynomial function is fitted and its zero-sublevel set, shown as the orange shaded region, approximately gives $\mathcal{R}$.

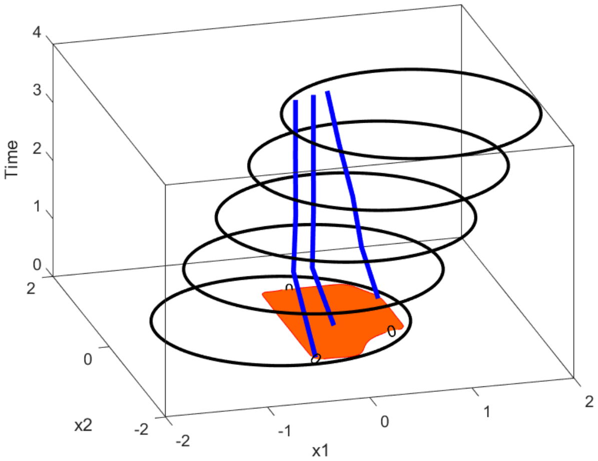

Figure 3.3: Figure showing an approximation of $L(V, 0) := \{x \in \mathbb{R}^n : V(x, 0) \leq 0\}$, shown in the shaded orange region, where V is the value function of the MSOP associated with (3.59). The z-axis represents time and the black circular lines represent the boundary of $\mathcal{A}_t$ for $t = 1, 2, 3, 4$. Three sample trajectories, shown in blue, start in $L(V, 0)$ and remain in the sets $\mathcal{A}_t$ for the time-steps $t = 1, 2, 3, 4$; giving numerical evidence that $L(V, 0)$ is indeed an approximation of the FTHMIS.

Chapter 4

MULTI-STAGE OPTIMIZATION PROBLEMS WITH FORWARD SEPARABLE COSTS

> The problems are solved, not by giving new information, but by arranging what we have known since long.
>
> Ludwig Wittgenstein

4.1 Background and Motivation

In this chapter, as was the case in Chapter 3, we consider Multi-Stage Optimization Problems (MSOPs) initialized at $x_0 \in \mathbb{R}^n$ of the following form:

$$\begin{aligned}
&(\mathbf{x},\mathbf{u}) \in \arg\inf J(\mathbf{x},\mathbf{u}) \text{ subject to:} \qquad (4.1)\\
&\quad \mathbf{u} = (u(0),...,u(T-1)), \mathbf{x} = (x(0),...,x(T))\\
&\quad x(0) = x_0,\;\; x(t+1) = f(x(t),u(t),t) \text{ for } t = 0,..,T-1\\
&\quad x(t) \in X_t \subset \mathbb{R}^n,\;\; u(t) \in U \subset \mathbb{R}^m \text{ for } t = 0,..,T.
\end{aligned}$$

In Chapter 3 we considered MSOPs of Form (4.1) with monotonically backward separable cost functions, functions that can be written as a nested composition of maps backwards in time taking the form

$$J(\mathbf{u},\mathbf{x}) = \phi_0(x(0),u(0),\phi_1(x(1),u(1),\dots\phi_T(x(T))\dots)).$$

In this chapter we consider MSOPs of Form (4.1) with forward separable cost functions, functions that can be written as a nested composition of maps forwards in time taking the form

$$J(\mathbf{u},\mathbf{x}) = \psi_T\left(x\left(T\right),\psi_{T-1}\left(x\left(T-1\right),u\left(T-1\right),\dots,\psi_0\left(x\left(0\right),u\left(0\right)\right)\dots.\right)\right).$$

Like monotonically backward separable functions, the class of forward separable functions contains functions that are not be additively separable, functions the form $J(\mathbf{u}, \mathbf{x}) = \sum_{t=0}^{T-1} c_t\left(x\left(t\right), u\left(t\right)\right) + c_T\left(x\left(T\right)\right)$. Therefore, in general, it is not possible to use classical DP theory (ie solving Bellman's Equation) to solve MSOPs with forward separable cost functions. Moreover, in general, it is also not possible to use methods developed in Chapter 3 to solve MSOPs with forward separable costs that we consider in this chapter. To see this let us consider a function of the following form,

$$J(\mathbf{u}, \mathbf{x}) := \max_{0 \le t \le T} d(x(t)) + \sum_{s=0}^{T-1} c_s(x(s), u(s)). \tag{4.2}$$

It was shown in Lemma 3.3 found in Chapter 3 that J is not a monotonically backward separable function (Defn. 3.3). Since J is not a monotonically backward separable function we are not able to solve the MSOP of Form (4.1) associated with the tuple $\{J, f, \{X_t\}_{0 \le s \le T}, U, T\}$ using the GBE (3.20). However, as shown in Example (4.4) the function J is forward separable and thus the MSOP of Form (4.1) associated with the tuple $\{J, f, \{X_t\}_{0 \le s \le T}, U, T\}$ is solvable using the methodology we will develop in this chapter.

In this chapter we propose a state augmentation method for solving MSOPs with forward separable cost functions that constructs equivalent MSOPs with additively separable objective functions. Intuitively, this state augmentation method increases the state space to include the necessary historic information required for a decision maker to act optimally at some future time step. Such reformulated MSOPs then satisfy the Principle of Optimality (Defn. 3.6) and can therefore be solved using classical DP theory (ie by solving Bellman's Equation). However, the resulting augmented-state MSOP has a higher dimensional state space than the original MSOP - an issue that can potentially render the augmented problem intractable due to the "curse of dimensionality". For this reason, we propose a complexity metric for the forward separable representation and show that in certain cases the dimensionality of the augmented system does not significantly exceed the dimensionality of the original problem - a case where Bellman's equation can be used effectively, see Powell (2007), and which we refer to as *Naturally Forward Separable* (NFS).

To illustrate the proposed methods, in Section 4.6 we consider battery scheduling for mitigating the effect of variability in renewable energy resources. Specifically, we consider historic data for renewable energy sources and design an input sequence for a battery that attempts to minimize energy costs based on time-of-use while also minimizing the maximum rate of energy consumption. Based on this model, we formulate the battery storage problem as a MSOP with a non additively-separable objective function consisting of both integrated time-of-use charges and a maximum term representing the demand charge. The fundamental mathematical challenge with MSOPs of this form is that, as previously shown in Lemma 3.3 (Chapter 3), problems which include maximum terms in the objective in general do not satisfy the *Principle of Optimality.* Thus MSOPs with cost functions including maximum terms cannot be solved by recursively solving the Bellman equation. To overcome this difficulty, we show that the battery scheduling problem is a special case of a MSOP with a NFS objective function. We then apply our state-augmentation technique to numerically solve the deterministic battery scheduling problem for given forecast solar data.

Remarkably, almost no work has been done on optimal use of batteries for reduction of demand charges. The exceptions include the heuristic algorithms of Maly and Kwan (1995) and the pioneering work of Braun and Lee (2006) , which considered *only* a demand charge. Recently this group used an ad-hoc algorithm to consider a combined demand/consumption charge in Cai *et al.* (2016) using detailed models of cooling/load. Furthermore, in Zeinalzadeh and Gupta (2016) a similar energy storage problem is solved using optimized curtailment and load shedding. An L_p

approximation of the demand charge was used in combination with multi-objective optimization in Kamyar and Peet (2016) and, in addition, the optimal use of building mass for energy storage was considered in Kamyar and Peet (2015), wherein a bisection on the demand charges was used. We note that none of these approaches resolve the fundamental mathematical problem of dynamic programming with non-additively separable cost functions.

4.2 Forward Separable Functions

In this section we define a class of functions called forward separable functions. We will show that for MSOPs with a forward separable cost functions, augmenting the state variables allows us to use Bellman's equation to obtain an optimal policy.

Forward separable functions were first defined in Li and Haimes (1987). Intuitively, this is the class of functions that can be separated into a nested composition of maps ordered forward in time. In the next definition we build upon the concept of forward separability by introducing the notion of augmented dimension.

Definition 4.1. *The function $J : \mathbb{R}^{m\times T} \times \mathbb{R}^{n\times(T+1)} \to \mathbb{R}$ is said to be* ***forward separable*** *if there exist representation maps $\psi_0 : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{d_0}$, $\psi_T : \mathbb{R}^n \times \mathbb{R}^{d_{T-1}} \to \mathbb{R}$, and $\psi_i : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ for $i = 1, \ldots, T-1$ such that*

$$J(\mathbf{u}, \mathbf{x}) = \psi_T(x(T), \psi_{T-1}(x(T-1), u(T-1), \psi_{T-2}(...., \\ \psi_1(x(1), u(1), \psi_0(x(0), u(0)))), ...))), \tag{4.3}$$

where $\mathbf{u} = (u(0), ..., u(T-1)) \in \mathbb{R}^{m\times T}$ and $u(i) \in \mathbb{R}^m$ for $i \in \{0, ..., T-1\}$; $\mathbf{x} = (x(0), ..., x(T)) \in \mathbb{R}^{n\times(T+1)}$ and $x(i) \in \mathbb{R}^n$ for $i \in \{0, ..., T\}$; $d_i \in \mathbb{N}$ for $i \in \{0, ..., T-1\}$.

Moreover, we say J is forward separable and has a ***representation dimension*** *of $l \in \mathbb{N}$ if there exists $\{\psi_i\}_{0\leq i\leq T}$ that satisfies Eq. (4.3) and $l = \max_{i\in\{0,...,T-1\}}\{d_i\}$*

where $d_i = \dim(Image\{\psi_i\})$.

Note: The representation dimension of a forward separable function is a property of the set $\{\psi_i\}_{0\leq i\leq T}$ chosen and not the function. However, the function itself does dictate which sets $\{\psi_i\}_{0\leq i\leq T}$ are feasible (satisfy Eq. (4.3)). The representation dimension of a forward separable function is not unique as there could be several sets $\{\psi_i\}_{0\leq i\leq T}$ that satisfy Eq. (4.3). Moreover, the forward separable property of an objective function is independent of the MSOP it is associated with; forward separability is solely a property of the function only and not whatever optimization problem it is being used as an objective function in.

Clearly, any additively separable function of the form $J(\mathbf{u}, \mathbf{x}) = \sum_{t=0}^{T-1} c_t(u(t), x(t)) + c_T(x(T))$ is forward separable and has a representation dimension of 1 using,

$$\begin{aligned}
&\psi_0(x, u) = c_0(x, u) \\
&\psi_i(x, u, w) = c_i(x, u) + w \quad \text{for } i = 1, \cdots, T-1 \\
&\psi_T(x, w) = c_T(x) + w.
\end{aligned} \tag{4.4}$$

4.2.1 MSOPs with Forward Separable Costs May not Satisfy the Principle of Optimality

Consider an MSOP of Form (4.1) associated with $\{J, f, \{X_t\}_{0\leq t\leq T}, U, T\}$. Suppose J is a forward separable function (Defn. 4.1) then the MSOP initialized at $x_0 \in \mathbb{R}^n$

takes the form,

$$\inf_{\mathbf{u},\mathbf{x}} \psi_T(x(T), \psi_{T-1}(x(T-1), u(T-1), \psi_{T-2}(....\psi_0(x(0), u(0)))....))) \tag{4.5}$$

subject to:

$x(t+1) = f(x(t), u(t), t)$ for $t = 0, .., T-1$,

$x(0) = x_0, \ x(t) \in X_t \subset \mathbb{R}^n$ for $t = 0, .., T$,

$u(t) \in U \subset \mathbb{R}^m$ for $t = 0, .., T-1$,

$\mathbf{u} = (u(0), ..., u(T-1))$ and $\mathbf{x} = (x(0), ..., x(T))$.

In Chapter 3 we showed in Lemma 3.3 that the Principle of Optimality (Defn. 3.6) can be used to show if a function J is monotonically backward separable (Defn. 3.3). If the objective function of an MSOP is monotonically backward separable then the MSOP can be solved using the GBE (3.20). However, MSOPs of Form (4.5) with forward separable costs may not satisfy the Principle of Optimality and thus cannot be solved using the GBE (3.20). For instance, it was shown in Lemma 3.3 that the MSOP given in Eq. (3.44) does not satisfy the Principle of Optimality. The objective function of the MSOP given in Eq. (3.44) takes the form

$$J(\mathbf{u}, \mathbf{x}) := \max_{0 \le t \le T} d(x(t)) + \sum_{s=0}^{T-1} c_s(x(s), u(s)). \tag{4.6}$$

We show in Example (4.4) that functions of the Form (4.6) are forward separable (Defn 4.1). Therefore, the MSOP given in Eq. (3.44) is an MSOP that does not satisfy the Principle of Optimality but is of the Form (4.5). Thus, in order to solve MSOPs of Form (4.5) we next develop a new solution strategy based on state space augmentation, that is independent of the methodology presented in Chapter 3.

4.2.2 How State Augmentation Solves MSOPs with Forward Separable Costs

For any MSOP with forward separable cost function, of Form (4.5), we may associate a new augmented-state MSOP which is shown to be equivalent to MSOP (4.7) (Lemma 4.1) and satisfies the Principle of Optimality (Corollary 4.1). The augmented-state MSOP initialized at $x_0 \in \mathbb{R}^n$ takes the following form,

$$\inf_{\mathbf{u},\mathbf{z}} \psi_T(z_1(T), z_2(T)) \tag{4.7}$$

$$\text{subject to:}$$

$$\begin{bmatrix} z_1(t+1) \\ z_2(t+1) \end{bmatrix} = \begin{bmatrix} f(z_1(t), u(t), t) \\ \psi_t(z_1(t), u(t), z_2(t)) \end{bmatrix} \; 0 \le t < T-1$$

$$\begin{bmatrix} z_1(0) \\ z_2(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ \emptyset \end{bmatrix}, \; z_1(t) \in X_t, \; u(t) \in U \text{ for } t = 0,..,T$$

$$\mathbf{u} = (u(0), ..., u(T-1)) \text{ and } \mathbf{z} = \left(\begin{bmatrix} z_1(0) \\ z_2(0) \end{bmatrix}, ..., \begin{bmatrix} z_1(T) \\ z_2(T) \end{bmatrix} \right),$$

where $z_1(t) \in \mathbb{R}^n$, $z_2(t) \in \mathbb{R}^{d_t}$, $d_t = \dim(Image\{\psi_{t-1}\})$ and $u(t) \in \mathbb{R}^m$ for all $t \in \{0, ..., T\}$.

Lemma 4.1 (Equivalence of MSOPs). *Consider the tuple $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\}$, where J is a forward separable function (Defn. 4.1) with representation maps $\{\phi_i\}_{t_0 \le i \le T}$. Consider the associated MSOP of Form* (4.5) *and its augmented MSOP of Form* (4.7)*. Then the following holds*

1. *$J^* = L^*$ where J^* and L^* are the optimal objective functions of the MSOPs given in Eqs.* (4.5) *and* (4.7) *respectively.*

2. *If $\mathbf{u} = (u(0), ..., u(T-1))$ and $\mathbf{x} = (x(0), ..., x(T))$ solve MSOP* (4.5) *and $\mathbf{w} = (w(0), ..., w(T-1))$ and $\mathbf{z} = \left(\begin{bmatrix} z_1(0) \\ z_2(0) \end{bmatrix}, ..., \begin{bmatrix} z_1(T) \\ z_2(T) \end{bmatrix} \right)$ solve MSOP* (4.7)*. Then $\mathbf{u} = \mathbf{w}$ and $x(t) = z_1(t)$ for all $t \in \{0, ..., T\}$.*

Proof. Suppose $(\mathbf{w}, \mathbf{z})$ solve the state augmented MSOP given in Eq. (4.7). First we show that $\mathbf{w}$ and $\mathbf{z}_1 := (z_1(0), ..., z_1(T))$ are feasible for the MSOP given in Eq. (4.5). Clearly $w(t) \in U$ for all $t \in \{0, ..., T\}$ and if we let $\mathbf{u} = \mathbf{w}$, $x(0) = x_0$, and $x(t+1) = f(x(t), u(t), t)$ for all $t \in \{0, ..., T\}$ we get $x(t) = z_1(t) \in X_t$ for all $t \in \{0, ..., T\}$, where $z_1(0) = x_0$ and $z_1(t+1) = f(z_1(t), u(t), t)$ for all $t \in \{0, ..., T\}$. Hence $\mathbf{u}$ and $\mathbf{x} = \mathbf{z}_1$ are feasible for the MSOP given in Eq. (4.5).

We now observe

$$\begin{aligned}
z_2(T) &= \psi_{T-1}(z_1(T-1), u(T-1), z_2(T-1)).\\
&\vdots\\
z_2(1) &= \psi_0(z_1(0), u(0)).\\
z_2(0) &= \emptyset.
\end{aligned}$$

Hence we have,

$$\begin{aligned}
L(\mathbf{w}, \mathbf{z}) &= \psi_T(z_1(T), z_2(T))\\
&\vdots\\
&= \psi_T(x(T), \psi_{T-1}(x(T-1), u(T-1), \psi_{T-2}(...., \psi_1(x(1), u(1), \psi_0(x(0), u(0)))), ...)))\\
&= J(\mathbf{u}, \mathbf{x}).
\end{aligned}$$

Thus if $(\mathbf{w}, \mathbf{z})$ solve the state augmented MSOP given in Eq. (4.7) with objective $L^* = \psi_T(z_1(T), z_2(T))$, then $(\mathbf{w}, \mathbf{z}_1)$ solve MSOP (4.5) with objective value J^*. □

As shown in Section 4.2.1 MSOPs with forward separable costs, of Form (4.5), may not satisfy the Principle of Optimality (Defn. 3.6). However, as we will show next the associated augmented MSOP, of Form 4.7, does satisfy the Principle of Optimality.

Corollary 4.1 (Augmented MSOPs satisfy the Principle of Optimality)**.** *The family of state augmented MSOP given in Eq.* (4.7) *satisfies the Principle of Optimality (Definition 3.6)*

Proof. MSOP (4.7) has a cost function that only depends on the terminal state $(z_1(T), z_2(T))^T$. Therefore the cost function of MSOP (4.7) is additively separable (Definition 3.2). It was shown in Lemma 3.1 that additively separable functions are monotonically backward separable with representation maps that are strictly monotonic (Eq. (3.24)). Therefore we deduce the family of MSOPs given in Eq. (4.7) satisfies the Principle of Optimality by Proposition 3.3. □

Lemma 4.1 tells us that for any MSOP with forward separable objective, given in Eq. (4.5), there exists an equivalent state augmented MSOP given in Eq. (4.7). Furthermore, Corollary 4.1 shows that MSOPs of Form (4.5) satisfy the Principle of Optimality. Therefore, a solution for any MSOP of Form (4.5) can be found by using DP methods that solve the associated augmented MSOP (4.7).

To understand the augmented approach intuitively, we note that DP methods break a multi-period planning problem into simpler sub-problems indexed by each time-stage. However, in order for a decision maker to make the optimal decision at each time-stage when faced with an exotic cost function that may depend on historic states, the decision maker may require past information about the system and not just the current state of the system. In this context, the augmented state contains the necessary information about the historic states taken by the system trajectories required by a decision maker to act optimally at each time-stage. However, as we will see next, by adding an augmented state we increase the state space dimension and hence increase the complexity of the MSOP.

Corollary 4.2. *Consider the tuple $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\}$. Suppose J is a forward separable function with representation dimension of $l \in \mathbb{N}$. Consider the associated MSOP of Form* (4.5)*. If the MSOP has state space dimension $n \in \mathbb{N}$ and input space of dimension $m \in \mathbb{N}$, then the associated state augmented MSOP* (4.7) *has a state* *space of dimension $l + n$ and input space of dimension m.*

Proof. The state space dimension of the MSOP (4.7) is $n + \max_{0 \le t \le T} d_t$, where $d_t = \dim(Image\{\phi_{t-1}\})$. From the definition of representation dimension, we have $\max_{t_0 \le t \le T} d_t = l$ and hence it follows that the state space dimension of the MSOP (4.7) is $n + \max_{t_0 \le t \le T} d_t = n + l$. □

4.3 MSOPs for which the use of State Augmentation is Tractable

It is well known that solving MSOPs by discretizing the state space and recursively solving Bellman's equation (3.27) is computationally intractable when the state space dimension is large; this is often called "the curse of dimensionality". In the previous Section 4.2.2, we proved that any MSOP, with forward separable cost function, of state space dimension $n \in \mathbb{N}$ can be converted to an equivalent MSOP, with additively separable costs and with state-space dimension $n + l \in \mathbb{N}$, where $l \in \mathbb{N}$ is the representation dimension of the forward separable cost function. However, for some representations, l may increase with respect to the MSOPs data; such as the state space dimension, input space dimension and the number of time-stages of the MSOP- thus triggering the curse of dimensionality. To address this problem, in this section, we define a class of forward separable functions, called Naturally Forward Separable (NFS) functions, with representation dimension, $l \in \mathbb{N}$, that is independent of the number of time-stages and the dimension of the state and input space.

Before we define NFS functions we motivate this new class of functions by showing that it is possible to represent any function as a forward separable function (Defn. 4.1). To do this we introduce some additional notation. Specifically, for a vector $v = (v_1, ..., v_n)^T \in \mathbb{R}^n$ we define $[v]_i^j = (v_i, ..., v_j)$ for $1 \le i < j \le n$.

Lemma 4.2. *Any function $J : \mathbb{R}^{m \times T} \times \mathbb{R}^{n \times (T+1)} \to \mathbb{R}$ is forward separable (Defn. 4.1)* *with a representation of dimension* $l(n, m, T) = T(n + m)$*.*

Proof. Consider a function $J : \mathbb{R}^{m\times T} \times \mathbb{R}^{n\times(T+1)} \to \mathbb{R}$. To show J is forward separable we define a forward separable representation $\{\psi_i\}_{i=0}^{T}$ which satisfy Eq. (4.3) as follows.

First, define $\psi_0 : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{n+m}$ as

$$\psi_0(x, u) = [x^T, u^T] = \left[x_1, ..., x_n, u_1, ..., u_m \right].$$

For $i \in \{1, ...T - 1\}$ we define $\psi_i : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{i(n+m)} \to \mathbb{R}^{(i+1)(n+m)}$ as

$$\psi_i(x, u, w) = \left[[w]_1^{ni}, x^T, [w]_{ni+1}^{(i)(n+m)}, u^T \right].$$

Lastly, define $\phi_T : \mathbb{R}^n \times \mathbb{R}^{T(n+m)} \to \mathbb{R}$ as

$$\psi_T(x, w) = J([[w]_1^{nT}, x], [w]_{nT+1}^{T(n+m)}).$$

Clearly, this definition of $\{\psi_i\}_{0\le i\le T}$ satisfies Eq. (4.3). Furthermore, it can be seen that the maximum dimension of the images of the maps $\{\psi_i\}_{0\le i\le T}$ is $T(n+m)$ showing the dimension of this representation of J is $l(n, m, T) = T(n + m)$. □

In Lemma 4.2 we showed that any function J is forward separable by naively taking the strategy of considering representation functions $\{\psi_i\}_{0\le i\le T}$ that act like memory functions; that is to store the entire historic state trajectory and input sequence used. If J is the objective function for some MSOP (4.5) then this approach would result in the associated equivalent state-augmented MSOP (4.7), having a very large state space dimension. Specifically, Corollary 4.2 shows that MSOP (4.7) has state space dimension $T(n + m) + n$. Clearly, for a large number of time-stages, $T \in \mathbb{N}$, MSOP (4.7) is intractable. For this reason we next define a special class of forward separable functions that have a representation with dimension independent of the number of time-stages.

Definition 4.2. *We say a function $J : \mathbb{R}^{m \times T} \times \mathbb{R}^{n \times (T+1)} \to \mathbb{R}$ is a Naturally Forward Separable (NFS) function if there exists representation maps, $\{\psi_i\}_{i=0}^{T}$, that satisfy Eq.* (4.3) *with representation dimension independent of n, m and T.*

In Chapter 3 we defined classes of MSOPs, $\mathcal{M}_{Addative}^{Discrete}$ and $\mathcal{M}_{Backward}^{Discrete}$, that we can tractably solve. We now add to this class by considering MSOPs with NFS cost functions (Defn. 4.2) .

Definition 4.3. *We say the five tuple $\{J, f, \{X_t\}_{0 \leq t \leq T}, U, T\}$ is a **naturally forward separable MSOP** or $\{J, f, \{X_t\}_{0 \leq t \leq T}, U, T\} \in \mathcal{M}_{Forward}^{Discrete}$ if J is a naturally forward separable function (Defn. 4.2), $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{N} \to \mathbb{R}^n$, $X_t \subset \mathbb{R}^n$, $U \subset \mathbb{R}^m$, and $T \in \mathbb{N}$. Each element of $\mathcal{M}_{Forward}^{Discrete}$ is associated with a MSOP of Form* (4.5).

4.3.1 An Algebra of Naturally Forward Separable Functions

Given a function, $J : \mathbb{R}^{m \times T} \times \mathbb{R}^{n \times (T+1)} \to \mathbb{R}$, it was shown in Section 3.3 that the Principle of Optimality can be used to show J is not monotonically backward separable. However, in the case of forward separability, there is no obvious way to determine whether J is NFS (Defn. 4.2). Instead, in this section, we show that the set of NFS functions form an algebra, closed under multiplication and under nonlinear transformations - implying that simple NFS functions ("building blocks") can be combined to construct new, more complex, NFS functions. In this way, one might approach the problem of showing J is NFS by finding representation maps by combining known NFS "building block" functions. Several examples of such "building blocks" can be found in Subsection 4.3.2.

Lemma 4.3 (The set of NFS functions is closed under addition, multiplication, and nonlinear mappings)**.** *Consider the function $U : \mathbb{R} \to \mathbb{R}$ and the NFS functions (Defn. 4.2), $J_1 : \mathbb{R}^{m_1 \times T_1} \times \mathbb{R}^{n_1 \times (T_1+1)} \to \mathbb{R}$ and $J_2 : \mathbb{R}^{m_2 \times T_2} \times \mathbb{R}^{n_2 \times (T_2+1)} \to \mathbb{R}$, with* *representation dimensions $l_1 \in \mathbb{N}$ and $l_2 \in \mathbb{N}$ respectively. The functions $G_1(\mathbf{u},\mathbf{x}) := J_1(\mathbf{u},\mathbf{x}) + J_2(\mathbf{u},\mathbf{x})$, $G_2(\mathbf{u},\mathbf{x}) := J_1(\mathbf{u},\mathbf{x}) \cdot J_2(\mathbf{u},\mathbf{x})$ and $G_3(\mathbf{u},\mathbf{x}) := U\left(J_1(\mathbf{u},\mathbf{x})\right)$ are NFS functions with representation dimension less than or equal to $l_1+l_2 \in \mathbb{N}$, $l_1+l_2 \in \mathbb{N}$, and $l_1 \in \mathbb{N}$, receptively.*

Proof. For simplicity let us consider the case where $T_1 = T_2$; other cases follow by the same argument. As J_1 and J_2 are forward separable functions there exist associated representation maps $\{g_i\}_{0\leq i\leq T_1}$ and $\{h_i\}_{0\leq i\leq T_2}$ such that J_1 and J_2 can be written in the Form (4.3) with associated representation dimensions l_1 and l_2, respectively. We now show that G_1 is forward separable by defining the associate representation $\{\psi_i\}_{0\leq i\leq T_1}$ such that G_1 can be written in the Form (4.3). Specifically, let

$$\psi_0(x,u) = \begin{bmatrix} g_0(x,u) \\ h_0(x,u) \end{bmatrix}, \tag{4.8}$$

$$\psi_i(x,u,w) = \begin{bmatrix} g_i(x,u,[w]_1^{d_{i-1}}) \\ h_i(x,u,[w]_{d_{i-1}+1}^{d_{i-1}+s_{i-1}}) \end{bmatrix} \text{ for } i \in \{1,....,T_1-1\},$$

$$\psi_{T_1}(x,w) = g_T(x,[w]_1^{d_{T_1-1}}) + h_T(x,[w]_{d_{T_1-1}+1}^{d_{T_1-1}+s_{T_1-1}}),$$

where $d_i = \dim(Image\{g_i\})$ and $s_i = \dim(Image\{h_i\})$ for $i \in \{0,...,T_1-1\}$.

We conclude that G_1 has a representation dimension, denoted l_{G_1}, such that

$$l_{G_1} = \max_{i\in\{0,...,T_1-1\}}\{d_i+s_i\} \leq \max_{i\in\{0,...,T_1-1\}}\{d_i\} + \max_{i\in\{0,...,T_1-1\}}\{s_i\} = l_1+l_2.$$

Furthermore, by a similar argument it can be shown that G_2 and G_3 are NFS with representation dimension less than or equal to l_1+l_2 and l_1 respectively. We are able to show this using the same representation maps $\{\psi_i\}_{0\leq i\leq T_1-1}$ from Eq. (4.8) with the terminal representation map for G_2 given by

$$\psi_{T_1}(x,w) = g_T\left(x,[w]_1^{d_{T_1-1}}\right) \cdot h_T\left(x,[w]_{d_{T_1-1}+1}^{d_{T_1-1}+s_{T_1-1}}\right).$$

For G_3 we use representation maps $\psi_t = g_t$ for $t \in \{0, ..., T_1 - 1\}$ and the terminal representation map for G_3 given by

$$\psi_{T_1}(x, w) = U\left(g_T\left(x, [w]_1^{d_{T_1-1}}\right)\right).$$

□

4.3.2 Examples: Naturally Forward Separable Functions

The first example of a NFS function (Defn. 4.2) is found in problems involving risk measures and certainty equivalents Bäuerle and Rieder (2013). In this case, we have the function $U(x) = \frac{1}{\gamma}e^{\gamma x}$ and apply the following:

Example 4.1. *For any functions $U : \mathbb{R} \to \mathbb{R}$ and $c_t : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$,*

$$J(\mathbf{u}, \mathbf{x}) = U\left(\sum_{t=0}^{T-1} c_t(x(t), u(t))\right)$$

is NFS with representation dimension 1.

Proof. The additively separable function $\sum_{t=0}^{T-1} c_t(x(t), u(t))$ is NFS using the representation maps given in Eq. (4.4). It therefore follows J is NFS by Lemma 4.3. □

Example 4.2. *The mixed p-norm function given by*

$$J(\mathbf{u}, \mathbf{x}) = \sum_{j=1}^{N}\left(\sum_{t=0}^{T-1} c_{j,t}(x(t), u(t))^{p_j}\right)^{\frac{1}{p_j}},$$

where $p_j > 0$ for all $j \in \{1, ..., N\}$, $c_{j,t} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^+$, and $N \in \mathbb{N}$, is NFS with representation dimension $N \in \mathbb{N}$.

Proof. Follows since J that can be written in the Form (4.3) using maps

$$\psi_0(x, u) = [c_{1,t_0}(x(0), u(0))^{p_1}, ..., c_{N,t_0}(x(0), u(0))^{p_N}]^T,$$

$$\psi_i(x, u, w) = [c_{1,i}(x(i), u(i))^{p_1} + w_1, ..., c_{N,i}(x(i), u(i))^{p_N} + w_N]^T \text{ for } i \in \{1, ..., T-1\},$$

$$\psi_T(x, w) = \sum_{j=1}^{N} w_j^{\frac{1}{p_j}}.$$

□

We next consider a function that appears as the objective function of an optimization problem in Domingo and Sniedovich (1993).

Example 4.3. *Consider the variance type function, $J : \mathbb{R}^{m\times T}\times\mathbb{R}^{n\times(T+1)} \to \mathbb{R}$ defined as*

$$J(\mathbf{u},\mathbf{x}) = \sum_{t=0}^{T}\left[a_t(x(t)) - \frac{1}{T}\sum_{s=0}^{T} a_s(x(s))\right]^2 \tag{4.9}$$

where $\mathbf{u} = (u(0), ..., u(T-1))$, $u(t) \in \mathbb{R}^m$, $\mathbf{x} = (x(0), ..., x(T))$, $x(t) \in \mathbb{R}^n$, and $a_t : \mathbb{R}^n \to \mathbb{R}$. Then J is NFS and has a representation dimension of 2.

Proof. Expanding the right hand side of (4.9) as in Domingo and Sniedovich (1993) we get,

$$\begin{aligned} J(\mathbf{u},\mathbf{x}) &= \sum_{t=0}^{T}\left[a_t^2(x(t)) - \frac{2}{T}a_t(x(t))\sum_{s=0}^{T}a_s(x(s)) + \frac{1}{T^2}\left(\sum_{s=0}^{T}a_s(x(s))\right)^2\right] \\ &= \sum_{t=0}^{T}a_t^2(x(t)) - \frac{1}{T}\left[\sum_{s=0}^{T}a_s(x(s))\right]^2. \end{aligned}$$

We now present functions J that can be written in the form of Eq. (4.3). We define $\psi_0 : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^2$ as

$$\psi_0(x,u) = \begin{bmatrix} a_1^2(x) \\ a_1(x) \end{bmatrix}.$$

We define $\psi_i : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^2 \to \mathbb{R}^2$ as

$$\psi_i(x,u,[w_1,w_2]^T) = \begin{bmatrix} w_1 + a_i^2(x) \\ w_2 + a_i(x) \end{bmatrix} \text{ for all } 1 \le i \le T-1.$$

Finally, $\psi_T : \mathbb{R}^n \times \mathbb{R}^2 \to \mathbb{R}$ is given by,

$$\psi_T(x,[w_1,w_2]^T) = (w_1 + a_T^2(x)) - \frac{1}{T}(w_2 + a_T(x))^2.$$

This definition of $\{\psi_i\}_{i=0}^T$ satisfies (4.3). Moreover it can be seen that the maximum dimension of the images of the maps $\{\psi_i\}_{i=0}^T$ is 2 showing the dimension of this representation of J is 2. □

We now show that the maximum function, that appears in the objective function of the battery scheduling problem in Section 4.6, is NFS. We also note that in Lemma 3.3 it was shown functions of this form are not monotonically backward separable (Defn. 3.3).

Example 4.4. *Consider the function* $J : \mathbb{R}^{m\times T} \times \mathbb{R}^{n\times(T+1)} \to \mathbb{R}$ *such that,*

$$J(\mathbf{u},\mathbf{x}) = \max\{\max_{0\le k\le T-1}\{d_k(u(k),x(k))\}, d_T(x(T))\} + \sum_{s=0}^{T-1} c_s(x(s),u(s)) + c_T(x(T))$$

where $\mathbf{u} = (u(0),...,u(T-1))$, $u(t) \in \mathbb{R}^m$, $\mathbf{x} = (x(0),...,x(T))$, $x(t) \in \mathbb{R}^n$, $c_k : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ *and* $d_k : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ *for* $0 \le k \le T-1$, $c_T : \mathbb{R}^n \to \mathbb{R}$ *and* $d_T : \mathbb{R}^n \to \mathbb{R}$. *Then* J *is NFS (Defn. 4.2) and has a representation dimension of 2.*

Proof. To show J is NFS we first show that J can be written in the Form (4.3) using the following representation maps. We define $\psi_0 : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^2$ as

$$\psi_0(x,u) = \begin{bmatrix} d_0(x,u) \\ c_0(x,u) \end{bmatrix}.$$

The function $\psi_i : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^2 \to \mathbb{R}^2$ is defined by,

$$\psi_i(x,u,[w_1,w_2]^T) = \begin{bmatrix} \max(d_i(x,u),w_1) \\ c_i(x,u) + w_2 \end{bmatrix} \text{ for all } 1 \le i \le T-1.$$

The function $\psi_T : \mathbb{R}^n \times \mathbb{R}^2 \to \mathbb{R}$ is defined by,

$$\psi_T(x,[w_1,w_2]^T) = \max(d_T(x),w_1) + c_T(x) + w_2.$$

The representation maps, $\{\psi_i\}_{0\le i\le T}$, satisfy Eq. (4.3). Moreover, it can be seen that the maximum dimension of the images of $\{\psi_i\}_{i=0}^T$ is 2. Thus the dimension of this representation of J is 2. □

4.4 Numerical Example: Solving MSOPs with NFS Cost Functions

Given a MSOP with a NFS (Defn. 4.2) cost function, with known representation maps, we have shown in Section 4.2 how to use state augmentation to construct an equivalent MSOP with additively separable (Defn. 3.2) cost functions. We have furthermore proposed a discretization scheme in Section 3.4 to solve MSOPs with infinite input and state spaces and additively separable cost functions. We now summarize these results by proposing the following steps for solving a given general MSOP. Given an MSOP of Form (4.5), associated with the tuple $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}_{Forward}^{Discrete}$, we do the following:

1. Find a NFS representation of the objective function (Eq. (4.3)) with associated representation maps $\{\psi_t\}_{t=0}^{T}$. One approach to this is to use Section 4.3.1 which details how to combine known NFS functions, with known representation maps, in order to find potential representation maps for other NFS functions.

2. Construct the associated augmented MSOP of Form (4.7).

3. Use discretization to approximate the augmented MSOP (4.7) to an discretized MSOP of Form (3.45).

4. Numerically solve the discretized MSOP (3.45) using recursive application of Bellman's Equation (3.27).

5. Construct a feasible policy for the original MSOP from an optimal policy ofthe discretized MSOP (3.45) using Eq. (3.47).

To illustrate how we use state augmentation and discretization methods we consider the following MSOP from Li and Haimes (1991).

$$\inf_{u(0),u(1),u(2)} x(3)^2[u(0)^2+u(1)^2+u(1)u(2)^2]^{\frac{1}{2}}+[u(0)^2+u(1)^2+u(1)u(2)^2]^2 \tag{4.10}$$
$$\text{subject to,}\quad x(t+1)=\frac{x(t)}{u(t)}\quad \text{for } t\in\{0,1,2\}$$
$$x(0)=10,\quad u(0),u(1),u(2)\geq 0.$$

Clearly the MSOP given in Eq. (4.10) can be written in the Form (4.5) with associated tuple $\{J,f,\{X_t\}_{0\leq t\leq T},U,T\}\in\mathcal{M}^{Discrete}_{Forward}$, where $J(\mathbf{u},\mathbf{x})=x(3)^2[u(0)^2+u(1)^2+u(1)u(2)^2]^{\frac{1}{2}}+[u(0)^2+u(1)^2+u(1)u(2)^2]^2$ is NFS with representation maps later given in Eq. (4.11), $f(x,u,t)=x/u$, $X_t=\mathbb{R}$, $U=[0,\infty)$, and $T=3$.

In Li and Haimes (1991) an analytic solution for Opt. (4.10) was found to be:

$$\mathbf{x}^*=\begin{bmatrix}10\\6.3943938\\5.782475\\3.8882658\end{bmatrix},\ \mathbf{u}^*=\begin{bmatrix}1.5638699\\1.105823\\1.4871604\end{bmatrix},\ J^*=74.767439.$$

The objective function in Opt. (4.10) is NFS (Defn. 4.2) and has a representation dimension of 2. This can be shown by writing the objective function J in the Form (4.3) using the functions,

$$\psi_0(x,u)=u^2,\quad \psi_1(x,u,w)=\begin{bmatrix}w+u^2\\u\end{bmatrix} \tag{4.11}$$
$$\psi_2\left(x,u,\begin{bmatrix}w_1\\w_2\end{bmatrix}\right)=w_1+w_2^2u^2,\quad \psi_3(x,w)=x^2\sqrt{w}+w^2.$$

The MSOP given in Eq. (4.10) can now be written in the form of the augmented

MSOP given in Eq. (4.7) as follows,

$$\min\{z_1(3)^2\sqrt{z_3(3)} + z_3(3)^2\} \tag{4.12}$$

subject to,

$$z_1(t+1) = \frac{z_1(t)}{u(t)},\ z_2(t+1) = \begin{cases} u(t) \text{ if t=1} \\ \\ 0 \text{ otherwise} \end{cases} \quad \text{for all } t \in \{0,1,2\},$$

$$z_3(1) = u(1)^2, \quad z_3(2) = z_3(1) + u(1)^2,$$

$$z_3(3) = z_3(2) + z_2(2)^2 u(2),$$

$$z_1(0) = 10,\ z_2(0) = 0,\ z_3(0) = 0 \quad u(0), u(1), u(2) \geq 0.$$

The MSOP given in Eq. (4.12) has a cost function that is additively separable (Defn. 3.2) and so can be solved using the discretization methods in Section 3.4. Moreover, by Lem. 4.1 the MSOP given in Eq. (4.12) is equivalent to the original MSOP given in Eq. (4.10).

Figure 4.1 shows the state trajectories associated with input sequences constructed using various discretization values $k \in \mathbb{N}$. It is seen that for $k = 200$ the algorithm produces a solution within three significant figures of the analytic optimal objective function for the MSOP given in Eq. (4.10).

4.5 Comparison: State Augmentation Methods vs GBE Methods

The sets of naturally monotonically backward separable functions (Defn. 3.3) and naturally forward separable functions (Defn. 4.2) are not disjoint. For instance, the function

$$J(\mathbf{u}, \mathbf{x}) = \max\left\{ \max_{0 \leq k \leq T-1} \{d_k(x(k), u(k))\}, d_T(x(T)) \right\}, \tag{4.13}$$

is both a naturally monotonically backward separable function (as shown in Example 3.1) and a naturally forward separable function (as shown in Example 4.4 setting

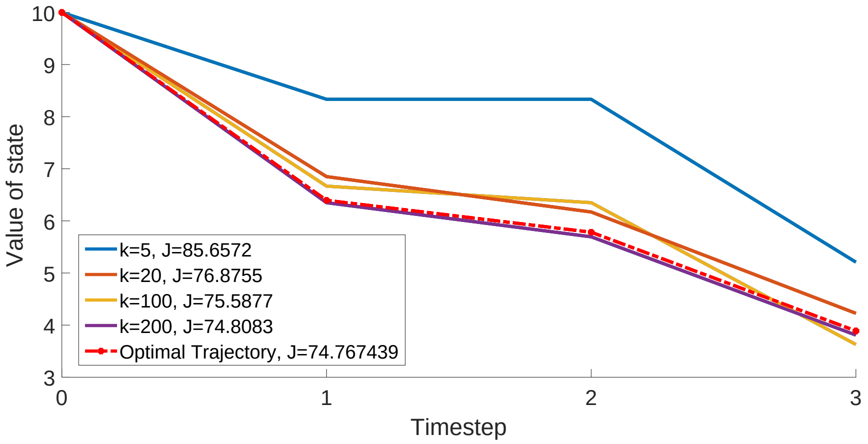

Figure 4.1: State trajectories associated with input sequences constructed from various discretization levels for the MSOP given in Eq. (4.10).

$c_k(x,u) \equiv 0$ for all $k \in \{0,...,T-1\}$ and $c_T(x) \equiv 0$). Thus as we will show next, the class of backward and forward separable MSOPs intersect.

Corollary 4.3. *Recalling $\mathcal{M}^{Discrete}_{Backward}$ is given in Defn. 3.4 and $\mathcal{M}^{Discrete}_{Forward}$ is given in Defn. 4.3 we have that*

1. $\mathcal{M}^{Discrete}_{Backward} \cap \mathcal{M}^{Discrete}_{Forward} \neq \emptyset$.

2. $\mathcal{M}^{Discrete}_{Forward} \not\subseteq \mathcal{M}^{Discrete}_{Backward}$.

Proof. The MSOP associated with the tuple $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\}$ where J is given in Eq. (4.13) is clearly an element of both $\mathcal{M}^{Discrete}_{Backward}$ and $\mathcal{M}^{Discrete}_{Forward}$ and thus $\mathcal{M}^{Discrete}_{Backward} \cap \mathcal{M}^{Discrete}_{Forward} \neq \emptyset$. On the other hand if J is of the form

$$J(\mathbf{u},\mathbf{x}) = \max\{\max_{0 \le k \le T-1}\{d_k(u(k),x(k))\}, d_T(x(T))\} + \sum_{s=0}^{T-1} c_s(x(s),u(s)) + c_T(x(T)), \tag{4.14}$$

then Example 4.4 shows $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\}$ is an element of $\mathcal{M}^{Discrete}_{Forward}$ and Lemma 3.3 shows there exist $\{d_k\}_{k=0}^T$ and $\{c_k\}_{k=0}^T$ such that $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\}$ is not an element of $\mathcal{M}^{Discrete}_{Backward}$. Therefore, $\mathcal{M}^{Discrete}_{Forward} \not\subseteq \mathcal{M}^{Discrete}_{Backward}$. □

For an MSOP with a cost function that is both backward and forward separable, such as the function given in Eq. (4.13), we can solve the MSOP by recursive application of the GBE (3.20) or we can solve the MSOP by constructing an equivalent augmented MSOP (4.7) that can be solved using the BE (3.27). State augmentation methods involve increasing the dimension of the state space and therefore the complexity of the MSOP. Therefore, we prefer to solve MSOPs without using state augmentation methods whenever possible. The Principle of Optimality (Defn. 3.6) can be used as a guide to determine whether a function is backward separable or not (as shown in Lemma 3.3) and hence determine if an MSOP can be solved using the GBE (3.20). Therefore the Principle of Optimality can be used as a guide to whether state augmentation methods are required to solve an MSOP.

On the other hand, as shown in Lemma 4.2, every function is forward separable. Therefore, although state augmentation methods typically have the disadvantage of large computation times, they do have the advantage that they can be used to solve a larger class of problems when compared to methods involving the use of the GBE (3.20). For instance, MSOPs with cost functions of the Form (4.14) are members of $\mathcal{M}_{Forward}^{Discrete}$, and hence can be tractably solved using state augmentation, but such MSOPs may not be members of $\mathcal{M}_{Backward}^{Discrete}$ (Lemma 3.3) and thus cannot be solved using the GBE (3.20).

We next provide a pathological example of an MSOP in $\mathcal{M}_{Backward}^{Discrete}$ but not clearly in $\mathcal{M}_{Forward}^{Discrete}$. We present the computation times for solving the MSOP using the GBE (3.20) and using state augmentation methods. Since the cost function of the MSOP is not clearly NFS (Defn. 4.2) we resort to the naive approach of writing the cost function in forward separable Form (4.3) using representation maps that store the entire historic state and input sequences, vastly increasing the complexity of the MSOP once augmented. This pathological example demonstrates how superior computation times can be achieved by methods that solve MSOPs without the use of state augmentation.

Consider an MSOP of Form (4.1) associated with tuple $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\}$, where $J(\mathbf{u}, \mathbf{x}) = \sqrt{x(0) + u(0) + \sqrt{.....\sqrt{x(T-1) + u(T-1) + \sqrt{x(T)}}}}$, $X_t = \{1, 2\}$, $U = \{0.5, 1\}$, $T \in \mathbb{N}$. Let us consider the MSOP initialized at $(x_0, t_0) = (2, 0)$,

$$\begin{aligned}
&\min_{\mathbf{u},\mathbf{x}} \sqrt{x(0) + u(0) + \sqrt{.....\sqrt{x(T-1) + u(T-1) + \sqrt{x(T)}}}} \\
&\text{subject to:} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (4.15) \\
&x(t+1) = \begin{cases} 2 \text{ if } u = 0.5 \\ 1 \text{ if } u = 1 \end{cases} \qquad \text{for } t = 0, .., T, \\
&x(0) = 2, \;\; x(t) \in \{1, 2\} \text{ for } t = 0, .., T, \\
&u(t) \in \{0.5, 1\} \text{ for } t = 0, .., T-1.
\end{aligned}$$

The cost function in the above MSOP is naturally monotonically backward separable (Defn. 3.3) and can be written in the Form (3.4) with representation maps

$$\phi_T(x) = \sqrt{x}, \;\; \phi_t(x, u, z) = \sqrt{x + u + z} \text{ for } t \in \{0, .., T-1\}. \qquad (4.16)$$

Moreover the cost function is also forward separable and can be written in the Form (4.3) with representation maps

$$\psi_0(x, u) = [x, u]^T, \quad \psi_t(x, u, z) = [z, x, u]^T, \qquad (4.17)$$
$$\psi_T(x, z) = \sqrt{z_1 + z_2 + \sqrt{....\sqrt{z_{2T-1} + z_{2T} + \sqrt{x}}}}.$$

We solved the MSOP in Eq. (4.15) using both the GBE and the state augmentation method, plotting the computation time results in Figure 4.2. We note that the discretization methods presented in Section 3.4 were not needed to solve this MSOP

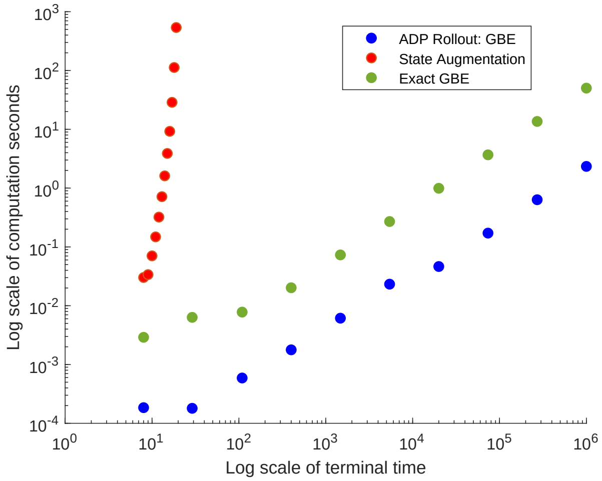

Figure 4.2: Log log graph showing computation time for solving MSOP (4.15) using state augmentation (red points), via exactly solving the GBE (green points), and via approximately solving the GBE using the rollout (blue points) algorithm versus the terminal time of the problem.

since the state and input spaces were already finite. The green points represent the computation time required to construct the value function by solving the GBE (3.20) with representation maps given in Eq. (4.16), and then to synthesize the optimal input sequence using Eq. (3.21). The red points represent the computation time required to construct the value function (Defn. 3.5) by solving Bellman's Equation (3.27) for the associated state augmented MSOP (4.7) and then to construct the optimal input sequence. The green points increases linearly as a function of the terminal time, $T \in \mathbb{N}$, of order $\mathcal{O}(T)$, whereas the red points increases exponentially with respect to T, of order $\mathcal{O}(2^T)$ (due to the fact that using representation maps, given in Eq. (4.17), results in an augmented state space of size 2^T). Moreover, Figure 4.2 also includes blue dots representing computation times required to solve the GBE approximately, as discussed in the next section.

4.5.1 Approximate Dynamic Programming Using the GBE

Rather than solving the MSOP (4.15) exactly using the GBE (3.20), as we did in the previous section, we now use an Approximate Dynamic Programming (ADP) (also known as Reinforcement Learning (RL)) algorithm to heuristically solve the MSOP and numerically show these algorithms can result in lower computational times when compared to methods that solve the GBE exactly. This demonstrates that MSOPs with monotonically backward separable cost functions can be heuristically solved using the same methods developed in the ADP literature with the aid of the methodology developed in this chapter.

Typically ADP methods use parametric function fitting (neural networks, linear combinations of basis functions, decision tree's, etc) to approximate the value function (Defn. 3.5) from data. The approximated value function is then used to synthesize a sub-optimal input sequence. To see how this works, suppose an ADP algorithm constructs some approximate value function, denoted $\tilde{V}$, then an approximate optimal input sequence, denoted $\tilde{\mathbf{u}} = (\tilde{u}(0), ..., \tilde{u}(T))$, can be constructed by solving

$$\tilde{u}(k) \in \arg \inf_{u \in \Gamma_{\tilde{x}(k),k}} \left\{ \phi_t(\tilde{x}(k), u, \tilde{V}(f(\tilde{x}(k), u, k), k+1)) \right\} \text{ for } k \in \{0, ..., T-1\}. \tag{4.18}$$

$$\tilde{x}(0) = x_0, \quad \tilde{x}(k+1) = f(\tilde{x}(k), \tilde{u}(k), k) \text{ for } k \in \{0, ..., T-1\}. \tag{4.19}$$

One way to obtain an approximate value function, $\tilde{V}$, is to use the rollout algorithm found in the textbook Bertsekas (1995). This algorithm supposes some base feedback law, $\mu_{base} : \mathbb{R}^n \times \mathbb{N} \to U$, that is "close" to the optimal feedback law is known and approximates the value function as follows. For for some $(x, t) \in X_t \times \{0, ..., T\}$

the approximate value function is given by

$$\begin{aligned}\tilde{V}(x,t) = \phi_t(x(t), u(t), &\phi_{t+1}(x(t+1), u(t+1), ...\phi_T(x(T))...)),\\ &\text{where } u(s) = \mu_{base}(x(s), s) \text{ for all } s \in \{t, ..., T-1\},\\ &\quad x(s+1) = f(x(s), u(s), t) \text{ for all } s \in \{t, ..., T-1\},\\ &\quad x(t) = x.\end{aligned}$$

Using the base policy $\mu_{base}(x,t) = \begin{cases} 1 \text{ if } t/4 \in \mathbb{N} \\ 0.5 \text{ otherwise} \end{cases}$ we used the rollout algorithm to solve the MSOP (4.15) for terminal times $T = 8$ to 10^6. Computation times are plotted as the blue points in Figure 4.2 showing better performance than solving the GBE exactly or using state augmentation. In this case the rollout algorithm yield the optimal input sequence but in general the rollout algorithm may yield suboptimal input sequences.

4.6 Application: Battery Scheduling

We apply our augmented DP methodology (developed in Section 4.2.2) to the scheduling of batteries in the presence of demand and Time-of-Use (TOU) charges and show that our proposed algorithm outperforms existing heuristics, such as Kamyar and Peet (2016) (approximately \$0.98 savings). To do this, we propose a simple model for the dynamics of battery storage. We then formulate the objective/cost function using electricity pricing plans which include demand charges. We then see that the battery scheduling problem can be formulated as an MSOP of the Form (4.5); which can be tractably solved as it has a NFS (Defn. 4.2) objective function. We will solve the battery scheduling problem in the deterministic case based on real historical solar data.

Table 4.1: List of constant values associated with MSOP (4.22) (prices constants correspond to Salt River Project E21 price plan).

Constant	Value	Constant	Value
α	0.999791667 (W/h)	t_{off}	41
η	0.92 (%)	p_{on}	0.0633×10^{-3} (\$/KWh)
$\bar{u}$	4000 (Wh)	p_{off}	0.0423×10^{-3} (\$/KWh)
$\underline{u}$	-4000 (Wh)	p_{d}	0.2973 (\$/KWh)
$\bar{e}$	8000 (Wh)	Δt	0.5 (h)
t_{on}	27		

Battery Dynamics We model the energy stored in the battery using the difference equation:

$$e(k+1) = \alpha(e(k) + \eta u(k)\Delta t), \tag{4.20}$$

where $e(k)$ denotes the energy stored in the battery at time step k, α is the bleed rate of the battery, η is the efficiency of the battery, $u(k)$ denotes the charging/discharging $(+/-)$ at time step k and Δt is the amount of time passed between each time step. Moreover we denote the maximum charge and discharge rate by $\bar{u}$ and $\underline{u}$ respectively. Thus we have the constraint that $u(k) \in [\underline{u}, \bar{u}] := U$ for all k. Similarly we also add the constraint $e(k) \in [\underline{e}, \bar{e}] := X$ for all k where $\underline{e}$ and $\bar{e}$ are the capacity constraints of the battery (typically $\underline{e} = 0$).

The Cost/Objective Function of the Battery Scheduling Problem: Let us denote $q(k)$ as the power supplied by the grid at time step k. Then,

$$q(k) = q_a(k) - q_s(k) + u(k), \tag{4.21}$$

where $q_a(k)$ and $q_s(k)$ are the power consumed by HVAC/appliances and the power supplied by solar photovoltaics at time step k respectively. It is assumed that both $q_a(k)$ and $q_s(k)$ are known a priori.

To define the cost of electricity we divide the day $t \in [0, T]$ into on-peak and off-peak periods. We define an off peak period starting from 12am till t_{on} and t_{off} till 12am. We define an on-peak period between t_{on} till t_{off}. The Time-of-Use (TOU, \$ per kWh) electricity cost during on-peak and off-peak is denoted by p_{on} and p_{off} respectively. We further simplify this as $p_k = p_{on}$ if $k \in T_{on}$ and $p_k = p_{off}$ if $k \in T_{off}$ where T_{on} and T_{off} are the on-peak and off-peak hours, respectively. These TOU charges define the first part of the objective function as:

$$\begin{aligned} J_{\text{TOU}}(\mathbf{u}, \mathbf{e}) &= p_{\text{off}} \sum_{k=0}^{t_{\text{on}}-1} q(k)\Delta t + p_{\text{on}} \sum_{k=t_{\text{on}}}^{t_{\text{off}}-1} q(k)\Delta t + p_{\text{off}} \sum_{k=t_{\text{off}}}^{T} q(k)\Delta t \\ &= \sum_{k\in[0,T]} p_k(q_a(k) - q_s(k))\Delta t + \sum_{k\in[0,T]} p_k u(k)\Delta t \end{aligned}$$

where the daily terminal timestep is $T = 24/\Delta t$.

We also include a demand charge, which is a cost proportional to the maximum rate of power taken from the grid during on-peak times. This cost is determined by p_d which is the price in \$ per kW. Thus it follows the demand charge will be:

$$J_D(\mathbf{u}, \mathbf{e}) = p_d \max_{k\in\{t_{\text{on}},\dots,t_{\text{off}}-1\}} \{q_a(k) - q_s(k) + u(k)\}.$$

4.6.1 Formulating the Battery Scheduling Problem as an MSOP

We may now define the MSOP for the battery scheduling problem in the presence of demand and Time-of-Use charges initialized by $(x_0, t_0) = (e_0, 0)$,

$$\begin{aligned} &\min_{\mathbf{u},\mathbf{e}} \{J_{TOU}(\mathbf{u}, \mathbf{e}) + J_D(\mathbf{u}, \mathbf{e})\} \text{ subject to} \qquad (4.22) \\ &e(k+1) = \alpha(e(k) + \eta u(k)\Delta t) \text{ for } k = 0, ..., T-1 \\ &e_0 = e0\ , e(k) \in [\underline{e}, \bar{e}],\ u(k) \in [\underline{u}, \bar{u}] \text{ for } k = 0, ..., T, \\ &\mathbf{u} = (u(0), ..., u(T-1)) \text{ and } \mathbf{e} = (e(0), ..., e(T)). \end{aligned}$$

Clearly, the MSOP in Eq. (4.22) is of Form (4.1) associated with the tuple $\{J, f, \{X_t\}_{0\le t\le T}, U, T\}$, where $J(\mathbf{u}, \mathbf{x}) = J_{\text{TOU}}(\mathbf{u}, \mathbf{x}) + J_D(\mathbf{u}, \mathbf{x})$, $f(x, u, t) = \alpha(x +$

$\eta u \Delta t)$, $X_t = [\underline{e}, \bar{e}]$, $U = [\underline{u}, \bar{u}]$, $T \in \mathbb{N}$. Now, J is of the form $J(\mathbf{u}, \mathbf{x}) = \max_{0 \le t \le T} d(x(t)) + \sum_{s=0}^{T-1} c_s(x(s), u(s))$ which was shown in Example 4.4 to be NFS (Defn. 4.2). Therefore, $\{J, f, \{X_t\}_{0 \le t \le T}, U, T\} \in \mathcal{M}^{Discrete}_{Forward}$ and thus the MSOP given in Eq. (4.22) can be solved using the state augmentation methods presented in Section 4.2.2.

Numerically Solving the Battery Scheduling Problem: We now solve the battery scheduling problem given in Eq. (4.22) using the methodology of Section 4.2.2 to construct an equivalent "augmented" MSOP of Form (4.7). We then use discretization schemes presented in Section 3.4 to solve this "augmented" MSOP.

We used solar and usage data obtained by local utility Salt River Project (SRP) in Tempe, AZ, for power variables q_s and q_a. We also use pricing data from SRP for the parameters p_{on}, p_{off} and p_d. Battery data obtained for the Tesla Powerwall was used to determine the parameters α, η, $\bar{u}$, $\underline{u}$ and $\bar{e}$. The results of simulating the numerically obtained input sequences are shown in Figure 4.3. The input sequence used for this simulation was created using our augmentation and discretization level of $k = 20$. Interpolation was used to aid in solving Bellman's Equation (3.27) and decrease the approximation error. These results show an improvement in accuracy over results obtained for a similar problem in Kamyar and Peet (2016) (approximately \$0.98 savings). As expected, we see the battery charges during off-peak and then discharges during on peak times to reduce ToU charges, while maintaining a reserve which it uses to keep consumption flat during on peak times, thereby minimizing the demand charge. As a result the power stabilizes during on peak times - becoming constant.

Figure 4.4 shows how the monthly cost decreases when we input sequences constructed from the associated discretized MSOP as the discretization level, $k \in \mathbb{N}$, is increased. Although we do not get a monotonically decreasing sequence of costs,

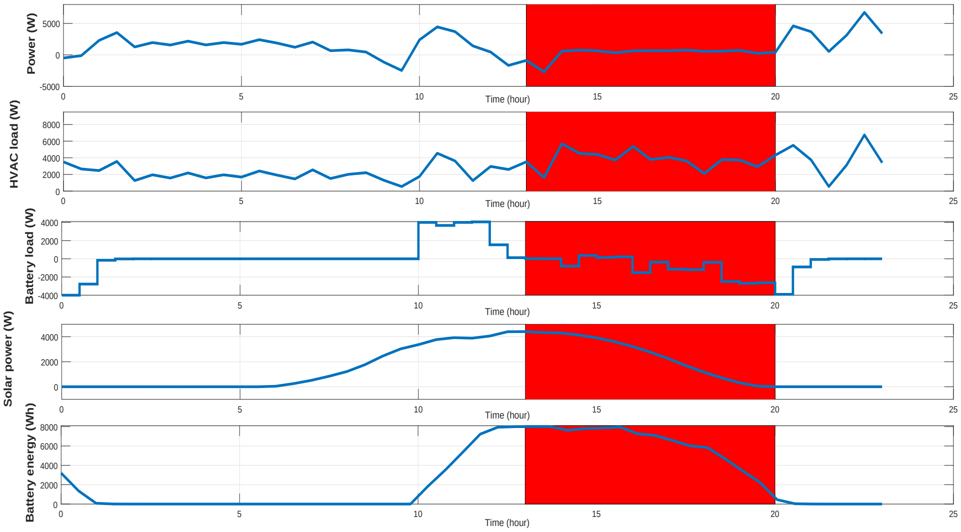

Figure 4.3: Graph showing state simulation from using an input sequence derived from approximately solving the battery scheduling problem with deterministic solar data. The maximum of the power is 0.7033(kw) and the cost is \$46.389.

the error does decrease as $k \to \infty$. Figure 4.5 also shows that augmenting and then following our proposed discretization scheme for the battery scheduling problem results in an input sequence that reduces the consumption demand peak as $k \in \mathbb{N}$ is increased. Figure 4.6 shows how the computational time required to solve the discretized battery scheduling problem appears to be of exponential nature with respect to $k \in \mathbb{N}$.

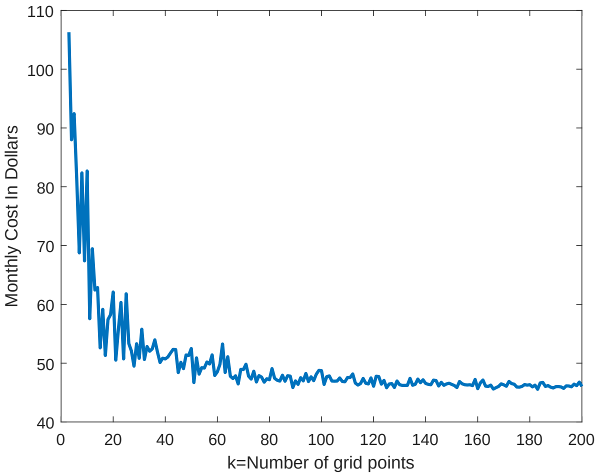

Figure 4.4: The resulting monthly cost from using an input sequence found by solving the discretized problem for optimal battery scheduling.

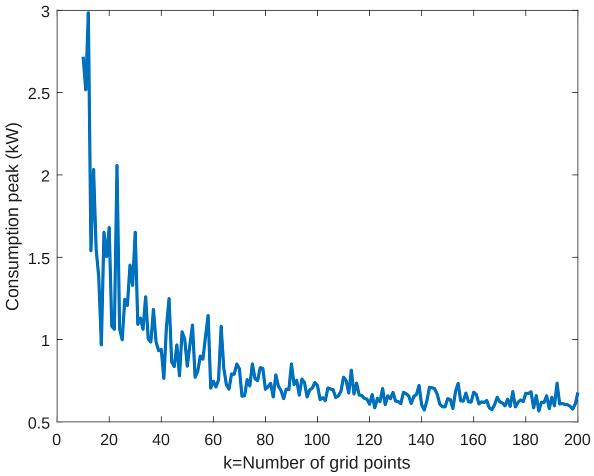

Figure 4.5: The resulting maximum demand from using an input sequence found by solving the discretized problem for optimal battery scheduling.

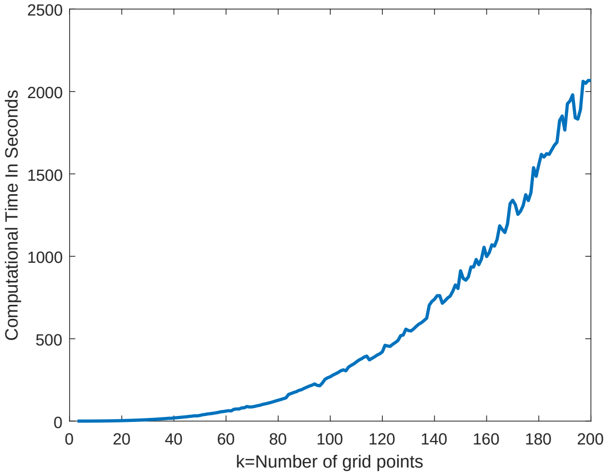

Figure 4.6: The computational time in seconds required to solve the discretized battery scheduling problem.

Part 2

CONTINUOUS TIME

Chapter 5

POLYNOMIAL APPROXIMATIONS OF VALUE FUNCTIONS

> Truth is much too complicated to allow anything but approximations.
>
> John Von Neumann

5.1 Background and Motivation

Consider a nested family of Optimal Control Problems (OCPs), each initialized by $(x_0, t_0) \in \mathbb{R}^n \times [0, T]$, and each an optimization problem of the form

$$(\mathbf{u}^*, x^*) \in \arg\inf_{\mathbf{u},x} \left\{ \int_{t_0}^{T} c(x(t), \mathbf{u}(t), t) dt + g(x(T)) \right\} \text{ subject to,}$$
$$\dot{x}(t) = f(x(t), \mathbf{u}(t)), \;\; \mathbf{u}(t) \in U, \text{ for all } t \in [t_0, T], \; x(t_0) = x_0. \tag{5.1}$$

The problem of solving OCPs (1.4) plays a central role in many practical applications, for instance in the design of non-pharmaceutical interventions in epidemics, see Kantner and Koprucki (2020), optimal train operation, see Khmelnitsky (2000), optimal maintenance strategies for manufacturing systems, see Huang *et al.* (2018), etc.

Solving OCPs directly can be challenging. Fortunately, the problem of solving a family of OCPs (1.4) can be reduced to the problem of solving a Partial Differential Equation (PDE), see Liberzon (2011). From the principle of optimality, if $(\mathbf{u}^*, x^*)$ solve the OCP for (x_0, t_0), then $(\tau_t \mathbf{u}^*, x^*(t))$ (where $\tau_t \mathbf{u}^*(s) = \mathbf{u}^*(t + s)$ for all $s \geq 0$) solves the OCP initialized at $(x^*(t), t)$ for any $t \in [t_0, T]$. This can be used to show that if a function, V, satisfies the Hamilton Jacobi Bellman (HJB) Partial Differential Equation (PDE), defined as

$$\nabla_t V(x,t) + \inf_{u \in U} \left\{ c(x,u,t) + \nabla_x V(x,t)^T f(x,u) \right\} = 0 \text{ for all } (x,t) \in \mathbb{R}^n \times (0,T),$$
$$V(x,T) = g(x) \quad \text{for all } x \in \mathbb{R}^n, \tag{5.2}$$

then necessary and sufficient conditions for $(\mathbf{u}^*, x^*)$ to solve OCP (5.1) initialized by (x_0, t_0) are

$$\mathbf{u}^*(t) = k(x^*(t), t), \ \dot{x}^*(t) = f(x^*(t), \mathbf{u}^*(t)), \quad \text{and} \quad x^*(t_0) = x_0,$$
$$\text{where } \ k(x,t) \in \arg \inf_{u \in U} \left\{ c(x,u,t) + \nabla_x V(x,t)^T f(x,u) \right\}. \tag{5.3}$$

For a given family of OCPs of Form (5.1), if V satisfies Eq. (5.2), then V is called the Value Function (VF) of the OCP. If V is the VF, then for any (x,t), the value $V(x,t)$ determines the optimal objective value of OCP (5.1) initialized by (x,t). Furthermore, the VF yields a solution to the OCP (5.1) initialized by (x_0, t_0) through application of Eq. (5.3). We call any $k : \Omega \times [0,T] \to U$ that satisfies Eq. (5.3) a controller and we say this controller is the optimal controller for the OCP when V is the VF of the OCP.

Thus knowledge of the VF allows us to solve the nested family of OCPs in (5.1). Unfortunately, to find the VF, we must solve the HJB PDE, given in Eq. (5.2), and this PDE has no analytic solution. In the absence of an analytic solution, we often parameterize a family of candidate VFs and search for one which satisfies the HJB PDE. However, this is a non-convex optimization problem since the HJB PDE is nonlinear. In this chapter we view the search for a VF through the lens of convex optimization. Moreover, given an OCP, we are particularly interested in computing a sub-VF, a function that is uniformly less than or equal to the VF of the OCP (ie a function $\tilde{V}$ such that $\tilde{V}(x,t) \leq V(x,t)$ for all $(x,t) \in \mathbb{R}^n \times [0,T]$ where V is the VF of the OCP). We consider what happens when we relax the nonlinear equality constraints imposed by the HJB PDE to linear inequality constraints and tighten the optimization problem's feasible set to polynomials. In this chapter we consider the following question:

Q1: Can we pose a sequence of convex optimization problems, each yielding a polynomial sub-VF that can be made arbitrarily "close" to the VF of the OCP?

Over the years, many numerical methods have been proposed for solving the HJB for a given OCP. Within this literature, a substantial number of the algorithms are based on a finite-dimensional projection of the spatial domain (gridding/meshing/discretization of the state space). In this class of algorithms we include (mixed) finite elements methods - an important example of which is Gallistl *et al.* (2020). Specifically, the approach in Gallistl *et al.* (2020) yields an approximate VF with an error bound on the first order mixed L^2 norm - a bound which converges as the number of elements is increased (assuming the Cordes condition holds). Other examples of this class of methods include the discretization approaches in Achdou *et al.* (2008); Kalise and Kunisch (2018). For example, in Achdou *et al.* (2008), we find an algorithm which yields an approximate VF with an L^∞ error bound which converges as the level of discretization increases. Alternative non-grid based algorithms include the method of characteristics found in Liberzon (2011), which can be used to compute evaluations of VF at fixed $(x, t) \in \mathbb{R}^n$, and max-plus methods found in McEneaney (2007). The result in McEneaney (2007) considers an OCP with linear dynamics and a cost function which is the point-wise maximum of quadratic functions. This max-plus approach yields an approximate VF with a converging error bound which holds on $x \in \mathbb{R}^n$, but increases with $|x|$.

While all of these numerical methods yield approximate VFs with associated approximation error bounds, the use of these functions for controller synthesis (see **Q2**) and reachable set estimation has been more limited (the connection between VFs, the HJB and reachable sets was made in Mitchell *et al.* (2005)). This is due to the fact that the approximate VFs obtained from such discretization methods are difficult to manipulate and apart from being close to the true VF, have relatively few provable properties (such as being uniformly less than or greater to the true VF ie being sub or super-VFs). Being a sub or super-VF is an important property of any approximate VF. As shown in Cor. 5.1, sub/super-VFs can yield outer bounds on reachable sets that can be used to certify that the underlying system does not transition into regions of the state space deemed unsafe; a useful tool in the safety analysis of dynamical systems.

To address these issues, in this chapter we focus on obtaining approximate VFs which are both polynomial and sub-VFs. Specifically, the use of polynomials ensures that the derivative of the approximated VFs can be efficiently computed (a useful property for solving the controller synthesis Eq. (5.3)), while the use of sub-VFs ensures that sublevel sets of the VF are guaranteed to contain the sublevel set of the true VF (see Cor. 5.1), and hence provide provable guarantees on the boundary of the reachable set (a useful property for safety analysis).

Substantial work on SOS relaxations of the HJB PDE for reachable set estimation and safety set analysis includes the carefully constructed optimization problems in Summers *et al.* (2013); Yin *et al.* (2018); Xue *et al.* (2019); Zhang *et al.* (2019) and includes, of course, our work in Jones and Peet (2019b,c). Such SOS relaxations of the HJB PDE can yield approximate VFs. However, there seems to be no prior work on using approximation theory to prove bounds on the sub-optimality of either controllers (see **Q2**) or corresponding reachable sets constructed from such approximated VFs. We note, however, that Xue *et al.* (2019) did establish the *existence* of a polynomial sub-solution to the HJB arbitrarily close to the true solution of the HJB in the framework of reachable sets. Treatments of the moment-based alternatives to the SOS approach includes Kamoutsi *et al.* (2017); Pakniyat and Vasudevan (2019); Korda *et al.* (2016); Zhao *et al.* (2017). Another duality-based approach, found in Chen and Ames (2019), considers a density-based dual to the VF and uses finite elements method to iteratively approximate the density and VF.

In this chapter we answer **Q1** by considering "sub-solutions" to the HJB PDE (5.2). Specifically, a "sub-solution", $\tilde{V}$, to the HJB PDE (5.2) satisfies the relaxed inequality constraint

$$\nabla_t \tilde{V}(x,t) + c(x,u,t) + \nabla_x \tilde{V}(x,t)^T f(x,u) \geq 0 \tag{5.4}$$

for all $u \in U$ and $(x,t) \in \mathbb{R}^n \times [0,T]$, which implies that if V is a VF, $\tilde{V}(x,t) \leq V(x,t)$ - i.e. $\tilde{V}$ is a sub-VF. Then given an OCP (5.1) and based on this relaxed version of the HJB PDE (5.4), we propose a sequence of SOS programming problems, indexed by the degree $d \in \mathbb{N}$ of the polynomial variables, and given in Eq. (5.59). The solution to each instance of the proposed sequence of optimization problems yields a polynomial P_d that is a sub-solution to the HJB PDE (5.2) (or sub-VF). We then show in Prop. 5.4 that for any VF V associated with the given OCP we have,

$$\lim_{d \to \infty} \|P_d - V\|_{L^1} = 0.$$

Furthermore, in Prop. 5.5 we show that this implies that the sublevel sets of $\{P_d\}_{d \in \mathbb{N}}$ converge to the sublevel sets of any VF, V, of the OCP (respect to the volume metric).

Our proposed method of approximately solving the HJB PDE by solving an SOS programming problem is implemented via Semi-Definite Programming (SDP). SDP problems can be solved to arbitrary accuracy in polynomial time using interior point methods, see Vandenberghe and Boyd (1996). However, the number of variables in the SDP problem associated with an n-dimensional and d-degree SOS problem is of the order n^d, see Ahmadi and Majumdar (2019), and therefore exponentially increases as $d \to \infty$. Fortunately there exist several methods that improve the scalability of SOS found in Ahmadi and Majumdar (2019); Zheng *et al.* (2019) but we do not discus such methods in this chapter.

5.2 Optimal Control Problems

The nested family of finite-time Optimal Control Problems (OCPs), each initialized by $(x_0, t_0) \in \mathbb{R}^n \times [0, T]$, are defined as:

$$\begin{aligned}
&(\mathbf{u}^*, x^*) \in \arg\inf_{\mathbf{u},x} \left\{ \int_{t_0}^{T} c(x(t), \mathbf{u}(t), t) dt + g(x(T)) \right\} \text{ subject to,} \\
&\dot{x}(t) = f(x(t), \mathbf{u}(t)) \quad \text{ for all } t \in [t_0, T], \qquad (5.5)\\
&(x(t), \mathbf{u}(t)) \in \Omega \times U \quad \text{ for all } t \in [t_0, T], \quad x(t_0) = x_0,
\end{aligned}$$

where $c : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ is referred to as the running cost; $g : \mathbb{R}^n \to \mathbb{R}$ is the terminal cost; $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is the vector field; $\Omega \subset \mathbb{R}^n$ is the state constraint set; $U \subset \mathbb{R}^m$ is the input constraint set; and T is the final time. For a given family of OCPs of Form (5.5) we associate the tuple $\{c, g, f, \Omega, U, T\}$.

In this chapter we consider a special class of OCPs of Form (5.5), where U is compact and c, g, f are locally Lipschitz continuous. We next recall the definition of local Lipschitz continuity.

Definition 5.1. *Consider sets $\Theta_1 \subset \mathbb{R}^n$ and $\Theta_2 \subset \mathbb{R}^m$. We say the function $F : \Theta_1 \to \Theta_2$ is **locally Lipschitz continuous** on Θ_1 and Θ_2, denoted $F \in LocLip(\Theta_1, \Theta_2)$, if for every compact set $X \subseteq \Theta_1$ there exists $K_X > 0$ such that for all $x, y \in X$*

$$||F(x) - F(y)||_2 \leq K_X ||x - y||_2. \qquad (5.6)$$

*If there exists $K > 0$ such that Eq. (5.6) holds for all $x, y \in \Theta_1$ we say F is **uniformly Lipschitz continuous**, denoted $F \in Lip(\Theta_1, \Theta_2)$.*

Definition 5.2. *We say the six tuple* $\{c, g, f, \Omega, U, T\}$ *is a Family of Lipschitz OCPs of Form* (5.5) *or* $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$ *if:*

1. $c \in LocLip(\Omega \times U \times [0, T], \mathbb{R})$.

2. $g \in LocLip(\Omega, \mathbb{R})$.

3. $f \in LocLip(\Omega \times U, \mathbb{R})$.

4. $U \subset \mathbb{R}^m$ *is compact.*

For $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$, if $\Omega = \mathbb{R}^n$ we say the family of associated OCPs is *state unconstrained*, and if $\Omega \neq \mathbb{R}^n$ we say the associated family of OCPs is *state constrained.*

5.3 Value Functions can Solve OCPs

In the following subsections, we establish that for every family of Lipschitz OCPs, as defined in Section 5.2, there exists a function, called the Value Function (VF), which:

(A) Is determined by the solution map - Eq. (5.10).

(B) Solves the Hamilton-Jacobi-Bellman (HJB) Partial Differential Equation (PDE) - Eq. (5.12).

(C) Can be used to construct a solution to the OCP.

5.3.1 Value Functions are Determined by the Solution Map

Consider a nonlinear Ordinary Differential Equation (ODE) of the form

$$\dot{x}(t) = f(x(t), \mathbf{u}(t)), \quad x(0) = x_0, \tag{5.7}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $\mathbf{u} : \mathbb{R} \to \mathbb{R}^m$, and $x_0 \in \mathbb{R}^n$.

Definition 5.3. *We say the function ϕ_f is a solution map of the ODE given in Eq.* (5.7) *on* $[0,T] \subset \mathbb{R}$ *if for all* $t \in [0,T]$

$$\frac{\partial \phi_f(x_0,t,\mathbf{u})}{\partial t} = f(\phi_f(x_0,t,\mathbf{u}),\mathbf{u}(t)), \text{ and } \phi_f(x_0,0,\mathbf{u}) = x_0.$$

Definition of Admissible Inputs: Given $\{c,g,f,\Omega,U,T\} \in \mathcal{M}_{Lip}^{Continuous}$ and associated family of OCPs of Form (5.5), we now use the solution map to define the set of admissible input signals for the OCP initialized at $(x_0,t_0) \in \Omega \times [0,T]$. For this we use the shift operator, denoted $\tau_s : L^2([0,T],\mathbb{R}^m) \to L^2([0,T-s],\mathbb{R}^m)$, where $s \in [0,T]$, and defined by

$$(\tau_s \mathbf{u})(t) := \mathbf{u}(s+t) \text{ for all } t \in [0,T-s]. \tag{5.8}$$

Definition 5.4. *For any* $(x_0,t_0) \in \mathbb{R}^n \times [0,T]$*, we say* $\mathbf{u}$ *is admissible, denoted* $\mathbf{u} \in \mathcal{U}_{\Omega,U,f,T}(x_0,t_0)$*, if* $\mathbf{u} : [t_0,T] \to U$ *and there exists a unique solution map,* ϕ_f*, such that*

$$\frac{\partial \phi_f(x_0,t-t_0,\tau_{t_0}\mathbf{u})}{\partial t} = f(\phi_f(x_0,t-t_0,\tau_{t_0}\mathbf{u}),\mathbf{u}(t)) \text{ for } t \in [t_0,T],$$
$$\phi_f(x_0,t-t_0,\tau_{t_0}\mathbf{u}) \in \Omega \text{ for } t \in [t_0,T], \text{ and } \phi_f(x_0,0,\tau_{t_0}\mathbf{u}) = x_0. \tag{5.9}$$

For a given family of OCPs of Form (5.5), we now define the associated VF using the solution map, ϕ_f. Lemma 5.1 then shows that VFs are locally Lipschitz continuous.

Definition 5.5. *For given* $\{c,g,f,\Omega,U,T\} \in \mathcal{M}_{Lip}^{Continuous}$ *we say* $V^* : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ *is a Value Function (VF) of the associated family of OCPs if for* $(x,t) \in \Omega \times [0,T]$*, the following holds*

$$V^*(x,t) = \inf_{\mathbf{u} \in \mathcal{U}_{\Omega,U,f,T}(x,t)} \Bigg\{ \tag{5.10}$$
$$\int_t^T c(\phi_f(x,s-t,\tau_t\mathbf{u}),\mathbf{u}(s),s)ds + g(\phi_f(x,T-t,\tau_t\mathbf{u}))\Bigg\},$$

where ϕ_f *is as in Eq.* (5.9)*. By convention if* $\mathcal{U}_{\Omega,U,f,T}(x,t) = \emptyset$ *then* $V^*(x,t) = \infty$.

Lemma 5.1 (Local Lipschitz continuity of VFs, see Bressan (2011))**.** *Consider some* $\{c, g, f, \mathbb{R}^n, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$*. Then if* V^* *satisfies Eq.* (5.10)*, we have that* $V^* \in LocLip(\mathbb{R}^n \times [0,T], \mathbb{R})$*.*

5.3.2 Value Functions are Solutions to the HJB PDE

Consider the family of OCPs associated with $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$. As shown in Bertsekas (1995), a sufficient condition for a function V^* to be a VF, is for V^* to satisfy the Hamilton Jacobi Bellman (HJB) PDE, given in Eq. (5.12). However, for a general family of OCPs of form $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$, solutions to the HJB PDE may not be differentiable, and hence classical solutions to the HJB PDE may not exist. For this reason, one typically uses a generalized notion of a solution to the HJB PDE called a viscosity solution, which is defined in Crandall (1997) as follows.

Definition 5.6. *Consider the first order PDE*

$$F(x, y(x), \nabla y(x)) = 0 \quad \textit{for all } x \in \Omega, \tag{5.11}$$

where $\Omega \subset \mathbb{R}^n$ *and* $F \in C(\Omega \times \mathbb{R} \times \mathbb{R}^n, \mathbb{R})$*.*

We say $y \in C(\Omega)$ *is a* ***viscosity sub-solution*** *of the PDE* (5.11) *if*

$$F(x, y(x), p) \leq 0 \quad \textit{for all } x \in \Omega \textit{ and } p \in D^+ y(x),$$

where $D^+ y(x) := \{p \in \mathbb{R} :$ *there exists* $\Phi \in C^1(\Omega, \mathbb{R})$ *such that*

$$\nabla \Phi(x) = p \textit{ and } y - \Phi \textit{ attains a local max at } x\}.$$

Similarly, $y \in C(\Omega)$ *is a* ***viscosity super-solution*** *of the PDE* (5.11) *if*

$$F(x, y(x), p) \geq 0 \quad \textit{for all } x \in \Omega \textit{ and } p \in D^- y(x)$$

where $D^- y(x) := \{p \in \mathbb{R} :$ *there exists* $\Phi \in C^1(\Omega, \mathbb{R})$ *such that*

$$\nabla \Phi(x) = p \textit{ and } y - \Phi \textit{ attains a local min at } x\}.$$

We say $y \in C(\Omega)$ is a ***viscosity solution*** *of* (5.11) *if it is both a viscosity sub and super-solution.*

Theorem 5.1 (Uniqueness of VFs, see Bressan (2011))**.** *Consider the family of OCPs associated with the tuple $\{c, g, f, \mathbb{R}^n, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$. Any function satisfying Eq.* (5.10) *is the unique viscosity solution of the HJB PDE*

$$\nabla_t V(x,t) + \inf_{u \in U} \left\{ c(x,u,t) + \nabla_x V(x,t)^T f(x,u) \right\} = 0 \textit{ for all } (x,t) \in \mathbb{R}^n \times [0,T]$$
$$V(x,T) = g(x) \quad \textit{for all } x \in \mathbb{R}^n. \tag{5.12}$$

Note that Lemma 5.1 and Theorem 5.1 are only valid in the absence of state constraints (the case when $\Omega = \mathbb{R}^n$). However, as we will show in Lemma 5.3, if the state constraints are sufficiently "loose", then the unconstrained and constrained solutions coincide.

5.3.3 VFs can Construct Optimal Controllers

Given an OCP, we next show if a "classical" differentiable solution to the HJB PDE (5.12) associated with the OCP is known then a solution to the OCP can be constructed using Eqs. (5.13) and (5.14). We will refer to any $k : \Omega \times [0,T] \to U$ that satisfies Eqs. (5.13) and (5.14) for some V as a controller and say this is the optimal controller of the OCP if V is the VF of the OCP.

Theorem 5.2 (Liberzon (2011))**.** *Consider the family of OCPs associated with tuple $\{c, g, f, \mathbb{R}^n, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$. Suppose $V \in C^1(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ solves the HJB PDE* (5.12)*. Then $\mathbf{u}^* : [t_0, T] \to U$ solves the OCP associated with $\{c, g, f, \mathbb{R}^n, U, T\}$ initialized at $(x_0, t_0) \in \mathbb{R}^n \times [0,T]$ if and only if*

$$\mathbf{u}^*(t) = k(\phi_f(x_0, t, \mathbf{u}^*), t) \textit{ for all } t \in [t_0, T], \tag{5.13}$$

$$\textit{where } k(x,t) \in \arg \inf_{u \in U} \{ c(x,u,t) + \nabla_x V(x,t)^T f(x,u) \}. \tag{5.14}$$

If the function V in Eq. (5.14) is not a VF the resulting controller may no longer construct a solution to the OCP. In Chapter 6 we will provide a bound on the performance of a constructed controller from a candidate VF based on how "close" the candidate VF is to the true VF under the Sobolev norm.

5.4 The Feasibility Problem of Finding VFs

Consider a family of OCPs associated with some $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$. Previously it was shown in Theorem 5.2 that if $V \in C^1(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ is a solution to the HJB PDE (5.12) then V may be used to solve the family of OCPs using Eqs. (5.13) and (5.14). The question, now, is how to find such a V.

Let us consider the problem of finding a value function as an optimization problem subject to constraints imposed by the HJB PDE (5.12). This yields the following feasibility problem:

$$\begin{aligned} &\text{Find } V \in C^1(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}), \\ &\qquad \text{such that } V \text{ satisfies (5.12).} \end{aligned} \tag{5.15}$$

Note that our optimization problem of Form (5.15) is non-convex and may not even have a solution with sufficient regularity. For these reasons, we next propose a convex relaxation of Problem (5.15). We first define sub-VFs and super-VFs that uniformly bound VFs either from above or bellow.

Definition 5.7. *We say the function $J : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ is a sub-VF to the family of OCPs associated with $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$ if*

$$J(x,t) \le V^*(x,t) \textit{ for all } t \in [0,T] \textit{ and } x \in \Omega,$$

for any V^* satisfying Eq.(5.10). Moreover if

$$J(x,t) \ge V^*(x,t) \textit{ for all } t \in [0,T] \textit{ and } x \in \Omega,$$

for any V^* satisfying Eq. (5.10)*, we say J is a super-VF.*

5.4.1 A Sufficient Condition for a Function to be a Sub-VF

We now propose "dissipation" inequalities, given in Eqs. (5.16) and (5.17), and show that if a differentiable function satisfies such inequalities then it must be a sub-value function.

Proposition 5.1. *For given $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$ suppose $J \in C^1(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ satisfies for all $(x, u, t) \in \Omega \times U \times (0, T)$*

$$\nabla_t J(x,t) + c(x,u,t) + \nabla_x J(x,t)^T f(x,u) \geq 0, \tag{5.16}$$

$$J(x,T) \leq g(x). \tag{5.17}$$

Then J is a sub-value function of the family of OCPs associated with $\{c, g, f, \Omega, U, T\}$.

Proof. Suppose $J \in C^1(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ satisfies Eqs. (5.16) and (5.17). Consider an arbitrary $(x_0, t_0) \in \Omega \times [0, T]$. If $\mathcal{U}_{\Omega,U,f,T}(x_0, t_0) = \emptyset$ then $V^*(x_0, t_0) = \infty$. Clearly in this case $J(x_0, t_0) < V^*(x_0, t_0)$ as J is continuous and therefore is finite over the compact region $\Omega \times [0, T]$. Alternatively if $\mathcal{U}_{\Omega,U,f,T}(x_0, t_0) \neq \emptyset$, then for any $\tilde{\mathbf{u}} \in \mathcal{U}_{\Omega,U,f,T}(x_0, t_0)$, we have the following by Defn. 5.4:

$$\phi_f(x_0, t - t_0, \tau_{t_0}\tilde{\mathbf{u}}) \in \Omega \text{ for all } t \in [t_0, T] \text{ and } \tilde{\mathbf{u}}(t) \in U \text{ for all } t \in [t_0, T].$$

Therefore (using the shorthand $\tilde{x}(t) := \phi_f(x_0, t - t_0, \tau_{t_0}\tilde{\mathbf{u}})$), by Eq. (5.16) we have for all $t \in [t_0, T]$

$$\nabla_t J(\tilde{x}(t), t) + c(\tilde{x}(t), \tilde{\mathbf{u}}(t), t) + \nabla_x J(\tilde{x}(t), t)^T f(\tilde{x}(t), \tilde{\mathbf{u}}(t)) \geq 0.$$

Now, using the chain rule we deduce

$$\frac{d}{dt} J(\tilde{x}(t), t) + c(\tilde{x}(t), \tilde{\mathbf{u}}(t), t) \geq 0 \text{ for all } t \in [t_0, T].$$

Then, integrating over $t \in [t_0, T]$, and since $J(\tilde{x}(T), T) \leq g(\tilde{x}(T))$ by Eq. (5.17), we have

$$J(x_0, t_0) \leq \int_{t_0}^{T} c(\tilde{x}(t), \tilde{\mathbf{u}}(t), t)dt + g(\tilde{x}(T)). \tag{5.18}$$

Since Eq. (5.18) holds for all $\tilde{\mathbf{u}} \in \mathcal{U}_{\Omega,U,f,T}(x_0, t_0)$, we may take the infimum over $\mathcal{U}_{\Omega,U,f,T}(x_0, t_0)$ to show that $J(x_0, t_0) \leq V^*(x_0, t_0)$. As this argument can be used for any $(x_0, t_0) \in \Omega \times [0, T]$ it follows J is a sub-value function. □

Definition 5.8. *For given* $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$ *we say a function* $J \in C^1(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ *is dissipative if it satisfies Inequalities* (5.16) *and* (5.17)*.*

Dissipative functions are viscosity sub-solutions (as per Defn. 5.6) to the HJB PDE (5.12). Moreover, by Prop. 5.1 a dissipative function is a sub-VF. However, a sub-VF need not be dissipative or a viscosity sub-solution to the HJB PDE.

5.4.2 A Convex Relaxation of the Problem of Finding VFs

The set of functions satisfying Eqs. (5.16) and (5.17) is convex as Eqs. (5.16) and (5.17) are linear in terms of the unknown variable/function J. Furthermore, for given $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$, any function which satisfies the HJB PDE (5.12) also satisfies Eqs. (5.16) and (5.17). This allows us to propose the following convex relaxation of the problem of finding a VF (Problem (5.15)):

$$\begin{aligned} &\text{Find } J \in C^1(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}), \\ &\qquad \text{such that } J \text{ satisfies (5.16) and (5.17).} \end{aligned} \tag{5.19}$$

5.4.3 A Polynomial Tightening of the Problem of Finding VFs

Problem (5.19) is convex. However, a function J, feasible for Problem (5.19) (and hence dissipative), may be arbitrarily far from the VF. For instance, in the case

$c(x,u,t) \geq 0$ and $0 \leq g(x) < M$, the constant function $J(x,t) \equiv -C$ is dissipative for any $C > M$. Thus, by selecting sufficiently large enough $C > M$, we can make $||J - V||$ arbitrary large, regardless of the chosen norm, $|| \cdot ||$.

To address this issue, we propose a modification of Problem (5.19), wherein we include an objective of Form $\int_{\Lambda \times [0,T]} w(x,t)J(x,t)dxdt$, parameterized by a compact domain of interest $\Lambda \subset \mathbb{R}^n$ and weight $w \in L^1(\Lambda \times [0,T], \mathbb{R}^+)$ (we use the weight, w, in Prop. 5.5). Specifically, for given $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$ and $d \in \mathbb{N}$, consider the optimization problem:

$$\begin{aligned} J_d \in \arg \max_{J \in \mathcal{P}_d(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})} & \int_{\Lambda \times [0,T]} w(x,t)J(x,t)dxdt \qquad (5.20) \\ \text{subject to:} & \nabla_t J(x,t) + c(x,u,t) + \nabla_x J(x,t)^T f(x,u) > 0 \text{ for } x \in \Omega, t \in (0,T), u \in U, \\ & J(x,T) < g(x) \text{ for all } x \in \Omega. \end{aligned}$$

Maximizing $\int_{\Lambda \times [0,T]} w(x,t)J(x,t)dxdt$ minimizes the weighted L^1 norm $\int_{\Lambda \times [0,T]} w(x,t)|V(x,t) - J(x,t)|dxdt$. The restriction to polynomial solutions $J \in \mathcal{P}_d(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ makes the problem finite-dimensional.

5.5 A Sequence of Dissipative Polynomials that Converge to the VF in Sobolev Space

For a given $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$, in Eq. (5.20), we proposed a sequence of optimization problems, indexed by $d \in \mathbb{N}$, each instance of which yields a dissipative function $J_d \in \mathcal{P}_d(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$. In this section, we prove that $\lim_{d \to \infty} ||J_d - V||_{L^1(\Lambda \times (0,T), \mathbb{R})} \to 0$ where V is the VF associated with the OCP $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$. To accomplish this proof, we divide the section into three subsections, wherein we find the following.

(A) In Prop. 5.2 we show that for any $V \in Lip(\Omega \times [0,T], \mathbb{R})$ that satisfies the dissipation-type inequality in Eq. (5.21) and any $\varepsilon > 0$ there exists a dissipative function $J_\varepsilon \in C^\infty(\Omega \times [0,T], \mathbb{R})$ such that $||J_\varepsilon - V||_{W^{1,p}(\Omega\times[0,T],\mathbb{R})} < \varepsilon$.

(B) In Theorem 5.3 we show that for every $\varepsilon > 0$, there exists $d \in \mathbb{N}$ and dissipative $P_\varepsilon \in \mathcal{P}_d(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ such that $||P_\varepsilon - V||_{W^{1,p}(\Omega\times[0,T],\mathbb{R})} < \varepsilon$, for any value function, V, associated with $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$.

(C) For any positive weight w, Prop. 5.3 shows that if J_d solves (5.20) for $d \in \mathbb{N}$, then $\lim_{d\to\infty} ||w(J_d - V)||_{L^1(\Lambda\times(0,T),\mathbb{R})} = 0$ for any VF, V, associated with $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$.

5.5.1 Existence of Smooth Dissipative Functions that Approximate the VF Arbitrarily well under the $W^{1,p}$ Norm

In this section we create a sequence of smooth (elements of $C^\infty(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$) functions that converges, with respect to the $W^{1,p}$ norm, to any Lipschitz function, V, satisfying the dissipation-type inequality in Eq. (5.21). This subsection uses some aspects of mollification theory found in Appendix B.

Approximation of Lipschitz functions satisfying a dissipation-type inequality We now show that for any Lipschitz function, V, satisfying the dissipation-type inequality in Eq. (5.21), V can be approximated arbitrarily well by a smooth function, J_ε, that also satisfies the dissipation-type inequality in Eq. (5.21). We use a similar proof strategy first appearing in Kurzwel (1963) and also later appearing in Wilson (1969); Lin *et al.* (1996); Teel and Praly (2000).

Lemma 5.2. *Let $E \subset \mathbb{R}^{n+1}$ be an open bounded set, $\Omega \subset \mathbb{R}^n$ be such that $\Omega \times (0,T) \subseteq E$, where $T > 0$, $U \subset \mathbb{R}^m$ be a compact set, $f \in Lip(\Omega \times U, \mathbb{R}^n)$, $c \in Lip(\Omega \times U \times [0,T], \mathbb{R})$, and $V \in Lip(E, \mathbb{R})$ such that*

$$\operatorname*{ess\,inf}_{(x,t)\in\Omega\times(0,T)}\{\nabla_t V(x,t)+\nabla_x V(x,t)^T f(x,u)+c(x,u,t)\}\geq 0, \tag{5.21}$$

where the derivatives, $\nabla_t V$ and $\nabla_x V$, are weak derivatives.

Then for any compact set $K \subset E$, $1 \leq p < \infty$ and for all $\varepsilon > 0$ there exits $J_\varepsilon \in C^\infty(K,\mathbb{R})$ such that

$$||V - J_\varepsilon||_{W^{1,p}(K,\mathbb{R})} < \varepsilon \textit{ and } \sup_{(x,t)\in K}|V(x,t) - J_\varepsilon(x,t)| < \varepsilon, \tag{5.22}$$

and for all $(x,t) \in K \cap (\Omega \times (0,T))$ and $u \in U$

$$\nabla_t J_\varepsilon(x,t) + \nabla_x J_\varepsilon(x,t)^T f(x,u) + c(x,u,t) \geq -\varepsilon. \tag{5.23}$$

Proof. Suppose V satisfies Eq. (5.21), $K \subset E$ is a compact set, $1 \leq p < \infty$, and $\varepsilon > 0$. By Rademacher's Theorem (Theorem C.4) V is weakly differentiable with essentially bounded derivative. Therefore $V \in W^{1,\infty}(E,\mathbb{R})$ and hence $V \in W^{1,p}(E,\mathbb{R})$. Now Prop. B.1 (Statements 3 and 4) can be used to show there exists $\sigma_1 > 0$ such that for any $0 \leq \sigma < \sigma_1$ we have

$$||V - [V]_{\sigma_1}||_{W^{1,p}(K,\mathbb{R})} < \varepsilon \text{ and } \sup_{(x,t)\in K}|V(x,t) - [V]_{\sigma_1}(x,t)| < \varepsilon. \tag{5.24}$$

Select $\sigma_2 > 0$ small enough so $K \subset < E >_{\sigma_2}$ (which can be done as E is open). Select $0 < \sigma_3 < \frac{\varepsilon}{L_V L_f + 2L_c}$, where $L_V, L_f, L_c > 0$ are the Lipschitz constant of the functions V, f, and c respectively. We now have the following for all $\sigma_4 < \min\{\sigma_3, \sigma_2\}$, $u \in U$

and $(x,t) \in K \cap (\Omega \times (0,T))$,

$$\begin{aligned}
&\nabla_t [V]_{\sigma_4}(x,t) + \nabla_x [V]_{\sigma_4}(x,t)^T f(x,u) + c(x,u,t) \qquad (5.25)\\
&= [\nabla_t V]_{\sigma_4}(x,t) + [\nabla_x V]_{\sigma_4}(x,t)^T f(x,u) + c(x,u,t)\\
&= \int_{B_{\sigma_4}(0)} \eta_{\sigma_4}(z_1,z_2) \bigg(\nabla_t V(x-z_1, t-z_2)\\
&\qquad\qquad + \nabla_x V(x-z_1,t-z_2)^T f(x-z_1,u) + c(x-z_1,u,t-z_2) \bigg) dz_1 dz_2\\
&\quad - \int_{B_{\sigma_4}(0)} \eta_{\sigma_4}(z_1,z_2) \nabla_x V(x-z_1,t-z_2)^T \bigg(f(x-z_1,u) - f(x,u) \bigg) dz_1 dz_2\\
&\quad - \int_{B_{\sigma_4}(0)} \eta_{\sigma_4}(z_1,z_2) \bigg(c(x-z_1,u,t-z_2) - c(x,u,t) \bigg) dz_1 dz_2\\
&\geq \operatorname*{ess\,inf}_{(z_1,z_2)\in B_{\sigma_4}(0)} \bigg\{ \nabla_t V(x-z_1,t-z_2)\\
&\qquad\qquad + \nabla_x V(x-z_1,t-z_2)^T f(x-z_1,u) + c(x-z_1,u,t-z_2) \bigg\}\\
&\quad - \operatorname*{ess\,sup}_{(z_1,z_2)\in B_{\sigma_4}(0)} \bigg\{ ||\nabla_x V(x-z_1,t-z_2)||_2 \bigg\} \operatorname*{ess\,sup}_{(z_1,z_2)\in B_{\sigma_4}(0)} \bigg\{ ||f(x-z_1,u) - f(x,u)||_2 \bigg\}\\
&\quad - \operatorname*{ess\,sup}_{(z_1,z_2)\in B_{\sigma_4}(0)} \bigg\{ |c(x-z_1,u,t-z_2) - c(x,u,t)| \bigg\}\\
&\geq -L_V \operatorname*{ess\,sup}_{(z_1,z_2)\in B_{\sigma_4}(0)} \bigg\{ ||f(x-z_1,u) - f(x,u)||_2 \bigg\}\\
&\quad - \operatorname*{ess\,sup}_{(z_1,z_2)\in B_{\sigma_4}(0)} \bigg\{ |c(x-z_1,u,t-z_2) - c(x,u,t)| \bigg\}\\
&\geq -L_V L_f \operatorname*{ess\,sup}_{(z_1,z_2)\in B_{\sigma_4}(0)} \bigg\{ ||z_1||_2 \bigg\} - L_c \operatorname*{ess\,sup}_{(z_1,z_2)\in B_{\sigma_4}(0)} \bigg\{ ||z_1||_2 + |z_2| \bigg\}\\
&= -(L_V L_f + 2L_c)\sigma_4 \geq -\varepsilon.
\end{aligned}$$

The first equality of Eq. (5.25) follows since $\nabla_t [V]_{\sigma_4}(x,t) = [\nabla_t V]_{\sigma_4}(x,t)$ and $\nabla_x [V]_{\sigma_4}(x,t) = [\nabla_x V]_{\sigma_4}(x,t)$ for all $(x,t) \in K \subset < E >_{\sigma_4}$ by Prop. B.1 (Statement 2). The first inequality follows by the monotonicity property of integration and the Cauchy Swartz inequality. Since V is Lipschitz $\operatorname{ess\,sup}_{(x,t)\in E} ||\nabla_x V(x,t)||_2 < L_V$ by Rademacher's Theorem (Theorem C.4). The second inequality follows by using

(5.21) together with $\operatorname{ess\,sup}_{(x,t)\in E} ||\nabla_x V(x,t)||_2 < L_V$. The third inequality follows by the Lipschitz continuity of f and c. Finally the fourth inequality follows by the fact $\sigma_4 < \sigma_3 < \frac{\varepsilon}{L_V L_f + L_c}$.

Now define $J_\varepsilon(x,t) := [V]_\sigma(x.t)$ where $0 < \sigma < \min\{\sigma_1, \sigma_4\}$. It follows that $J_\varepsilon \in C^\infty(K,\mathbb{R})$ by Prop. B.1 (Statement 1). Moreover J_ε satisfies Eqs. (5.22) and (5.23) by Eqs. (5.24) and (5.25). □

In Lemma 5.2 we showed that for any given function, $V \in Lip(E,\mathbb{R})$, any compact subsets $K \subset E$, any $\varepsilon > 0$, and any $1 \le p < \infty$, there exists a smooth function, J_ε, satisfying Eq. (5.23), such that $||V - J_\varepsilon||_{W^{1,p}(K,\mathbb{R})} < \varepsilon$. We next show this "local" result over compact subsets, K, can be extended to a "global" results over the entire domain, E. To do this we use Theorem C.6, stated in Chapter C. Given an open cover of E, Theorem C.6 states that there exists a family of functions, called a partition of unity. In the next proposition we use partitions of unity together with the "local" approximates of the Lipschitz function, V, to construct a smooth "global" approximation of V over the entire domain E.

Proposition 5.2. *Let $E \subset \mathbb{R}^{n+1}$ be an open bounded set, $\Omega \subset \mathbb{R}^n$ be such that $\Omega \times (0,T) \subseteq E$, where $T > 0$, $U \subset \mathbb{R}^m$ be a compact set, $f \in Lip(\Omega \times U, \mathbb{R}^n)$, $c \in Lip(\Omega \times U \times [0,T], \mathbb{R})$, and $V \in Lip(E,\mathbb{R})$ satisfies Eq. (5.21). Then for all $1 \le p < \infty$ and $\varepsilon > 0$ there exits $J \in C^\infty(E,\mathbb{R})$ such that*

$$||V - J||_{W^{1,p}(E,\mathbb{R})} < \varepsilon \text{ and } \sup_{(x,t)\in E} |V(x,t) - J(x,t)| < \varepsilon, \tag{5.26}$$

and for all $(x,u,t) \in \Omega \times U \times (0,T)$

$$\nabla_t J(x,t) + \nabla_x J(x,t)^T f(x,u) + c(x,u,t) \ge -\varepsilon. \tag{5.27}$$

Proof. Let us consider the family of sets $E_i = \{x \in E : \sup_{y\in\partial E} ||x-y||_2 < \frac{1}{i}\}$ for $i \in \mathbb{N}$. It follows $\{E_i\}_{i=1}^\infty$ is an open cover (Defn. C.1) for E and thus by Theorem C.6 there exists a smooth partition of unity, $\{\psi_i\}_{i=1}^{\infty} \subset C^{\infty}(E, \mathbb{R})$, that satisfies Statements 1 to 4 of Theorem C.6.

For $\varepsilon > 0$ Lemma 5.2 shows that for each $i \in \mathbb{N}$ there exists a function $J_i \in C^{\infty}(\overline{E_i}, \mathbb{R})$ such that

$$\sup_{(x,t)\in E_i} |V(x,t) - J_i(x,t)| < \frac{\varepsilon}{2^{i+1}(1+\tau_i+\theta_i)}, \tag{5.28}$$

$$||V - J_i||_{W^{1,p}(\overline{E_i},\mathbb{R})} < \frac{\varepsilon}{2^{i+1}(1+\tau_i+\theta_i)}, \tag{5.29}$$

$$\begin{aligned}\nabla_t J_i(x,t) + \nabla_x J_i(x,t)^T f(x,u) + c(x,u,t) \geq -\frac{\varepsilon}{2^{i+1}(1+\tau_i+\theta_i)} \\ \text{for all } (x,t) \in \overline{E_i} \cap (\Omega \times (0,T)), u \in U,\end{aligned} \tag{5.30}$$

where we denote $\tau_i := \sup_{(x,u,t)\in\Omega\times U\times(0,T)}\{|\nabla_t \psi_i(x,t) + \nabla_x \psi_i(x,t)^T f(x,u)|\} \geq 0$ and $\theta_i := \left(\max_{|\alpha|\leq 1} \sup_{(x,t)\in E} |D^{\alpha}\psi_i(x,t)|^p\right)^p \geq 0$; which is well defined and finite as $\Omega \times U \times (0,T)$ is bounded and ψ_i is smooth.

Now, let us define $J(x,t) := \sum_{i=1}^{\infty} \psi_i(x,t) J_i(x,t)$, we will show $J \in C^{\infty}(E, \mathbb{R})$ and that J satisfies Eqs. (5.26) and (5.27).

It follows $J \in C^{\infty}(E, \mathbb{R})$ by Theorem C.6. To see this we note for each $i \in \mathbb{N}$ we have $\psi_i \in C^{\infty}(E, \mathbb{R})$ and $\psi_i(x,t) = 0$ outside E_i implying $\psi_i J_i \in C^{\infty}(E, \mathbb{R})$. Moreover, for each $(x,t) \in E$ there exists an open set, $S \subseteq E$, where only a finite number of ψ_i are nonzero. Therefore it follows that the function J is a finite sum of infinitely differentiable functions and thus J is also infinitely differentiable.

We now show J satisfies Eq. (5.26). We first show $||V - J||_{W^{1,p}(E,\mathbb{R})} < \varepsilon$:

$$\begin{aligned}||V - J||_{W^{1,p}(E,\mathbb{R})} &= ||V - \sum_{i=1}^{\infty} \psi_i J_i||_{W^{1,p}(E,\mathbb{R})} \qquad (5.31)\\ &= ||\sum_{i=1}^{\infty} \psi_i (V - J_i)||_{W^{1,p}(E,\mathbb{R})} \le \sum_{i=1}^{\infty} ||\psi_i (V - J_i)||_{W^{1,p}(E,\mathbb{R})}\\ &= \sum_{i=1}^{\infty} ||\psi_i (V - J_i)||_{W^{1,p}(\bar{E}_i,\mathbb{R})} \le \sum_{i=1}^{\infty} \theta_i ||V - J_i||_{W^{1,p}(\bar{E}_i,\mathbb{R})}\\ &< \sum_{i=1}^{\infty} \left(\frac{\varepsilon + \theta_i}{2^{i+1}(1 + \tau_i + \theta_i)} \right) < \varepsilon.\end{aligned}$$

The second equality of Eq. (5.31) follows since partitions of unity satisfy $\sum_{i=1}^{\infty} \psi_i(x,t) \equiv 1$ by Theorem C.6. The first inequality follows by the triangle inequality. The third equality follows since partitions of unity satisfy $\psi_i(x,t) = 0$ outside of E_i for all $i \in \mathbb{N}$ by Theorem C.6. The third inequality follows by Eq. (5.29). The fourth inequality follows as $\sum_{i=1}^{\infty} \frac{1}{2^i} = 1$. Now, by a similar augment to Eq. (5.31), using Eq. (5.28) rather than Eq. (5.29), it also follows $\sup_{(x,t)\in E} |V(x,t) - J(x,t)| < \varepsilon$ and thus J satisfies Eq. (5.26).

Next we will show J satisfies Eq. (5.27). Before doing this we first prove a preliminary identity. Specifically,

$$\sum_{i=1}^{\infty} \bigg(\nabla_t \psi_i(x,t) + \nabla_x \psi_i(x,t)^T f(x,u) \bigg) = 0, \qquad (5.32)$$

for all $(x,t) \in \Omega \times (0,T) \subseteq E$ and $u \in U$. This follows because only finitely many ψ_i's are non-zero for each $(x,t) \in E$ and thus it follows $\sum_{i=1}^{\infty} \psi_i(x,t)$ is a finite sum of infitely differentiable functions. Therefore, we can interchange derivatives and summations, thus since $\sum_{i=1}^{\infty} \psi_i(x,t) \equiv 1$ it follows that $\nabla_t \bigg(\sum_{i=1}^{\infty} \psi_i(x,t) \bigg) = \sum_{i=1}^{\infty} \nabla_t \psi_i(x,t) = 0$. Similarly for each $j \in \{1,...,n\}$ we have $\sum_{i=1}^{\infty} \frac{\partial \psi_i(x,t)}{\partial x_j} = 0$ which implies $\sum_{i=1}^{\infty} \nabla_x \psi_i(x,t) = 0 \in \mathbb{R}^n$.

Now, it follows J satisfies Eq. (5.27) since

$$\begin{aligned}
&\nabla_t J(x,t) + \nabla_x J(x,t)^T f(x,u) + c(x,u,t) \qquad (5.33)\\
&= \sum_{i=1}^{\infty} \bigg(\psi_i(x,t)(\nabla_t J_i(x,t) + \nabla_x J_i(x,t)^T f(x,u) + c(x,u,t)) \bigg)\\
&\quad + \sum_{i=1}^{\infty} \bigg(J_i(x,t)(\nabla_t \psi_i(x,t) + \nabla_x \psi_i(x,t)^T f(x,u)) \bigg)\\
&\geq \frac{-\varepsilon}{2} + \sum_{i=1}^{\infty} (J_i(x,t) - V(x,t))(\nabla_t \psi_i(x,t) + \nabla_x \psi_i(x,t)^T f(x,u)) \geq -\varepsilon,
\end{aligned}$$

for all $(x,t) \in \Omega \times (0,T) \subseteq E$ and $u \in U$. The first equality of Eq. (5.33) follows by the chain rule and the fact $\sum_{i=1}^{\infty} \psi_i(x,t) {\equiv} 1$. The first inequality follows by Eqs. (5.30) and (5.32). The second inequality follows by Eq. (5.28) and $\sum_{i=1}^{\infty} \frac{1}{2^i} {=} 1$. □

5.5.2 Existence of Dissipative Polynomials that can Approximate the VF Arbitrarily well under the $W^{1,p}$ Norm

Previously, in Prop. 5.2, we showed for any $V \in Lip(\Omega \times [0,T], \mathbb{R})$ satisfying Eq. (5.21) there exists a smooth function J that also satisfies Eq. (5.21) and approximates V with arbitrary accuracy under the Sobolev norm. We now use this result to show for any VF, associated with some family OCPs $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$, there exists a dissipative polynomial, V_l, that approximates the VF arbitrarily well with respect to the Sobolev norm. Our proof uses Theorem C.2, found in Chapter C, that shows differentiable functions, such as J, can be approximated up to their first order derivatives over compact sets arbitrarily well by polynomials. Prop. 5.2 only gives the existence of a smooth approximation, J, when the VF is Lipschitz continuous. Lemma 5.1 shows the VF, associated with a family of OCPs, is locally Lipschitz when $\Omega = \mathbb{R}^n$ (which is not a compact set). Unfortunately, Theorem C.2 can only be used for polynomial approximation over compact sets. Thus, before proceeding we first give a sufficient condition for a VF, associated with a family of OCPs with compact state constraints, to be Lipschitz continuous over some set $\Lambda \subset \Omega$.

Lipschitz continuity of VFs associated with a family of state constrained OCPs Consider the family of OCPs $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$. If the state is constrained ($\Omega \neq \mathbb{R}^n$), the associated VF can be discontinuous and is no longer uniquely defined as the viscosity solution of the HJB PDE. Next, in Lemma 5.3, we give a sufficient condition that when satisfied implies VFs, associated with a family of state constrained OCPs, are equal to the unique locally Lipschitz continuous VF of the state unconstrained OCP over some subset $\Lambda \subseteq \Omega$, and hence are Lipschitz continuous over Λ. To state Lemma 5.3 we first define the forward reachable set.

Definition 5.9. *For $X_0 \subset \mathbb{R}^n$, $\Omega \subseteq \mathbb{R}^n$, $U \subset \mathbb{R}^m$, $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and $S \subset \mathbb{R}^+$, define*

$$FR_f(X_0, \Omega, U, S) := \Big\{ y \in \mathbb{R}^n \ : \textit{there exists}\, x \in X_0, T \in S, \textit{and} \\ \mathbf{u} \in \mathcal{U}_{\Omega,U,f,T}(x, 0) \ \textit{such that} \ \phi_f(x, T, \mathbf{u}) = y \Big\}.$$

Lemma 5.3. *Consider $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$ and any function $V_1 : \Omega \times [0, T] \to \mathbb{R}$ that satisfies Eq. (5.10). Let $V_2 : \mathbb{R}^n \times [0, T] \to \mathbb{R}$ be the VF for the unconstrained problem $\{c, g, f, \mathbb{R}^n, U, T\}$. If $\Lambda \subseteq \Omega$ is such that*

$$FR_f(\Lambda, \mathbb{R}^n, U, [0, T]) \subseteq \Omega, \tag{5.34}$$

then $V_1(x, t) = V_2(x, t)$ for all $(x, t) \in \Lambda \times [0, T]$.

Proof. To show $V_1(x, t) = V_2(x, t)$ for all $(x, t) \in \Lambda \times [0, T]$ we must prove $\mathcal{U}_{\Omega,U,f,T}(x, t) = \mathcal{U}_{\mathbb{R}^n,U,f,T}(x, t)$ for all $(x, t) \in \Lambda \times [0, T]$.

For any $(x, t) \in \Lambda \times [0, T]$ if $\mathbf{u} \in \mathcal{U}_{\Omega,U,f,T}(x, t)$ then clearly $\mathbf{u} \in \mathcal{U}_{\mathbb{R}^n,U,f,T}(x, t)$, thus $\mathcal{U}_{\Omega,U,f,T}(x, t) \subseteq \mathcal{U}_{\mathbb{R}^n,U,f,T}(x, t)$. On the other hand if $\mathbf{u} \in \mathcal{U}_{\mathbb{R}^n,U,f,T}(x, t)$ then by

Defn. 5.4 it follows $\mathbf{u}(s) \in U$ for all $s \in [t,T]$ and that there exists a unique map, denoted by $\phi_f(x,s,\mathbf{u})$, that satisfies the following for all $s \in [t,T]$

$$\frac{\partial \phi_f(x,s-t,\tau_t \mathbf{u})}{\partial s} = f(\phi_f(x,s-t,\tau_t \mathbf{u}), \mathbf{u}(s)), \;\; \phi_f(x,0,\tau_t \mathbf{u}) = x.$$

To show $\mathbf{u} \in \mathcal{U}_{\Omega,U,f,T}(x,t)$ we need $\phi_f(x,s-t,\tau_t \mathbf{u}) \in \Omega$ for all $s \in [t,T]$, which is equivalent to

$$\phi_f(x,s,\tilde{\mathbf{u}}) \in \Omega \text{ for all } s \in [0,T-t], \tag{5.35}$$

where $\tilde{\mathbf{u}} = \tau_t \mathbf{u} \in \mathcal{U}_{\Omega,U,f,T-t}(x,0)$. Eq. (5.35) then follows trivially by Eq. (5.34). $\square$

Alternative sufficient conditions that imply a VF, associated with some family of state constrained OCPs, is Lipschitz continuous and the unique viscosity solution of the HJB PDE include: the Inward Pointing Constraint Qualification (IPCQ) found in Soner (1986) and Frankowska and Mazzola (2013), the Outward Pointing Constraint Qualification (OPCQ) found in Frankowska and Vinter (2000), and epigraph characterization of VFs found in Altarovici *et al.* (2013).

Approximation of VFs by dissipative polynomials Considering a family of OCPs $\{c,g,f,\Omega,U,T\} \in \mathcal{M}_{Lip}^{Continuous}$, and assuming there exists a set $\Lambda \subseteq \Omega$ that satisfies Eq. (5.34), we now prove the existence of dissipative polynomial functions that can approximate the any VF of $\{c,g,f,\Omega,U,T\} \in \mathcal{M}_{Lip}^{Continuous}$ arbitrarily well under the Sobolev norm.

Theorem 5.3. *For given $\{c,g,f,\Omega,U,T\} \in \mathcal{M}_{Lip}^{Continuous}$ suppose $\Lambda \subseteq \Omega$ is a bounded set that satisfies Eq.* (5.34)*, then for any function V satisfying Eq.* (5.10)*, $1 \le p < \infty$,* *and* $\varepsilon > 0$ *there exists* $V_l \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ *such that*

$$\|V - V_l\|_{W^{1,p}(\Lambda \times [0,T], \mathbb{R})} < \varepsilon, \tag{5.36}$$

$$\sup_{(x,t) \in \Lambda \times [0,T]} |V(x,t) - V_l(x,t)| < \varepsilon, \tag{5.37}$$

$$V_l(x,t) \leq V(x,t) \textit{ for all } t \in [0,T] \textit{ and } x \in \Omega, \tag{5.38}$$

$$\nabla_t V_l(x,t) + c(x,u,t) + \nabla_x V_l(x,t)^T f(x,u) > 0 \tag{5.39}$$

$$\textit{for all } x \in \Omega, t \in (0,T), u \in U,$$

$$V_l(x,T) < g(x) \textit{ for all } x \in \Omega. \tag{5.40}$$

Proof. Let $\varepsilon > 0$. Suppose V satisfies Eq. (5.10). Rather than approximating V, defined for a family of OCPs on the compact set Ω, we instead approximate the unique VF, denoted by V^*, associated with the family of OCPs where $\Omega = \mathbb{R}^n$. It is easier to approximate V^* compared to V as V^* has the following useful properties: By Lemma 5.1, V^* is locally Lipschitz continuous; and by Theorem 5.1, V^* is the unique viscosity solution of the HJB PDE (5.12). Furthermore, as Λ satisfies Eq. (5.34), Lemma 5.3 implies

$$V^*(x,t) = V(x,t) \text{ for all } (x,t) \in \Lambda \times [0,T]. \tag{5.41}$$

This proof is structured as follows. We first use Prop. 5.2 to approximate V^* by an infinitely differentiable function denoted as J_δ. Then using Theorem C.2, found in Chapter C, we approximate J_δ by a polynomial P_δ. Finally, to ensure Inequalities (5.39) and (5.40) are satisfied, a correction term ρ is subtracted from P_δ, creating the function $V_l(x,t) := P_\delta(x,t) - \rho(t)$ that we show satisfies Eqs. (5.36) to (5.40).

Since Ω is compact, there exists some open bounded set $E \subset \mathbb{R}^{n+1}$ of finite measure which contains $\overline{\Omega \times (0,T)}$. Since $V^* \in LocLip(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ (by Lemma 5.1) and $E \subset \mathbb{R}^n$ is bounded it follows $V^* \in Lip(E \times [0,T], \mathbb{R})$. Then by Rademacher's theorem (See Theorem C.4 in Chapter C), V^* is differentiable almost everywhere in E. Moreover, as V^* is the unique viscosity solution to the HJB PDE, the following holds for all $u \in U$ and almost everywhere in $(x,t) \in \Omega \times (0,T) \subset E$.

$$\begin{aligned}&\nabla_t V^*(x,t) + c(x,u,t) + \nabla_x V^*(x,t)^T f(x,u)\\ &\geq \nabla_t V^*(x,t) + \inf_{u \in U}\{c(x,u,t) + \nabla_x V^*(x,t)^T f(x,u)\} = 0\end{aligned}$$

This implies that the following holds for all $u \in U$

$$\operatorname*{ess\,inf}_{(x,t)\in\Omega\times(0,T)} \left\{\nabla_t V^*(x,t) + \nabla_x V^*(x,t)^T f(x,u) + c(x,u,t)\right\} \geq 0.$$

Therefore, we conclude that V^* satisfies Eq. (5.21). Thus, by Prop. 5.2, for any $\delta > 0$ there exists $J_\delta \in C^\infty(E,\mathbb{R})$ such that

$$\|V^* - J_\delta\|_{W^{1,p}(E,\mathbb{R})} < \delta, \tag{5.42}$$

$$\nabla_t J_\delta(x,t) + \nabla_x J_\delta(x,t)^T f(x,u) + c(x,u,t) \geq -\delta \text{ for all } (x,t) \in \Omega \times (0,T). \tag{5.43}$$

In particular, let us choose $\delta > 0$ such that

$$\delta < \frac{\varepsilon}{2 + (2 + 4T + 2MT)(T\mu(\Lambda))^{\frac{1}{p}}}, \tag{5.44}$$

where $M := \sup_{(x,u)\in\Omega\times U} ||f(x,u)||_2 < \infty$ and $\mu(\Lambda) < \infty$ is the Lebesgue measure of Λ.

We now approximate $J_\delta \in C^\infty(E,\mathbb{R})$ by a polynomial function. Theorem C.2, found in Chapter C, shows there exists $P_\delta \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ such that for all $(x,t) \in E$

$$|J_\delta(x,t) - P_\delta(x,t)| < \delta. \tag{5.45}$$

$$|\nabla_t J_\delta(x,t) - \nabla_t P_\delta(x,t)| < \delta. \tag{5.46}$$

$$||\nabla_x J_\delta(x,t) - \nabla_x P_\delta(x,t)||_2 < \delta. \tag{5.47}$$

$$\|J_\delta - P_\delta\|_{W^{1,p}(E,\mathbb{R})} < \delta. \tag{5.48}$$

Now,

$$\begin{aligned}\|V^* - P_\delta\|_{W^{1,p}(E,\mathbb{R})} &= \|V^* - J_\delta + J_\delta - P_\delta\|_{W^{1,p}(E,\mathbb{R})}\\ &\leq \|V^* - J_\delta\|_{W^{1,p}(E,\mathbb{R})} + \|J_\delta - P_\delta\|_{W^{1,p}(E,\mathbb{R})} < 2\delta, \end{aligned} \tag{5.49}$$

where the first inequality follows by the triangle inequality, and the second inequality follows from Eq. (5.42) and Eq. (5.48).

By a similar argument to Inequality (5.49) we deduce,

$$\sup_{(x,t)\in E} |V^*(x,t) - P_\delta(x,t)| < 2\delta. \tag{5.50}$$

Furthermore,

$$\begin{aligned}&\nabla_t P_\delta(x,t) + \nabla_x P_\delta(x,t)^T f(x,u) + c(x,u,t)\\ &\geq \left(\nabla_t P_\delta(x,t) + \nabla_x P_\delta(x,t)^T f(x,u) + c(x,u,t)\right)\\ &\qquad - \delta - \left(\nabla_t J_\delta(x,t) + \nabla_x J_\delta(x,t)^T f(x,u) + c(x,u,t)\right)\\ &= -\delta + \left(\nabla_t P_\delta(x,t) - \nabla_t J_\delta(x,t)\right) - \left(\nabla_x J_\delta(x,t) - \nabla_x P_\delta(x,t)\right)^T f(x,u)\\ &> -\delta - \delta - ||\nabla_x J_\delta(x,t) - \nabla_x P_\delta(x,t)||_2 ||f(x,u)||_2\\ &> -(2+M)\delta \text{ for all } (x,t) \in \Omega \times (0,T), \end{aligned} \tag{5.51}$$

The first inequality of Eq. (5.51) follows by Inequality Eq. (5.43). The second inequality follows by Eq. (5.46) and the Cauchy Schwarz inequality. The third inequality follows by Eq. (5.47).

Moreover, we have that

$$\begin{aligned}P_\delta(x,T) &= P_\delta(x,T) - V^*(x,T) + V^*(x,T)\\ &< g(x) + 2\delta \text{ for all } x \in \Omega. \end{aligned} \tag{5.52}$$

This inequality follows from the fact that $V^*(x,T) = g(x)$ since V^* satisfies the boundary condition in the HJB PDE (5.12), and Eq. (5.50).

We now construct V_l from P_δ. Let us denote the correction function $\rho(t) := (2+M)(T-t)\delta + 2\delta$, where $M = \sup_{(x,u)\in\Omega\times U} ||f(x,u)||_2$. We define V_l as

$$V_l(x,t) := P_\delta(x,t) - \rho(t). \tag{5.53}$$

We now find that V_l satisfies Inequality (5.39) since we have

$$\begin{aligned}&\nabla_t V_l(x,t) + c(x,u,t) + \nabla_x V_l(x,t)^T f(x,u)\\ &= \bigg(\nabla_t P_\delta(x,t) + \nabla_x P_\delta(x,t)^T f(x,u) + c(x,u,t)\bigg) + (2+M)\delta\\ &> 0, \qquad \text{for all } (x,t) \in \Omega \times (0,T),\end{aligned}$$

where the above inequality follows from Eq. (5.51).

We next show V_l satisfies Inequality (5.40):

$$V_l(x,T) = P_\delta(x,T) - 2\delta < g(x) \text{ for all } x \in \Omega,$$

where the above inequality follows by Eq. (5.52).

Now, since V_l satisfies Eqs. (5.39) and (5.40) it follows V_l satisfies Eq. (5.38) by Prop. 5.1.

To show that V_l satisfies Inequality (5.36), we first we derive a bound on the norm of the correction function ρ.

$$\begin{aligned}||\rho||_{W^{1,p}(\Lambda\times[0,T],\mathbb{R})} &= \bigg(\int_{\Lambda\times[0,T]} |(2+M)(T-t)\delta + 2\delta|^p dxdt\bigg)^{\frac{1}{p}}\\ &\quad + \bigg(\int_{\Lambda\times[0,T]} |(2+M)\delta|^p dxdt\bigg)^{\frac{1}{p}} \le (2+4T+2MT)(T\mu(\Lambda))^{\frac{1}{p}}\delta.\end{aligned}$$

Now, by Eqs. (5.41), (5.44) and (5.49),

$$\begin{aligned}\|V - V_l\|_{W^{1,p}(\Lambda\times[0,T],\mathbb{R})} &= \|V^* - V_l\|_{W^{1,p}(\Lambda\times[0,T],\mathbb{R})} \qquad (5.54)\\ &= \|V^* - P_\delta - \eta\|_{W^{1,p}(\Lambda\times[0,T],\mathbb{R})}\\ &\leq \|V^* - P_\delta\|_{W^{1,p}(E,\mathbb{R})} + \|\eta\|_{W^{1,p}(\Lambda\times[0,T],\mathbb{R})}\\ &\leq 2\delta + (2+4T+2MT)(T\mu(\Lambda))^{\frac{1}{p}}\delta < \varepsilon.\end{aligned}$$

By a similar argument to Eq. (5.54) we deduce V_l satisfies Eq. (5.37)

We conclude that V_l, defined in Eq. (5.53), satisfies Eqs. (5.37), (5.38), (5.39), and (5.40) thus completing the proof. □

5.5.3 Our Family Of Optimization Problems Yield A Sequence Of Polynomials That Converge To A VF Under The L^1 Norm

Consider some $\{c,g,f,\Omega,U,T\} \in \mathcal{M}_{Lip}^{Continuous}$ and suppose the sequence $\{J_d\}_{d\in\mathbb{N}}$ solves each instance of the optimization problem given in Eq. (5.20) for $d \in \mathbb{N}$. We next use Theorem 5.3 to show that the sequence, $\{J_d\}_{d\in\mathbb{N}}$ converges to any VF associated with the family of OCP's $\{c,g,f,\Omega,U,T\} \in \mathcal{M}_{Lip}^{Continuous}$ with respect to the weighted L^1 norm as $d \to \infty$.

Proposition 5.3. *For given $\{c,g,f,\Omega,U,T\} \in \mathcal{M}_{Lip}^{Continuous}$ and positive integrable function $w \in L^1(\Omega\times[0,T],\mathbb{R}^+)$ suppose $\Lambda \subseteq \Omega$ satisfies Eq.* (5.34) *then*

$$\lim_{d\to\infty}\int_{\Lambda\times[0,T]} w(x,t)|V(x,t) - J_d(x,t)|dxdt = 0, \qquad (5.55)$$

where V is any function satisfying Eq. (5.10)*, and $J_d \in \mathcal{P}_d(\mathbb{R}^n\times\mathbb{R},\mathbb{R})$ is any solution to Optimization Problem* (5.20) *for $d \in \mathbb{N}$.*

Proof. Suppose V satisfies the theorem statement. To show Eq. (5.55) we must show that for any $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$\int_{\Lambda\times[0,T]} w(x,t)|V(x,t) - J_d(x,t)|dxdt < \varepsilon \text{ for all } d \geq N.$$

Since by assumption Λ satisfies Eq. (5.34), we can use Theorem 5.3 (from Section 5.5.2) to show that for any $\delta > 0$ there exists dissipative $V_l \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ feasible to Optimization Problem (5.20) and is such that

$$\operatorname*{ess\,sup}_{(x,t)\in\Lambda\times[0,T]} |V(x,t) - V_l(x,t)| < \delta.$$

For our given $\epsilon > 0$, by selecting $\delta < \varepsilon / \int_{\Lambda\times[0,T]} w(x,t)dxdt$ (Note if $\int_{\Lambda\times[0,T]} w(x,t)dxdt = 0$, Eq. (5.55) already holds and the proof is complete) we have a V_l such that

$$\begin{aligned} &\int_{\Lambda\times[0,T]} w(x,t)|V(x,t) - V_l(x,t)|dxdt \qquad (5.56)\\ &\leq \int_{\Lambda\times[0,T]} w(x,t)dxdt \operatorname*{ess\,sup}_{(x,t)\in\Lambda\times[0,T]} |V(x,t) - V_l(x,t)| \\ &< \delta \int_{\Lambda\times[0,T]} w(x,t)dxdt < \varepsilon. \end{aligned}$$

Now define $N := \deg(V_l)$ and denote the solution to Problem (5.20) for $d \geq N$ as $J_d \in \mathcal{P}_N(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$. As V_l is feasible to Problem (5.20) for all $d \geq N$, it follows the objective function evaluated at J_d is greater than or equal to the objective function evaluated at V_l; that is

$$\int_{\Lambda\times[0,T]} w(x,t)J_d(x,t)dxdt \geq \int_{\Lambda\times[0,T]} w(x,t)V_l(x,t)dxdt \text{ for } d \geq N. \qquad (5.57)$$

$$\begin{aligned} \text{Now, } &\int_{\Lambda\times[0,T]} w(x,t)|V(x,t) - J_d(x,t)|dxdt \qquad (5.58)\\ &= \int_{\Lambda\times[0,T]} w(x,t)V(x,t) - w(x,t)J_d(x,t)dxdt \\ &\leq \int_{\Lambda\times[0,T]} w(x,t)|V(x,t) - V_l(x,t)|dxdt < \varepsilon \text{ for all } d \geq N. \end{aligned}$$

The equality in Eq. (5.58) follows since $J_d(x,t) \leq V(x,t)$ for all $(x,t) \in \Omega \times [0,T]$ (Prop. 5.1). The first inequality follows by a combination of Eq. (5.57) and the inequality $V_l(x,t) \leq V(x,t)$ for all $(x,t) \in \Omega \times [0,T]$. Finally, the second inequality follows by Eq. (5.56). □

5.6 A Family of SOS Problems that Yield Polynomials that Converge to the VF

Consider some $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$ and denote $\{J_d\}_{d\in\mathbb{N}}$ as the sequence of solutions to the optimization problem found in Eq. (5.20). We have shown in Prop. 5.3that the sequence of functions, $\{J_d\}_{d\in\mathbb{N}}$, converge to any VF associated with the family of OCPs $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$ with respect to the L^1 norm. The indexed polynomial optimization problems in Eq. (5.20) may now be readily tightened to more tractable SOS optimization problems.

Specifically, for each $d \in \mathbb{N}$, we tighten the polynomial optimization problem in Eq. (5.20) to the SOS optimization problem given in Eq. (5.59). We later show in Prop. 5.4 that the sequence of solutions to the SOS problem given in Eq. (5.59) yield polynomials, $\{P_d\}_{d\in\mathbb{N}}$, indexed by degree $d \in \mathbb{N}$, that converge to the VF (with respect to the L^1 norm) as $d \to \infty$.

For our SOS implementation we consider a special class of OCPs, given next in Defn. 5.10. This class has the property that functions c, g, f are polynomial, and sets Ω and U are semi-algebraic.

Definition 5.10. *We say the six tuple $\{c, g, f, \Omega, U, T\}$ is a polynomial optimal control problem or $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Poly}^{Continuous}$ if the following holds*

1. $c \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}, \mathbb{R})$ *and* $g \in \mathcal{P}(\mathbb{R}^n, \mathbb{R})$.

2. $f \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^m, \mathbb{R}^n)$.

3. *There exists* $h_\Omega \in \mathcal{P}(\mathbb{R}^n, \mathbb{R})$ *such that* $\Omega = \{x \in \mathbb{R}^n : h_\Omega(x) \geq 0\}$.

4. *There exists* $h_U \in \mathcal{P}(\mathbb{R}^m, \mathbb{R})$ *such that* $U = \{u \in \mathbb{R}^m : h_U(u) \geq 0\}$.

Note polynomials are locally Lipschitz continuous, that is $\mathcal{P}(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}) \subset LocLip(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$. Therefore $\mathcal{M}_{Poly}^{Continuous} \subset \mathcal{M}_{Lip}^{Continuous}$.

For given $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Poly}^{Continuous}$, $d \in \mathbb{N}$, $\Lambda \subset \mathbb{R}^n$ and $w \in L^1(\Lambda \times [0,T], \mathbb{R}^+)$, we thus propose an SOS tightening of Optimization Problem (5.20) as follows:

$$
\begin{aligned}
& P_d \in \arg \max_{P \in \mathcal{P}_d(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})} c^T \alpha && (5.59)\\
& \text{subject to: } k_0, k_1 \in \sum_{SOS}^{d}, \quad s_i \in \sum_{SOS}^{d} \text{ for } i = 0,1,2,3,\\
& P(x,t) = c^T Z_d(x,t),\\
& k_0(x) = g(x) - P(x,T) - s_0(x) h_\Omega(x),\\
& k_1(x,u,t) = \nabla_t P(x,t) + c(x,u,t) + \nabla_x P(x,t)^T f(x,u)\\
& \qquad\qquad - s_1(x,u,t) h_\Omega(x) - s_2(x,u,t) h_U(u) - s_3(x,u,t) \cdot (Tt - t^2),
\end{aligned}
$$

where $\alpha_i = \int_{\Lambda \times [0,T]} w(x,t) Z_{d,i}(x,t) dxdt$, and recalling $Z_d : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^{\mathcal{N}_d}$ is the vector of monomials of degree $d \in \mathbb{N}$. Note that solutions to Opt. (5.59) may not be feasible to Opt. (5.20) due to the strict inequalities of the latter problem.

5.6.1 We Can Numerically Construct A Sequence Of Polynomials That Converge The VF

For a given family of OCPs, we now show that the sequence of solutions to the SOS Opts. (5.59) converges locally to the VF of the associated OCPs with respect to the L^1 norm.

Proposition 5.4. *For given $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Poly}^{Continuous}$ and positive integrable function $w \in L^1(\Omega \times [0,T], \mathbb{R}^+)$ suppose $\Lambda \subseteq \Omega$ satisfies Eq.* (5.34) *then*

$$
\lim_{d \to \infty} \int_{\Lambda \times [0,T]} w(x,t) |V(x,t) - P_d(x,t)| dxdt = 0, \qquad (5.60)
$$

where V is any function satisfying Eq. (5.10) *and $P_d \in \mathcal{P}_d(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ is any solution to Problem* (5.59) *for $d \in \mathbb{N}$.*

Proof. To show Eq. (5.60) we show that for any $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $d \geq N$

$$\int_{\Lambda \times [0,T]} w(x,t)|V(x,t) - P_d(x,t)|dxdt < \varepsilon.$$

As it is assumed Λ satisfies Eq. (5.34) we are able to use Prop. 5.3 that shows for any $\varepsilon > 0$ there exists $N_1 \in \mathbb{N}$ such that for all $d \geq N_1$

$$\int_{\Lambda \times [0,T]} w(x,t)|V(x,t) - J_d(x,t)|dxdt < \varepsilon, \tag{5.61}$$

where J_d is a solution to Optimization Problem (5.20) for $d \in \mathbb{N}$.

In particular let us fix some $d_1 \geq N_1$. Since J_{d_1} solves Problem (5.20) it must satisfy the constraints of Problem (5.20). Thus we have

$$\begin{aligned} &k_0(x) := g(x) - J_{d_1}(x,T) > 0 \text{ for all } x \in \Omega, \\ &k_1(x,u,t) := \nabla_t J_{d_1}(x,t) + c(x,u,t) + \nabla_x J_{d_1}(x,t)^T f(x,u) > 0 \\ &\qquad\qquad\qquad\qquad\qquad\qquad \text{for all } (x,u,t) \in \Omega \times U \times [0,T]. \end{aligned}$$

Since k_0 and k_1 are strictly positive functions over the compact semialgebriac set $\Omega \times U \times [0,T] = \{(x,u,t) \in \mathbb{R}^{n+m+1} : h_\Omega(x) \geq 0, h_U(u) \geq 0, t(T-t) \geq 0\}$, Putinar's Positivstellensatz (stated in Theorem C.5, Chapter C) shows that there exist $s_0, s_1, s_2, s_3, s_4, s_5 \in \sum_{SOS}$ such that

$$\begin{aligned} k_0 - h_\Omega s_0 &= s_1, \\ k_1 - h_\Omega s_2 - h_U s_3 - h_T s_4 &= s_5, \end{aligned} \tag{5.62}$$

where $h_T(t) := (t)(T-t)$.

Let us denote $N_2 := \max_{i \in \{0,1,2,3,4,5\}} deg(s_i)$. By Eq. (5.62) it follows that if J_{d_1} is feasible to Problem (5.59) for $d \geq \max\{d_1, N_2\}$. Therefore, for $d \geq \max\{d_1, N_2\}$, the objective function evaluated at the solution to Problem (5.59) must be greater than or equal to objective function evaluated at J_{d_1}. That is by writing the solution to Problem (5.59) as $P_d(x,t) = c_d^T Z_d(x,t)$ and writing J_{d_1} as $J_{d_1} = b_{d_1}^T Z_{d_1}(x,t)$ we get that for $d \geq \max\{d_1, N_2\}$

$$c_d^T \alpha \geq b_{d_1}^T \alpha. \tag{5.63}$$

Now using Eqs. (5.63) and (5.61) it follows for all $d \geq \max\{d_1, N_2\}$

$$\begin{aligned}
&\int_{\Lambda \times [0,T]} w(x,t)|V(x,t) - P_d(x,t)|dxdt \\
&\qquad = \int_{\Lambda \times [0,T]} w(x,t)V(x,t)dxdt - \int_{\Lambda \times [0,T]} w(x,t)P_d(x,t)dxdt \\
&\qquad = \int_{\Lambda \times [0,T]} w(x,t)V(x,t)dxdt - c_d^T \alpha \\
&\qquad \leq \int_{\Lambda \times [0,T]} w(x,t)V(x,t)dxdt - b_{d_1}^T \alpha \\
&\qquad = \int_{\Lambda \times [0,T]} w(x,t)|V(x,t) - J_{d_1}(x,t)|dxdt < \varepsilon,
\end{aligned}$$

where the above inequality follows using Prop. 5.1, which shows $P_d(x,t) \leq V(x,t)$ and $J_d(x,t) \leq V(x,t)$ for all $(x,t) \in \Omega \times [0,T]$, as P_d and J_d satisfy Inequalities (5.16) and (5.17) (since they both satisfy the constraints of Optimization Problem (5.20)). □

5.6.2 We can Numerically Construct a Sequence of Sublevel Sets that Converge to the VFs Sublevel Set

For a given family of OCPs, Prop. 5.4 shows the SOS optimization problem, given in Eq. (5.59), yields a sequence of polynomials, $\{P_d\}_{d \in \mathbb{N}}$, a sequence that converges to the VF (denoted by V), where convergence is with respect to the L^1 norm, and where the VF is associated with the given family of OCPs. We next extend this convergence result by showing that, for any $\gamma \in \mathbb{R}$, the sequence $\{P_d\}_{d \in \mathbb{N}}$ yields a sequence of γ-sublevel sets, where the sequence of γ-sublevel sets converges to the γ-sublevel set of the value function, V, where convergence is with respect to the volume metric.

For sets $A, B \subset \mathbb{R}^n$, let us recall the volume metric (defined in Appendix A) as $D_V(A,B)$, where

$$D_V(A,B) := \mu((A/B) \cup (B/A)),$$

where we recall $\mu(A) := \int_{\mathbb{R}^n} \mathbb{1}_A(x) dx$ is the Lebesgue measure. Note that Lemma A.1 shows that D_V is a metric.

Proposition 5.5. *Consider $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Poly}^{Continuous}$ and $w(x,t) = \delta(t-s)$ where $s \in [0,T]$ and δ is the Dirac delta function. Suppose $\Lambda \subseteq \Omega$ satisfies Eq.* (5.34)*. Then we have the following for all $\gamma \in \mathbb{R}$:*

$$\lim_{d \to \infty} D_V\bigg(\{x \in \Lambda : V(x,s) \le \gamma\}, \{x \in \Lambda : P_d(x,s) \le \gamma\} \bigg) = 0, \tag{5.64}$$

where V is any function satisfying Eq. (5.10)*, and P_d is any solution to Problem* (5.59) *for $d \in \mathbb{N}$.*

Proof. To show Eq. (5.64) we use Prop. A.1. Let us consider the family of functions, $\{P_d \in \mathcal{P}_d(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}) : d \in \mathbb{N}\}$, where P_d solves the optimization problem given in Eq. (5.59) for $d \in \mathbb{N}$ and $w(x,t) = \delta(t-s)$.

From the definition of Problem (5.59), we have that P_d satisfies the Inequalities in (5.16) and (5.17). Therefore, by Prop. 5.1, we have that $P_d(x,t) \le V(x,t)$ for all $(x,t) \in \Omega \times [0,T]$, where V is any function satisfying Eq. (5.10). Since $\Lambda \subseteq \Omega$ satisfies Eq. (5.34), and although the Dirac Delta function is not a member of $L^1(\Omega \times [0,T], \mathbb{R})$, a similar argument to Prop. 5.4 implies that

$$\lim_{d \to \infty} \int_\Lambda |V(x,s) - P_d(x,s)| dx dt = \lim_{d \to \infty} \int_{\Lambda \times [0,T]} \delta(t-s) |V(x,t) - P_d(x,t)| dx dt = 0.$$

We now apply Prop. A.1 to deduce Eq. (5.64). □

5.7 Numerical Examples: Using our SOS Algorithm to Approximate VFs

In this section we use the SOS programming problem as defined in Eq. (5.59) to numerically approximate the VFs associated with several different OCPs. We first approximate a known VF. Then, in Subsection 5.7.1, we approximate another unknown VF for reachable set estimation.

Example 5.1. *Let us consider the tuple* $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Poly}^{Continuous}$*, where* $c(x, u, t) \equiv 0$*,* $g(x) = x$*,* $f(x, u) = xu$*,* $\Omega = (-R, R) = \{x \in \mathbb{R} : x^2 < R^2\}$*,* $U = (-1, 1) = \{u \in \mathbb{R} : u^2 < 1\}$*, and* $T = 1$*. It was shown in Liberzon (2011) that the VF associated with* $\{c, g, f, \mathbb{R}^n, U, T\}$ *can be analytically found as*

$$V^*(x,t) = \begin{cases} \exp(t-1)x \textit{ if } x > 0, \\ \exp(1-t)x \textit{ if } x < 0, \\ 0 \textit{ if } x = 0. \end{cases} \tag{5.65}$$

We note that V^* *is not differentiable at* $x = 0$*. However,* V^* *satisfies the HJB PDE away from* $x = 0$*. This problem shows that the VF can be non-smooth even for simple OCPs with polynomial vector field and cost functions.*

In Fig. 5.1 we have plotted the point wise error, $e(x, t) := V^*(x, t) - P_d(x, t)$*, where* P_d *is the solution to the SOS Optimization Problem* (5.59) *for* $d = 16$*,* $T = 1$*,* $\Lambda = [-0.5, 0.5]$*,* $w(x, t) \equiv 1$*,* $h_\Omega(x) = 2.4^2 - x^2$ *and* $h_U(u) = 1 - u^2$*. The figure shows* $e(x, t) \geq 0$ *for all* $(x, t) \in [-0.5, 0.5] \times [0, 1]$ *verifying that, as expected by Prop. 5.1,* P_d *is a sub-VF. Moreover,* $0 < e(x, t) < 0.1125$ *for all* $(x, t) \in [-0.5, 0.5] \times [0, 1]$ *implying* $||V^* - P_d||_\infty < 0.1125$*, showing that we get a tight VF approximation in the* L^∞ *norm (even though we optimize for the* L^1 *norm).*

In Fig. 5.2 we have plotted the function $F(d) := ||V^* - V_d||_{L^1(\Lambda,\mathbb{R})}$ *where* V^* *is given in Eq.* (5.65) *and* V_d *is the solution to the SOS Optimization Problem* (5.59)

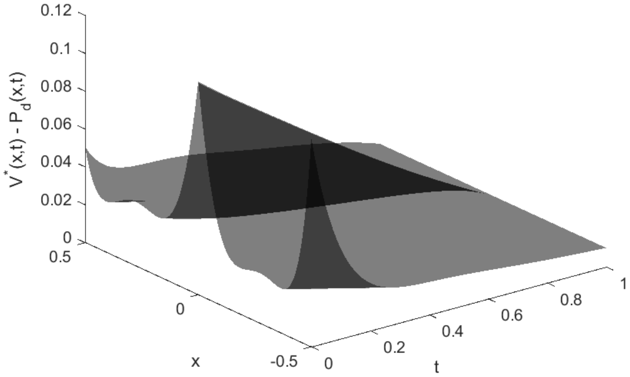

Figure 5.1: Plot associated with Example 5.1 showing point wise error, $e(x,t) := V^*(x,t) - P_d(x,t)$ where V^* is given in Eq. (5.65) and P_d solves the SOS Problem (5.59) for $d = 16$.

for $d = 4$ to 20, where $\Lambda = [-0.5, 0.5]$, $w(x,t) \equiv 1$, $h_\Omega(x) = 2.4^2 - x^2$ and $h_U(u) = 1 - u^2$. All solutions, V_d, of Problem (5.59) were sub-value functions as expected. Moreover, the figure shows by increasing the degree $d \in \mathbb{N}$ the resulting sub-VF, V_d, better approximates V^*, however convergence does slow after $d = 5$.

5.7.1 Application: Reachable Set Estimation

We next present several reachable set results required in our numerical approximation of the Lorenz attractor (Example 5.2). Similarly to forward reachable sets (Defn. 5.9) we now define backward reachable sets.

Definition 5.11. *For $X_0 \subset \mathbb{R}^n$, $\Omega \subseteq \mathbb{R}^n$, $U \subset \mathbb{R}^m$, $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and $S \subset \mathbb{R}^+$, let*

$$BR_f(X_0, \Omega, U, S) := \Big\{ y \in \mathbb{R}^n \ : \ \text{there exists } x \in X_0, T \in S, \\ \text{and } \mathbf{u} \in \mathcal{U}_{\Omega,U,f,T}(y,0) \text{ such that } \phi_f(y,T,\mathbf{u}) = x \Big\}.$$

Theorem 5.4 (VFs characterize backward reachable sets, see Jones and Peet (2019b)).

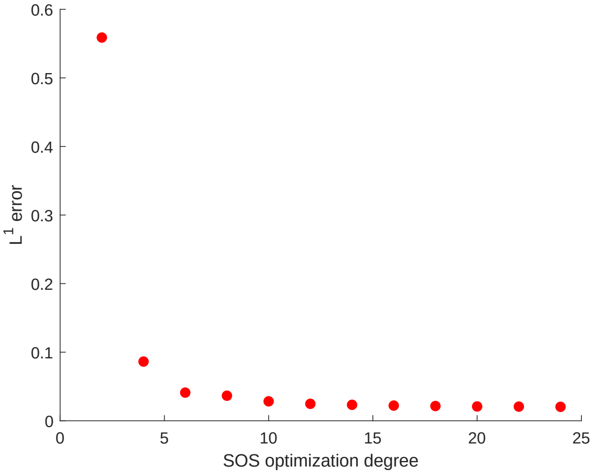

Figure 5.2: Scatter plot associated with Example 5.1 showing the L^1 norm error: $||V^* - P_d||_{L^1(\Lambda \times [0,T], \mathbb{R})}$, where V^* is given in Eq. (5.65) and P_d solves the SOS Problem (5.59) for $d = 4$ to 24. The smallest L^1 norm error occurred at $d = 24$ with a value of 0.020316.

Given $\{0, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}$ *define* $X_0 := \{x \in \mathbb{R}^n : g(x) < 0\}$*. Then*

$$BR_f(X_0, \Omega, U, \{T\}) = \{x \in \Omega : V^*(x,0) < 0\}, \tag{5.66}$$

where $V^* : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ *is any function that satisfies Eq.* (5.10)*.*

Corollary 5.1 (Sub-VFs contain reachable sets)**.** *Given* $\{0, g, f, \Omega, U, T\} \in \mathcal{M}_{Lip}$ *and suppose* $V_l : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ *is a sub-VF (Defn. 5.7), then*

$$BR_f(X_0, \Omega, U, \{T\}) \subseteq \{x \in \Omega : V_l(x,0) < 0\}, \tag{5.67}$$

where $X_0 := \{x \in \mathbb{R}^n : g(x) < 0\}$*.*

Lemma 5.4 (Equivalence of computation of backward and forward reachable sets, see Jones and Peet (2019b))**.** *Suppose* $X_0 \subset \mathbb{R}^n$*,* $\Omega \subset \mathbb{R}^n$*,* $U \subset \mathbb{R}^m$*,* $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$*, and* $T \in \mathbb{R}^+$*. Then*

$$FR_{-f}(X_0, \Omega, U, \{T\}) = BR_f(X_0, \Omega, U, \{T\}).$$

We now numerically solve the SOS programming problem in Eq. (5.59) obtaining an approximate VF that can be used to estimate the reachable set of the Lorenz system. The problem of estimating the Lorenz attractor has previously been studied in Jones and Peet (2019c); Li *et al.* (2005); Wang *et al.* (2012a); Goluskin (2020).

Example 5.2. *Let us consider the Lorenz system defined by the three dimensional second order nonlinear ODE:*

$$\begin{aligned}\dot{x}_1(t) &= \sigma(x_2(t) - x_1(t)), \\ \dot{x}_2(t) &= x_1(t)(\rho - x_3(t)) - x_2(t), \\ \dot{x}_3(t) &= x_1(t)x_2(t) - \beta x_3(t),\end{aligned} \tag{5.68}$$

where $\sigma = 10$, $\beta = 8/3$, $\rho = 28$. We make a coordinate change so the Lorenz attractor is located in a unit box by defining

$$\bar{x}_1 := 50x_1, \quad \bar{x}_2 := 50x_2, \quad \bar{x}_3 := 50x_3 + 25. \tag{5.69}$$

The ODE (5.68) *can then be written in the form $\dot{x}(t) = \tilde{f}(x(t), \mathbf{u}(t))$ using $\tilde{f}(x) = [50\sigma(x_2 - x_1), 50x_1(\rho - 50x_3 - 50(25)) - 50x_2, 50^2x_1x_2 - 50\beta x_3 - 25\beta]^T$. Note, as $\tilde{f}$ is independent of any input $u \in U$ without loss of generality we will set $U = \emptyset$. The problem of estimating the Lorenz attractor is then equivalent to the problem of estimating $FR_{\tilde{f}}(\mathbb{R}^n, \mathbb{R}^n, U, \{\infty\})$. In this section we estimate $FR_{\tilde{f}}(\mathbb{R}^n, \mathbb{R}^n, U, \{\infty\})$ by estimating $FR_{\tilde{f}}(X_0, \Lambda, U, \{T\})$ for some $T < \infty$, $\Lambda \subset \mathbb{R}^3$, $X_0 := \{x \in \mathbb{R}^3 : g(x) < 0\}$, and $g \in \mathcal{P}(\mathbb{R}^n, \mathbb{R})$.*

Figure 5.3 shows the set $\{x \in \mathbb{R}^3 : P(x, 0) < 0\}$ where P is the solution to the SOS Optimization Problem (5.59) *for $d = 4$, $T = 0.5$, $f(x) = -\tilde{f}(x)$ for all $x \in \Omega := \{x \in \mathbb{R}^n : h_\Omega(x) \geq 0\}$ and $f(x) = 0$ for all $x \in \partial\Omega$ (freezing the dynamics on $\partial\Omega$ helps to ensure Eq. (5.34) is satisfied, improving numerical performance), $h_U \equiv 0$, $h_\Omega(x) = 1 - x_1^2 - x_2^2 - x_3^2$, $c \equiv 0$, $g(x) = (x_1 + 0.6)^2 + (x_2 - 0.6)^2 + (x_3 - 0.2)^2 - 0.1^2$, $\Lambda = [-0.4, 0.4] \times [-0.5, 0.5] \times [-0.4, 0.6]$, and $w(x, t) = \delta(t)$ where δ is the Dirac delta function. Prop. 5.1 shows P is a sub-VF. Then Cor. 5.1 shows $BR_f(X_0, \Lambda, U, \{T\}) \subseteq$*

$\{x \in \mathbb{R}^3 : P(x,0) < 0\}$ *and hence* $FR_{\tilde{f}}(X_0, \Lambda, U, \{T\}) = BR_f(X_0, \Lambda, U, \{T\}) \subseteq \{x \in \mathbb{R}^3 : P(x,0) < 0\}$ *by Lem. 5.4. Thus the* 0*-sublevel set of* P *contains the forward reachable set. Moreover, Figure 5.3 provides numerical evidence that the* 0*-sublevel set of* P *approximates the Lorenz attractor accurately.*

Note, given an OCP with VF denoted by V^*, Prop. 5.4 shows that the sequence of polynomial solutions to the SOS Problem (5.59), indexed by $d \in \mathbb{N}$, converges to V^* with respect to the L^1 norm as $d \to \infty$. Moreover, Prop. 5.5 shows that this sequence of polynomial solutions yields a sequence of sublevel sets that converges to $\{x \in \mathbb{R}^n : V^*(x,0) \leq 0\}$ with respect to the volume metric as $d \to \infty$. However, Theorem. 5.4 shows reachable sets are characterized by the "strict" sublevel sets of VFs, $\{x \in \mathbb{R}^n : V^*(x,0) < 0\}$. Counterexample A.1 (Chapter A) shows that a sequence of functions that converges to some function V with respect to the L^1 norm may not yield a sequence of "strict" sublevel sets that converges to the "strict" sublevel set of V. Therefore we conclude that the sequence of "strict" sublevel sets obtained by solving the SOS Problem (5.59) may in general not converge to the desired reachable set. However, in practice there is often little difference between the sets $\{x \in \mathbb{R}^n : V^*(x,0) \leq 0\}$ and $\{x \in \mathbb{R}^n : V^*(x,0) < 0\}$. Example 5.2 shows how accurate estimates of reachable sets can be obtained by solving the SOS Problem (5.59). Moreover, these reachable set estimations are guaranteed to contain the true reachable set by Cor. 5.1, a property useful in safety analysis, see Yin *et al.* (2018).

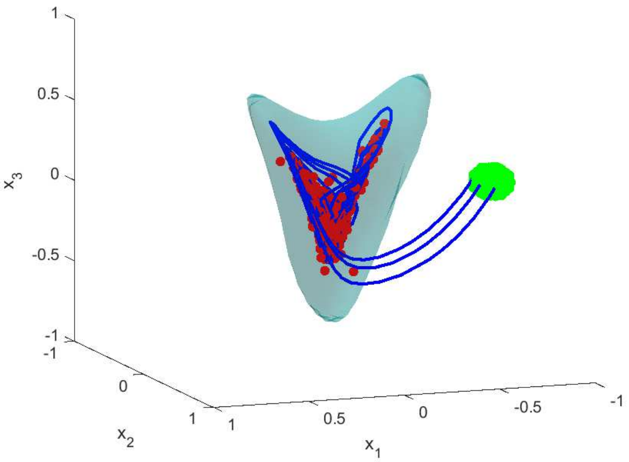

Figure 5.3: Forward reachable set estimation from Example 5.2. The transparent cyan set represents the 0-sublevel set of the solution to the SOS Problem (5.59), the 20^3 green points represent initial conditions, the 20^3 red points represent where the initial conditions transition to after $t = 0.5$ under scaled dynamics from the ODE (5.68) (found using Matlab's `ODE45` function), and the three blue curves represents three sample trajectories terminated at $t = 0.5$ and initialized at three randomly selected green initial conditions.

Chapter 6

PERFORMANCE BOUNDS OF CONTROLLERS CONSTRUCTED FROM APPROXIMATE VALUE FUNCTIONS

> Although this may seem a paradox, all exact science is dominated by the idea of approximation.
>
> Bertrand Russell

6.1 Background and Motivation

Previously in Chapter 5 we considered the problem of solving a nested family of OCPs of Form (5.1). For a given OCP Theorem 5.2 shows that the value function (Defn. 5.5) of the OCP can be used to construct an optimal controller. The goal of Chapter 5 was to solve the following question:

Q1: Can we pose a sequence of convex optimization problems, each yielding a polynomial sub-VF that can be made arbitrarily "close" to the VF of the OCP?

In Chapter 5 we answered **Q1** by proposing a sequence of d-degree SOS programming problem given in Eq. (5.59). Now that we have a technique for approximating the VF of some OCP we may want to use this approximate VF to construct an optimal controller. This leads us to the following question:

Q2: Can we bound the sub-optimality in performance of a controller constructed from some function V by the "distance" between V and the VF of the OCP?

The use of approximate VFs to construct controllers has been well-treated in the literature, although such controllers often: apply only to OCPs with specific structure (typically dynamics are affine in the input variable, see Ribeiro *et al.* (2020) for linearization techniques that approximate non-input affine dynamics by input affine dynamics); do not have associated performance bounds; and/or assume differentiability of the VF. For example, in Jiang and Jiang (2015); Abu-Khalaf and Lewis (2005); Baldi *et al.* (2012, 2015); Zhu *et al.* (2017) policy iteration methods are proposed that alternate between finding approximations of the VF based on a controller and using the approximate VF to synthesizing controller. Also in Abu-Khalaf and Lewis (2005) it was shown that the proposed policy iteration method converges under the rather restrictive assumption that the true VF is differentiable. Alternatively, grid based approaches that synthesize controllers can be found in Kang and Wilcox (2017); Kunisch *et al.* (2004). However, the method in Kang and Wilcox (2017) is only shown to yield a function that converges to the VF but no performance bound is given for the controller. In Kunisch *et al.* (2004), convergence to the optimal controller is demonstrated numerically in certain cases, but no provable performance bound is given.

There are also results within the SOS framework for optimization of polynomials that use approximate VFs to construct controllers. For example, in Leong *et al.* (2014) it was shown that the objective value of a specific class of OCP's using a controller constructed from a given approximate VF was bounded from above by the approximated VF. However, this bound was conservative and no method was given for refinement of the bound. In Jennawasin *et al.* (2011) a method for approximating VFs by sub and super-VFs that are also SOS polynomials is given, however, no VF approximation error bounds or resulting controller synthesis performance bound is given. Alternatively, in Cunis *et al.* (2020) a bilinear SOS optimization framework is proposed, which iterates between finding a Lyapunov function and finding a controller to maximize the region of attraction. However, this work does not consider OCPs or

VFs per se.

Despite this extensive literature, to the best of our knowledge, there exists no way of constructing approximate VFs for which the performance of the associated controller can be proven to be arbitrarily close to optimal (although such bounds exist for discrete time systems over infinite time horizons, see Bertsekas and Tsitsiklis (1995)). For such a result to exist in continuous-time over finite time horizons, then, we need some way of bounding sub-optimality of the performance of the controller based on distance of the approximated VF to the true VF.

To address this need, in this chapter we answer **Q2** by showing that for any V, we can construct a candidate solution to the OCP (5.1), $\mathbf{u}(t) = k(x(t), t)$, given by the controller defined in Eq. (5.3). We then show in Thm. 6.1 that the corresponding objective value of the OCP (5.1) evaluated at $\mathbf{u}$ is within $C\|V^* - V\|_{W^{1,\infty}}$ of the optimal objective, where V^* is the true VF of the OCP and $C > 0$ is given in Eq. (6.3). This result implies approximation of value functions in the $W^{1,\infty}$ norm results in feedback controllers with performance that can be made arbitrarily close to optimality. Note, this result may be of broad interest since it does not require V to be a solution to our proposed SOS Problem (5.59) and hence provides a bound on the sub-optimality of controllers constructed from any approximate VF.

6.2 Performance Bounds

Given an OCP, if an associated differentiable VF is known then a solution to the OCP can be constructed using Theorem 5.2. However, in general, it is challenging to find a VF analytically. Rather than computing a true VF, we consider a candidate VF which is "close" to a true VF under some norm. This motivates us to ask the question: how well will a controller constructed from a candidate VF perform? To answer this question we next define the loss/performance of an input. For $\{c, g, f, \Omega, U, T\} \in$

$\mathcal{M}_{Lip}^{Continuous}$ (Defn. 5.2) we denote the loss/performance function as,

$$L(x_0,\mathbf{u}) := \int_0^T c(\phi_f(x_0,s,\mathbf{u}),\mathbf{u}(s),s)ds + g(\phi_f(x_0,T,\mathbf{u})) \tag{6.1}$$
$$- \inf_{\mathbf{u}\in\mathcal{U}_{\Omega,U,f,T}(x_0,0)} \left\{ \int_0^T c(\phi_f(x_0,s,\mathbf{u}),\mathbf{u}(s),s)ds + g(\phi_f(x_0,T,\mathbf{u}))\right\}.$$

Clearly, $L(x_0,\mathbf{u}) \geq 0$ for all $(x_0,\mathbf{u}) \in \Omega \times \mathcal{U}_{\Omega,U,f,T}(x_0,0)$.

Theorem 6.1. *Consider $\{c,g,f,\mathbb{R}^n,U,T\} \in \mathcal{M}_{Lip}^{Continuous}$. Suppose $J \in C^2(\mathbb{R}^n \times \mathbb{R},\mathbb{R})$ and $\Omega \subset \mathbb{R}^n$ is an open set such that for $x_0 \in \mathbb{R}^n$ we have $FR_f(x_0,\mathbb{R}^n,U,[0,T]) \subseteq \Omega$. In this case we have*

$$L(x_0,\mathbf{u}_J) \leq C||J-V^*||_{W^{1,\infty}(\Omega\times[0,T],\mathbb{R})}, \tag{6.2}$$
$$\text{where } C := 2\max\left\{1,T,T\max_{1\leq i\leq n}\sup_{(x,t)\in\Omega\times U}|f_i(x,u)|\right\}, \tag{6.3}$$

V^* is the unique viscosity solution to the HJB PDE (5.12),

$$\mathbf{u}_J(t) = k_J(\phi_f(x_0,t,\mathbf{u}_J),t), \tag{6.4}$$

and k_J is any function such that

$$k_J(x,t) \in \arg\inf_{u\in U}\{c(x,u,t) + \nabla_x J(x,t)^T f(x,u)\}. \tag{6.5}$$

Proof. Now, for any $J \in C^2(\mathbb{R}^n \times \mathbb{R},\mathbb{R}) \subset LocLip(\mathbb{R}^n \times \mathbb{R},\mathbb{R})$, we wish to show that Eq. (6.2) holds. To do this, we will show that J is the true VF for some modified OCP. Before constructing this modified OCP, for any $F \in LocLip(\mathbb{R}^n \times \mathbb{R},\mathbb{R})$, let us define

$$H_F(x,t,u) := \nabla_t F(x,t) + c(x,u,t) + \nabla_x F(x,t)^T f(x,u),$$
$$\tilde{H}_F(x,t) := \inf_{u\in U} H_F(x,t,u),$$

where $\nabla_t F$ and $\nabla_x F$ are weak derivatives, known to exist by Rademacher's Theorem (Thm. C.4).

Then, by construction, J satisfies the following PDE

$$\nabla_t J(x,t) + \inf_{u \in U} \left\{ c(x,u,t) - \tilde{H}_J(x,t) + \nabla_x J(x,t)^T f(x,u) \right\} = 0 \quad \text{for all } (x,t) \in \mathbb{R}^n \times [0,T]. \tag{6.6}$$

Eq. (6.6) implies that J satisfies the HJB PDE associated with $\{\tilde{c}, \tilde{g}, f, \mathbb{R}^n, U, T\}$, where $\tilde{c}(x,u,t) := c(x,u,t) - \tilde{H}_J(x,t)$ and $\tilde{g}(x) := J(x,T)$. Note that since $c \in LocLip(\Omega \times U \times [0,T], \mathbb{R})$, $f \in LocLip(\Omega \times U, \mathbb{R})$, and $\frac{\partial}{\partial x_i} J \in LocLip(\Omega \times [0,T], \mathbb{R})$ for all $i \in \{1, ...n+1\}$ (since $J \in C^2(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$) it follows that $H_J \in LocLip(\Omega \times U \times [0,T], \mathbb{R})$. By Lemma C.4 we then deduce $\tilde{H}_J \in LocLip(\Omega \times [0,T], \mathbb{R})$ and thus $\{\tilde{c}, \tilde{g}, f, \mathbb{R}^n, U, T\} \in \mathcal{M}_{Lip}^{Continuous}$.

Since H_J is independent of $u \in U$, we have that

$$\arg \inf_{u \in U} \{\tilde{c}(x,u,t) + \nabla_x J(x,t)^T f(x,u)\} = \arg \inf_{u \in U} \{c(x,u,t) + \nabla_x J(x,t)^T f(x,u)\},$$

and therefore we are able to deduce by Theorem 5.2 that $\mathbf{u}_J$ (given in Eq. (6.4)) solves the modified OCP associated with $\{\tilde{c}, \tilde{g}, f, \mathbb{R}^n, U, T\}$ with initial condition $x_0 \in \mathbb{R}^n$. Thus for all $\mathbf{u} \in \mathcal{U}_{\Omega,U,f,T}(x_0, 0)$ we have that

$$\begin{aligned} &\int_0^T \tilde{c}(\phi_f(x_0, s, \mathbf{u}_J), \mathbf{u}_J(s), s) ds + \tilde{g}(\phi_f(x_0, T, \mathbf{u}_J)) \\ &\qquad \leq \int_0^T \tilde{c}(\phi_f(x_0, s, \mathbf{u}), \mathbf{u}(s), s) ds + \tilde{g}(\phi_f(x_0, T, \mathbf{u})). \end{aligned} \tag{6.7}$$

By substituting $\tilde{c}(x,u,t) = c(x,u,t) - \tilde{H}_J(x,t)$ and $\tilde{g}(x) = J(x,T)$ into Inequality (6.7) and noting that $V^*(x,T) = g(x)$, we have the following for all $\mathbf{u} \in \mathcal{U}_{\Omega,U,f,T}(x_0,0)$:

$$\begin{aligned}
&\int_0^T c(\phi_f(x_0,s,\mathbf{u}_J),\mathbf{u}_J(s),s)ds + g(\phi_f(x_0,T,\mathbf{u}_J)) \qquad (6.8)\\
&\qquad\qquad - \int_0^T c(\phi_f(x_0,s,\mathbf{u}),\mathbf{u}(s),s)ds - g(\phi_f(x_0,T,\mathbf{u}))\\
&\leq \int_0^T \tilde{H}_J(\phi_f(x_0,s,\mathbf{u}_J),s) - \tilde{H}_J(\phi_f(x_0,s,\mathbf{u}),s)ds\\
&\qquad\qquad + V^*(\phi_f(x_0,T,\mathbf{u}_J),T) - J(\phi_f(x_0,T,\mathbf{u}_J),T)\\
&\qquad\qquad + J(\phi_f(x_0,T,\mathbf{u}),T) - V^*(\phi_f(x_0,T,\mathbf{u}),T)\\
&< T \operatorname*{ess\,sup}_{s\in[0,T]}\{\tilde{H}_J(\phi_f(x_0,s,\mathbf{u}_J),s) - \tilde{H}_J(\phi_f(x_0,s,\mathbf{u}),s)\}\\
&\qquad\qquad + 2\sup_{y\in\Omega}\{|V^*(y,T) - J(y,T)|\}\\
&\leq T\left(\operatorname*{ess\,sup}_{(y,s)\in\Omega\times[0,T]}\{\tilde{H}_J(y,s)\} - \operatorname*{ess\,inf}_{(y,s)\in\Omega\times[0,T]}\{\tilde{H}_J(y,s)\}\right)\\
&\qquad\qquad + 2\operatorname*{ess\,sup}_{(y,s)\in\Omega\times[0,T]}\{|V^*(y,s) - J(y,s)|\}.
\end{aligned}$$

The second and third inequalities of Eq. (6.8) follow because $\phi_f(x_0,t,\mathbf{u}) \in \Omega$ for all $(t,\mathbf{u}) \in [0,T]\times \in \mathcal{U}_{\Omega,U,f,T}(x_0,0)$ (since it is assumed $FR_f(x_0,\mathbb{R}^n,U,[0,T]) \subseteq \Omega$), and because $\sup_{y\in\Omega}\{|V^*(y,T)-J(y,T)|\} = \operatorname{ess\,sup}_{y\in\Omega}\{|V^*(y,T)-J(y,T)|\}$ holds by Lemma C.6 (since V^* and J are both continuous, and Ω is open).

We now split the remainder of the proof into three parts. In Part 1 of the proof, we derive an upper bound for $\operatorname{ess\,sup}_{(y,s)\in\Omega\times[0,T]}\{\tilde{H}_J(y,s)\}$. In Part 2 of the proof, we find a lower bound for $\operatorname{ess\,inf}_{(y,s)\in\Omega\times[0,T]}\{\tilde{H}_J(y,s)\}$. In Part 3 of the proof, we use the two bounds derived in Part 1 and 2 of the proof, combined with Inequality (6.8) to verify Eq. (6.2) and complete the proof.

Before proceeding with Parts 1 to 3 of the proof we introduce some notation for the set of points where the VF is differentiable,

$$S_{V^*} := \{(x,t) \in \Omega\times[0,T] : V^* \text{ is differentiable at } (x,t)\}.$$

Lemma 5.1 shows that $V^* \in Lip(\Omega \times [0,T], \mathbb{R}) \subset LocLip(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ and Rademacher's Theorem (Thm. C.4) states that Lipschitz functions are differentiable almost everywhere. It follows, therefore, that $\mu((\Omega \times [0,T])/S_{V^*}) = 0$, where μ is the Lebesgue measure.

Part 1 of Proof: For each $(y,s) \in S_{V^*}$ let us consider some family of points $k^*_{y,s} \in U$ such that

$$k^*_{y,s} \in \arg\inf_{u \in U} \left\{ c(y,u,s) + \nabla_x V^*(y,s)^T f(y,u) \right\}.$$

Note, $k^*_{y,s}$ exists for each fixed $(y,s) \in S_{V^*}$ by the extreme value theorem since $U \subset \mathbb{R}^m$ is compact, c, f are continuous, and $\nabla_x V^*$ is independent of $u \in U$ and bounded by Rademacher's Theorem (Thm. C.4).

Now for all $(y,s) \in S_{V^*}$ it follows that

$$\tilde{H}_J(y,s) = \inf_{u \in U} H_J(y,s,u) \leq H_J(y,s,k^*_{y,s}). \tag{6.9}$$

Moreover, since V^* is the viscosity solution to the HJB PDE by Theorem 5.1, we have that

$$H_{V^*}(y, k^*_{y,s}, s) = 0 \quad \text{for all } (y,s) \in S_{V^*}. \tag{6.10}$$

Combing Eqs. (6.9) and (6.10) it follows that

$$\begin{aligned}
&\tilde{H}_J(y,s) \leq H_J(y,s,k^*_{y,s}) - H_{V^*}(y,k^*_{y,s},s) \\
&= \nabla_t J(y,s) - \nabla_t V^*(y,s) + (\nabla_x J(y,s) - \nabla_x V^*(y,s))^T f(y,k^*_{y,s}) \\
&\leq |\nabla_t J(y,s) - \nabla_t V^*(y,s)| + \max_{1 \leq i \leq n} |f_i(y,k^*_{y,s})| \sum_{i=1}^n \left| \frac{\partial}{\partial x_i}(J(y,s) - V^*(y,s)) \right| \qquad (6.11)
\end{aligned}$$

for all $(y,s) \in S_{V^*}$. As Eq. (6.11) is satisfied for all $(y,s) \in S_{V^*}$ and $\mu((\Omega \times$

$[0,T])/S_{V^*}) = 0$ it follows Eq. (6.11) holds almost everywhere. Therefore

$$\begin{aligned}&\operatorname*{ess\,sup}_{(y,s)\in\Omega\times[0,T]} \tilde{H}_J(y,s) \\ &\qquad \le \max\left\{1, \max_{1\le i\le n} \sup_{(x,t)\in\Omega\times U} |f_i(x,u)|\right\} ||V^* - J||_{W^{1,\infty}(\Omega\times[0,T])}.\end{aligned} \tag{6.12}$$

<u>Part 2 of Proof:</u> If k_J satisfies Eq. (6.5), then

$$\tilde{H}_J(y,s) = \inf_{u\in U} H_J(y,s,u) = H_J(y,s,k_J(y,s)) \text{ for all } (y,s)\in S_{V^*}. \tag{6.13}$$

Moreover, since V^* is a viscosity solution to the HJB PDE (5.12), we have by Theorem 5.1 that

$$H_{V^*}(y,s,k_J(y,s)) \ge \inf_{u\in U} H_{V^*}(y,s,u) = 0 \text{ for all } (y,s)\in S_{V^*}. \tag{6.14}$$

Combining Eqs. (6.13) and (6.14) it follows that

$$\begin{aligned}&\tilde{H}_J(y,s) \ge H_J(y,s,k_J(y,s)) - H_{V^*}(y,s,k_J(y,s)) \\ &= \nabla_t J(y,s) - \nabla_t V^*(y,s) + (\nabla_x J(y,s) - \nabla_x V^*(y,s))^T f(y,k_J(y,s)) \\ &\ge -|\nabla_t J(y,s) - \nabla_t V^*(y,s)| \\ &\qquad - \max_{1\le i\le n} |f_i(y,k_J(y,s))| \sum_{i=1}^n \left|\frac{\partial}{\partial x_i}(J(y,s) - V^*(y,s))\right|\end{aligned} \tag{6.15}$$

for all $(y,s)\in S_{V^*}$. Therefore, since $\mu((\Omega\times[0,T])/S_{V^*}) = 0$, we have that Inequality (6.15) holds almost everywhere. Thus

$$\operatorname*{ess\,inf}_{(y,s)\in\Omega\times[0,T]} \tilde{H}_J(y,s) \ge -\max\left\{1, \max_{1\le i\le n} \sup_{(x,t)\in\Omega\times U} |f_i(x,u)|\right\} ||V^* - J||_{W^{1,\infty}(\Omega\times[0,T])}. \tag{6.16}$$

<u>Part 3 of Proof:</u>

Combining Inequalities (6.8), (6.12) and (6.16) it follows that

$$\int_0^T c(\phi_f(x_0,s,\mathbf{u}_J),\mathbf{u}_J(s),s)ds + g(\phi_f(x_0,T,\mathbf{u}_J)) \tag{6.17}$$
$$- \int_0^T c(\phi_f(x_0,s,\mathbf{u}),\mathbf{u}(s),s)ds - g(\phi_f(x_0,T,\mathbf{u}))$$
$$< C||J-V^*||_{W^{1,\infty}} \text{ for all } \mathbf{u} \in \mathcal{U}_{\Omega,U,f,T}(x_0,0),$$

where $C := 2\max\left\{1, T, T\max_{1\le i\le n}\sup_{(x,t)\in\Omega\times U}|f_i(x,u)|\right\}$. Now as Inequality (6.17) holds for all $\mathbf{u} \in \mathcal{U}_{\Omega,U,f,T}(x_0,0)$ we can take the infimum and deduce Inequality (6.2).

□

6.3 Application: Using SOS to Approximate VFs for Controller Construction

Given an OCP, in Theorem 6.1 we showed that the performance of a controller constructed from a candidate VF is bounded by the $W^{1,\infty}$ norm between the true VF of the OCP and the candidate VF. We next demonstrate through numerical examples that the performance of a controller constructed from a typical solutions to the SOS Problem (5.59) is significantly higher than that predicted by this bound.

Consider tuple $\{c,g,f,\mathbb{R}^n,U,T\} \in \mathcal{M}_{Poly}^{Continuous}$, where the cost function is of the form $c(x,u,t) = c_0(x,t) + \sum_{i=1}^m c_i(x,t)u_i$, the dynamics are of the form $f(x,u) = f_0(x) + \sum_{i=1}^m f_i(x)u_i$, and the input constraints are of the form $U = [a_1,b_1] \times ... \times [a_m,b_m]$. Since any rectangular set can be represented as $U = [-1,1]^m$ using the substitution $\tilde{u}_i = \frac{2u_i - 2b_i}{b_i - a_i}$ for $i \in \{1,...,m\}$, without loss of generality we assume $U = [-1,1]^m$. Now, given an OCP associated with $\{c,g,f,\mathbb{R}^n,U,T\} \in \mathcal{M}_{Poly}^{Continuous}$ suppose $V \in C^1(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$ solves the HJB PDE (5.12), then by Theorem 5.2 a solution to the OCP initialized at $x_0 \in \mathbb{R}^n$ can be found as

$$\mathbf{u}^*(t) := k(\phi_f(x_0, t, \mathbf{u}^*), t), \text{ where} \tag{6.18}$$

$$k(x,t) \in \arg \inf_{u \in [-1,1]^m} \left\{ \sum_{i=1}^m c_i(x,t) u_i + \nabla_x V(x,t)^T f_i(x) u_i \right\}. \tag{6.19}$$

Since the objective function in Eq. (6.19) is linear in the decision variables $u \in \mathbb{R}^m$, and since the constraints have the form $u_i \in [-1, -1]$, it follows that Eqs. (6.18) and (6.19) can be reformulated as

$$\mathbf{u}^*(t) := k(\phi_f(x_0, t, \mathbf{u}^*), t), \text{ where} \tag{6.20}$$

$$k_i(x,t) = -\operatorname{sign}(c_i(x,t) + \nabla_x V(x,t)^T f_i(x)). \tag{6.21}$$

In the following numerical examples we approximately solve OCPs of this form (with cost functions and dynamics affine in the input variable) by constructing a controller from the solution, P_d, to the SOS Problem (5.59) for some $d \in \mathbb{N}$. We construct such controllers by replacing V with P_d in Eqs. (6.20) and (6.21). We will consider OCPs with no state constraints and initial conditions inside some set $\Lambda \subseteq \mathbb{R}^n$. We select $\Omega = B_R(0)$ with $R > 0$ sufficiently large enough so Eq. (5.34) is satisfied. That is, no matter what control we use, the solution map starting from any $x_0 \in \Lambda$ will not be able to leave the state constraint set Ω. In this case the solution to the state constrained problem, $\{c, g, f, \Omega, U, T\}$, is equivalent to the solution of the state unconstrained problem, $\{c, g, f, \mathbb{R}^n, U, T\}$.

To evaluate the performance of our constructed controller, $\mathbf{u}$, we approximate the objective/cost function of the OCP evaluated at $\mathbf{u}$ (ie the cost associated with $\mathbf{u}$)

using the Riemann sum:

$$\int_0^T c(\phi_f(x_0, t,\mathbf{u}), t)dt + g(\phi_f(x_0, T, \mathbf{u})) \tag{6.22}$$
$$\approx \sum_{i=1}^{N-1} c(\phi_f(x_0, t_i, \mathbf{u}), t_i)\Delta t_i + g(\phi_f(x_0, t_N, \mathbf{u})),$$

where $0 = t_0 < ... < t_N = T$, $\Delta t_i = t_{i+1} - t_i$ for all $i \in \{1, ..., N-1\}$, and $\{\phi_f(x_0, t_i, \mathbf{u})\}_{i=0}^N$ can be found using Matlab's `ode45` function.

Example 6.1. *Let us consider the following OCP from Jacobson* et al. *(1970):*

$$\min_{\mathbf{u}} \int_0^5 x_1(t)dt \tag{6.23}$$
$$\textit{subject to: } \begin{bmatrix} \dot{x_1}(t) \\ \dot{x_2}(t) \end{bmatrix} = \begin{bmatrix} x_2(t) \\ \mathbf{u}(t) \end{bmatrix}, \ \mathbf{u}(t) \in [-1, 1] \textit{ for all } t \in [0, 5].$$

We associate this problem with the tuple $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Poly}^{Continuous}$ *where* $c(x,t) = x_1$, $g(x) \equiv 0$, $f(x,u) = [x_2, u]^T$, $U = [-1, 1]$, *and* $T = 5$. *By solving the SOS Optimization Problem* (5.59) *for* $d = 3$, $\Lambda = [-0.6, 0.6] \times [-1, 1]$, $w(x,t) \equiv 1$, $h_\Omega(x) = 10^2 - x_1^2 - x_2^2$ *and* $h_U(u) = 1 - u^2$, *it is possible to obtain a polynomial sub-value function* P. *By replacing* V *with* P *in Eqs.* (6.20) *and* (6.21) *it is then possible to construct a controller,* k_P, *that yields a candidate solution to the OCP as* $\tilde{\mathbf{u}}(t) = k_P(x(t), t)$.

For initial condition $x_0 = [0, 1]^T$ *we use Matlab's* ***ode45*** *to find the set* $\{\phi_{\tilde{f}}(x_0, t, \tilde{\mathbf{u}}) \in \mathbb{R}^2 : t \in [0, T]\}$ *(recalling* ϕ_f *denotes the solution map (Defn. 5.4)), which is shown in the phase plot in Figure 6.1. For* $N = 10^8$ *Eq.* (6.22) *was used to find the cost associated with a fixed input,* $\mathbf{u}(t) \equiv 1$, *as* 354.17, *whereas the cost of using* $\mathbf{u}(t) \equiv -1$ *was found to be* 41.67. *The cost of using our derived input* $\tilde{\mathbf{u}}$ *was found to be* 0.2721, *an improvement when compared to the cost* 0.2771 *found in Jacobson* et al. *(1970). Note, it may be possible that the results of the algorithm proposed in Jacobson* et al.

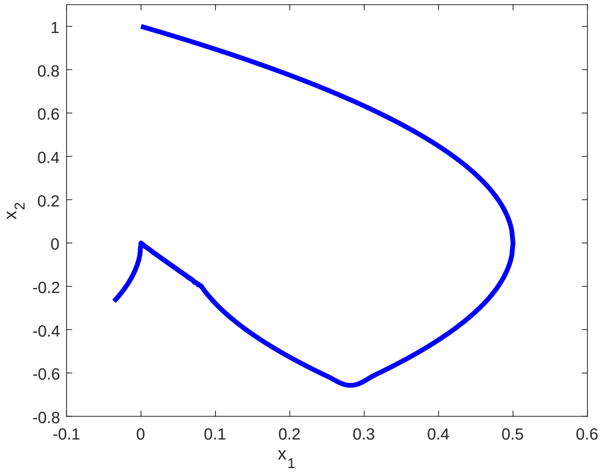

Figure 6.1: The phase plot of Example 6.1 found by constructing the controller given in Eq. (6.18) using the solution to the SOS Problem (5.59).

(1970) may be improved by selecting different tunning parameters of the algorithm. We have assumed that the authors of Jacobson et al. *(1970) have selected the tunning parameters for which their algorithm performs best.*

Example 6.2. *Consider an OCP found in Jacobson* et al. *(1970) and Dadebo* et al. *(1998) which has the same dynamics as Eq.* (6.23) *but a different cost function. The associated tuple is* $\{c, g, f, \Omega, U, T\} \in \mathcal{M}_{Poly}^{Continuous}$ *where* $c(x,t) = x_1^2 + x_2^2$*,* $g(x) \equiv 0$*,* $f(x,u) = [x_2, u]^T$*,* $U = [-1,1]$*, and* $T = 5$*. By solving the SOS Optimization Problem* (5.59) *for* $d = 4$*,* $\Lambda = [-0.5, 1.1] \times [-1.1, .5]$*,* $w(x,t) \equiv 1$*,* $h_\Omega(x) = 10^2 - x_1^2 - x_2^2$ *and* $h_U(u) = 1 - u^2$*, we obtain the polynomial sub-VF* P*. Similarly to Example 6.1 we construct a controller from the polynomial sub-VF* P *using Eqs.* (6.20) *and* (6.21)*. Using Eq.* (6.22)*, the fixed input* $\mathbf{u}(t) \equiv +1$ *was found to have cost* 446.03*. The fixed input* $\mathbf{u}(t) \equiv -1$ *cost was found to be* 67.48*. The controller derived from* P *was found to have cost* 0.7255*, an improvement compared to a cost of* 0.75041 *found in Dadebo* et al. *(1998) and* 0.8285 *found in Jacobson* et al. *(1970).Also note, it may be possible that the results of the algorithms proposed in Jacobson* et al. *(1970); Dadebo* et al. *(1998) may be improved by selecting different tunning parameters of the algorithms.*

We have assumed that the authors of Jacobson et al. *(1970); Dadebo* et al. *(1998) have selected the tunning parameters for which their algorithm performs best.*

Example 6.3. *As in Moyalan* et al. *(2021) let us consider the (scaled) Van der Pol oscillator:*

$$\dot{x}_1(t) = 2x_2(t), \tag{6.24}$$
$$\dot{x}_2(t) = 10x_2(t)(0.21 - 1.2^2 x_1(t)) - 0.8x_1(t) + u(t),$$

where $u(t) \in [-1,1]$. Let us consider OCPs of Form (5.5) *governed by the dynamics given in Eq.* (6.24) *with $\Omega = \mathbb{R}^n$, $U = [-1,1]$ and cost functions of the form $c(x,u,t) = ||x-q||_2^2$ and $g(x) = ||x-q||_2^2$, where $q = [-0.4, 0]$ or $q = [0;0]$. Clearly any solution to the OCP is an input $\mathbf{u}$ that forces the systems trajectories towards the point $q \in \mathbb{R}^2$.*

By solving the SOS Optimization Problem (5.59) *twice for $q = [-0.4, 0]$ and $q = [0;0]$ with $d = 14$, $T = 10$, f, c, g as defined previously, $\Lambda = [-1,1]^2$, $w(x,t) \equiv 1$, $h_\Omega(x) = 2.1 - x_1^2 - x_2^2$, and $h_U(u) = 1 - u^2$ we obtain polynomial sub-value functions P_1 and P_2 respectively. By replacing V with P_i, for $i \in \{1,2\}$, in Eqs.* (6.20) *and* (6.21) *we then construct controllers, k_{P_i}, that yield candidate solution to the OCPs, $\tilde{\mathbf{u}}_\mathbf{i}(t) = k_{P_i}(x(t),t)$ $i \in \{1,2\}$.*

For initial condition $x_0 = [0.75, 0.75]^T$ and terminal time $T = 10$ we use Matlab's ***ode45*** *to find the curves $\{\phi_{\tilde{f}}(x_0, t, \tilde{\mathbf{u}}_\mathbf{i}) \in \mathbb{R}^2 : t \in [0,T]\}$ for $i = 1,2$ (recalling ϕ_f denotes the solution map (Defn. 5.4)), which is shown as the blue and red curves respectively in the phase plot in Figure 6.2. Moreover, for comparison we have also plotted the solution trajectory under the fixed input $\mathbf{u}(t) \equiv 0$ as the green curve, which demonstrates the shape of the Van-der-Pol limit cycle. As expected the input $\mathbf{u}_1$ drives the system to the point $q = [-0.4;0]$ with terminal state, shown as the black dot in Figure 6.2, as $[-0.430; 0.112]$. Moreover, the input $\mathbf{u}_2$ drives the system to the point $q = [0;0]$ with terminal state $[-0.012; 0.007]$.*

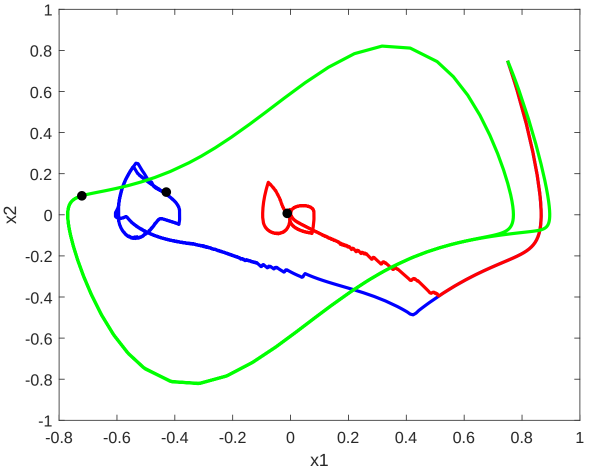

Figure 6.2: Graph showing the phase plot of Example 6.3 found by constructing controllers given by Eq. (6.20) using the solution to the SOS Problem (5.59). The blue curve shows the $T = 10$ solution trajectory initialized at $(0.75, 0.75)$ of the ODE (6.24) driven by the controller found by considering costs $c(x, u, t) = ||x - q||_2^2$ and $g(x) = ||x - q||_2^2$, where $q = (-0.4, 0)$. The red curve shows the $T = 10$ solution trajectory initialized at $(0.75, 0.75)$ of the ODE (6.24) driven by the controller found by considering the same costs but with $q = (0, 0)$. The green curve is the $T = 10$ solution trajectory initialized at $(0.75, 0.75)$ of the ODE (6.24) under the input $\mathbf{u}(t) \equiv 0$. Terminal states for each trajectory are given by the black dots. Costs associated with each trajectory can be found in Table 6.1.

Table 6.1 shows the $T = 10$ *cost of using various inputs when* $q = [-0.4, 0]$ *or* $q = [0; 0]$. *All costs were calculated using Eq.* (6.22) *for initial condition* $[0.75; 0.75]$. *The costs of using* $\mathbf{u}_1$ *and* $\mathbf{u}_2$ *are shown in the* $\mathbf{u}_{SOS}$ *row under columns* $q = [-0.4, 0]$ *or* $q = [0; 0]$ *respectively. As expected the inputs derived using SOS out perform (have lower cost) compared to constant inputs.*

Table 6.1: This table shows the corresponding costs of various inputs for the OCPs of Form (5.5) given in Example 6.3.

Input $\mathbf{u}$	Cost for $q = [-0.4; 0]$	Cost for $q = [0; 0]$
$\mathbf{u}_{SOS}$	0.21473	0.078919
$\mathbf{u}(t) \equiv 0$	0.84466	1.0037
$\mathbf{u}(t) \equiv +1$	1.1824	2.444
$\mathbf{u}(t) \equiv -1$	4.5615	2.4681

Chapter 7

CONVERSE LYAPUNOV FUNCTIONS AND CONVERGING INNER APPROXIMATIONS TO MAXIMAL REGIONS OF ATTRACTION OF NONLINEAR SYSTEMS

> In order to solve this differential equation you look at it until a solution occurs to you.
>
> George Polya

7.1 Background and Motivation

For a given equilibrium point, a Region of Attraction (ROA) of a nonlinear Ordinary Differential Equation (ODE) is defined as a set of initial conditions for which the solution map of the ODE tends to that equilibrium point. The *maximal* ROA of an equilibrium point, meanwhile, is defined as the ROA which contains all other ROAs of that equilibrium point. Specifically, for an ODE $\dot{x}(t) = f(x(t))$, we denote the solution map (known to exist when f is Lipschitz continuous) of the ODE by $\phi_f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ which satisfies

$$\frac{d}{dt}\phi_f(x,t) = f(\phi_f(x,t)) \text{ for all } x \in \mathbb{R}^n \text{ and } t \geq 0,$$
$$\phi_f(x,0) = x \text{ for all } x \in \mathbb{R}^n,$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is such that $f(0) = 0$. The maximal ROA is then defined as

$$ROA_f := \{x \in \mathbb{R}^n : \lim_{t \to \infty} ||\phi_f(x,t)||_2 = 0\}.$$

The problem of computing sets which accurately approximate the maximal ROA with respect to some set metric plays a central role in the stability analysis of many engineering applications. For instance, knowledge of the ROA provides a metric for the susceptibility of the F/A-18 Hornet aircraft experiencing an unsafe out of control flight departure phenomena, called falling leaf mode, see Chakraborty *et al.* (2011a,b).

If the matrix $A \in \mathbb{R}^{n \times n}$ is Hurwitz (the real part of the eigenvalues of A are all negative) then the associated linear system, with vector field $f(x) = Ax$, has a maximal ROA that can be found exactly as $ROA_f = \mathbb{R}^n$. In the more general case of nonlinear systems there is no known general analytical formula for ROA_f. However, for particular nonlinear systems, such as those arising from gradient flow dynamics, the maximal ROA can be expressed analytically, see Mohammadi *et al.* (2018). In the absence of an analytical formula for ROA_f in recent years there has been considerable interest in discovering numerical methods for approximating ROA_f rather than finding ROA_f exactly.

Lyapunov‘s second method is arguably the most widely used technique for finding ROAs associated with an ODE, see Kellett (2015). Rather than solving the ODE directly to find a closed form expression of the solution map, ROAs can be computed indirectly by searching for a “generalized energy function”, called a Lyapunov function. A Lyapunov function of an ODE is any function that is positive everywhere, apart from the origin where it is zero, and is strictly decreasing along the solution map of the ODE. Specifically, if we can find a function V such that $V(0) = 0$ and $V(x) > 0$ for all $x \neq 0$, then if $\nabla V(x)^T f(x)$ is negative over the sublevel set $\{x \in \mathbb{R}^n : V(x) \leq a\}$ we have that $\{x \in \mathbb{R}^n : V(x) \leq a\} \subseteq ROA_f$ is a ROA, see Hahn (1967). For linear systems, $f(x) = Ax$ where $A \in \mathbb{R}^{n \times n}$, a necessary and sufficient condition for $ROA_f = \mathbb{R}^n$ is that there exists a quadratic Lyapunov function of form $V(x) = x^T P x$ where $P > 0$. Thus, in this case, the problem of finding the maximal ROA of a linear system is reduced to solving the Linear Matrix Inequality (LMI) $A^T P + PA < 0$ for $P > 0$.

In the case of nonlinear systems a common approach for finding Lyapunov functions has been to generalize the search from quadratic functions, $V(x) = x^T Px$, to Sum-of-Square (SOS) polynomials functions, $V(x) = Z_d(x)^T PZ_d(x)$ where Z_d is the degree $d \in \mathbb{N}$ monomial vector. Then, to find a Lyapunov function we must solve an SOS optimization problem, rather than solving an LMI (as was the case for linear systems). Over the years, many SOS optimization problems have been proposed for ROA estimation, see Tan and Packard (2008); Zheng *et al.* (2018); Anderson and Papachristodoulou (2015); Colbert and Peet (2018). Recently in Cunis *et al.* (2020), SOS was used to estimate the region of attraction of an uncrewed aircraft; in Valmorbida and Anderson (2017) an SOS based algorithm was proposed to construct a rational Lyapunov function that yields an estimate of the ROA; in Awrejcewicz *et al.* (2021) a recursive procedure for constructing the polynomial Lyapunov functions was proposed.

Despite the recent success of modern attempts to find accurate approximations of the maximal ROA, to the best of our knowledge, a numerical algorithm that can be proven to provide an approximation of the maximal ROA arbitrarily well with respect to any set metric has yet to be proposed. Many of the current numerical methods for finding ROAs use SOS programming to find polynomial Lyapunov functions. However, barring any assumptions on the existence of a sufficiently smooth Lyapunov function, it is currently unknown how well polynomial functions can approximate the maximal ROA of a given nonlinear ODE. Concerningly, several counter examples, found in Ahmadi *et al.* (2011); Ahmadi and El Khadir (2018), show that there exist globally asymptotically stable systems ($ROA_f = \mathbb{R}^n$) with polynomial vector fields, but for which there does not exist any associated polynomial Lyapunov function that can certify global asymptotic stability (not even locally in the case of Ahmadi and El Khadir (2018)). On the other hand, for systems that are locally exponentially stable it has been shown in Peet (2009) that there always exists a polynomial Lyapunov function that can certify local exponential stability. This result has been extended in Leth *et al.* (2017) to show that there always exist polynomial Lyapunov functions that can certify a system is locally rationally stable (a weaker form of stability than exponential stability) under the assumption that there exists a smooth Lyapunov function (that need not be polynomial). Furthermore, for systems with homogeneous vector fields it has been shown in Ahmadi and El Khadir (2019) that there always exists a rational Lyapunov function that is the solution to some SOS problem.

For work that is concerned with using SOS to approximate the maximal ROA of locally exponentially stable systems we mention Jones *et al.* (2017). It was shown in Jones *et al.* (2017) that under the assumption that there exists a sufficiently smooth Lyapunov function, there exists a polynomial Lyapunov function that yields a sublevel set that approximates ROA_f arbitrarily well with respect to the Hausdorff metric. We note that the conservatism of the assumption that there exists a sufficiently smooth Lyapunov function is currently unknown. Moreover, the proposed algorithm for approximating the maximal ROA found in Jones *et al.* (2017) is only conjectured to yield an arbitrarily close approximation of the maximal ROA but has yet to be proven.

In this chapter our goal is to design an algorithm that approximates the maximal ROA of a given locally exponentially stable ODE arbitrarily well. In order to achieve this goal we propose a new converse Lyapunov function (given in Eq. (7.11)) whose 1-sublevel set is equal to ROA_f. Our proposed converse Lyapunov function is shown to be sufficiently smooth - meaning it can be approximated by a polynomial. After proposing such a converse Lyapunov function, we are then able to design a sequence of SOS Optimization Problems (7.69) and prove that this sequence yields a sequence of polynomials that converges to our proposed converse Lyapunov function uniformly from above in the L^1 norm. Finally, we show that since this sequence of polynomials converges to our proposed converse Lyapunov function in the L^1 norm from above, their associated sequence of 1-sublevel sets must also converge in the volume metric to the 1-sublevel set of our proposed converse Lyapunov function (which is equal to the maximal region of attraction of the ODE). Therefore, for a given locally exponentially stable ODE, the goal of this chapter is: 1) To establish the existence of a globally Lipschitz continuous converse Lyapunov function whose 1-sublevel set is equal to ROA_f. 2) To propose the first numerical algorithm that can approximate the maximal ROA arbitrarily well with respect to some set metric. Furthermore, our numerical algorithm yields an inner approximation of ROA_f (that is solution maps initialized inside our approximation of ROA_f asymptotically coverage to the origin); a useful property for the safety analysis of dynamical systems.

The rest of this chapter is organized as follows. In Section 7.2 we define the maximal region of attraction of an ODE in terms of the solution map of the ODE. In Section 7.3 we formulate the problem of approximating the region of attraction as an optimization problem. In Section 7.4 we propose a globally Lipschitz continuous Lyapunov function that characterizes the maximal region of attraction. In Section 7.5 we propose a convex optimization problem for the approximation of our proposed converse Lyapunov function in the L^1-norm. In Section 7.7 we tighten this optimization problem to an SOS programming problem. Finally, several numerical examples are given in Section 7.8 and our conclusion is given in Section 7.9.

7.2 Regions of Attraction are Defined Using Solution Maps of Nonlinear ODEs

Consider a nonlinear Ordinary Differential Equation (ODE) of the form

$$\dot{x}(t) = f(x(t)), \quad x(0) = x_0 \in \mathbb{R}^n, \quad t \in [0, \infty), \tag{7.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is the vector field and $x_0 \in \mathbb{R}^n$ is the initial condition. Note that, throughout this chapter we will assume $f(0) = 0$ so the origin is an equilibrium point (Note that in Chapter 8 this may not be the case).

Given $D \subset \mathbb{R}^n$, $I \subset [0, \infty)$, and an ODE (7.1) we say any function $\phi_f : D \times I \to \mathbb{R}^n$ satisfying

$$\begin{aligned} &\frac{\partial \phi_f(x,t)}{\partial t} = f(\phi_f(x,t)) \text{ for } (x,t) \in D \times I, \qquad (7.2)\\ &\phi_f(x,0) = x \text{ for } x \in D,\\ &\phi_f(\phi_f(x,t),s) = \phi_f(x,t+s) \text{ for } x \in D\ t,s \in I \text{ with } t+s \in I, \end{aligned}$$

is a solution map of the ODE (7.1) over $D \times I$. For simplicity throughout this chapter we will assume there exists a unique solution map to the ODE (7.1) over all $(x,t) \in \mathbb{R}^n \times [0,\infty)$ (uniqueness and existence of a solution map sufficient for the purposes of this chapter, such as for initial conditions inside some invariant set, like the Region of Attraction (7.4), and for all $t \geq 0$, can be shown to hold under minor smoothness assumption on f, see Khalil (1996)).

We now use the solution map of the ODE (7.1) to define notions of stability.

Definition 7.1. *We say the set $U \subset \mathbb{R}^n$ is an asymptotically stable set of the ODE* (7.1) *if:*

1. *U contains a neighborhood of the origin.*
2. *For any $x \in U$ we have that $\phi_f(x,t) \in U$ for all $t \in [0,\infty)$ and $\lim_{t\to\infty} \phi_f(x,t) = 0$.*

Furthermore, if there also exists $\delta, \mu > 0$ such that for any $x \in U$ we have that

$$||\phi_f(x,t)||_2 \leq \mu e^{-\delta t} ||x||_2 \textit{ for all } t \geq 0, \qquad (7.3)$$

then we say $U \subset \mathbb{R}^n$ is an exponentially stable set of the ODE (7.1)*.*

Definition 7.2. *The (Maximal) Region of Attraction (ROA) of the ODE* (7.1) *is defined as the following set:*

$$ROA_f := \{x \in \mathbb{R}^n : \lim_{t \to \infty} ||\phi_f(x,t)||_2 = 0\}. \tag{7.4}$$

The ROA of the ODE (7.1) can be thought of as the "maximal" asymptotically stable set. That is if $U \subset \mathbb{R}^n$ is an asymptotically stable set of the ODE (7.1) then $U \subseteq ROA_f$. Moreover, as we will show next, the ROA is an open set.

Lemma 7.1 (Lemma 8.1 in Khalil (1996))**.** *Consider an ODE of Form* (7.1)*. The set ROA_f (Defined in Eq.* (7.4)*) is open.*

Before proceeding we introduce some useful notation for the η-ball set entry times of solution maps. For a given function $\phi_f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, $x \in ROA_f$, and $\eta > 0$ we denote

$$F_\eta(x) := \inf\{t \geq 0 : \phi_f(x,t) \in B_\eta(0)\}. \tag{7.5}$$

We now state two important properties of solution maps used in many of the proofs presented in this chapter.

Lemma 7.2 (Exponential divergence of solution maps. Page 392 in Hirsch *et al.* (2004))**.** *Suppose $f \in C^2(\mathbb{R}^n, \mathbb{R})$ and there exists $\theta, R > 0$ such that $||D^\alpha f(x)||_2 < \theta$ for all $x \in B_R(0)$ and any $||\alpha||_1 \leq 2$, where $\alpha \in \mathbb{N}^n$. Then the solution map satisfies the following inequality*

$$||\phi_f(x,t) - \phi_f(y,t)||_2 \leq e^{\theta t}||x-y||_2 \textit{ for } t \geq 0 \textit{ and } x, y \in ROA_f. \tag{7.6}$$

Lemma 7.3 (Smoothness of the solution map. Page 149 in Hirsch *et al.* (2004))**.** *If $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ then the solution map is such that $\phi_f \in C^1(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$.*

7.3 The Problem of Approximating The ROA

Consider $f \in C^2(\mathbb{R}^n, \mathbb{R}^n)$. The goal of this chapter is to compute an optimal (with respect to some set metric) inner approximation of ROA_f (given in Defn. 7.2). That is, we would like to solve the following optimization problem:

$$\min_{X \in \mathcal{C}} \{D(ROA_f, X)\} \tag{7.7}$$
$$\text{such that } X \subseteq ROA_f,$$

where $\mathcal{C} \subset P(\mathbb{R}^n)$ is some constraint set (recalling from Chapter 2 that $P(\mathbb{R}^n)$ is the power set of $\mathbb{R}^n$) and $D : \{Y : Y \subset \mathbb{R}^n\} \times \{Y : Y \subset \mathbb{R}^n\} \to \mathbb{R}$ is some set metric. Note, if the constraint set contains all subsets of $\mathbb{R}^n$, that is $\mathcal{C} = P(\mathbb{R}^n)$, then trivially the optimization problem is solved by the region of attraction, $X = ROA_f$.

The optimization problem given in Eq. (7.7) is fundamentally "geometric in nature" since it is solved by finding a subset of Euclidean space, $X \subset \mathbb{R}^n$. In this chapter we reformulate the optimization problem given in Eq. (7.7) as an optimization problem that is "algebraic in nature", being solved by a function rather than a set. In order to formulate such an "algebriac" optimization problem we first propose a converse Lyapunov function (given later in Eq. (7.11)), denoted here as W, whose 1-sublevel set is equal to ROA_f; that is $ROA_f = \{x \in \mathbb{R}^n : W(x) < 1\}$. Then rather than finding the set "closest" to ROA_f, we find the "closest" d-degree polynomial to W with respect to the L^1 norm. Thus we consider the following "algebriac" problem:

$$P_d \in \arg \min_{J \in \mathcal{P}_d(\mathbb{R}^n, \mathbb{R})} \int_\Lambda |J(x) - W(x)| dx \tag{7.8}$$
$$\text{such that } W(x) \leq J(x) \text{ for all } x \in \Omega,$$

where $ROA_f \subseteq \Lambda \subseteq \Omega \subset \mathbb{R}^n$. Then, Cor. A.1 (found in Appendix A) can be used to show that $\{x \in \Lambda : P_d(x) < 1\}$ converges to $\{x \in \Lambda : W(x) < 1\} = ROA_f$ as $d \to \infty$

with respect to the volume metric (given in Eq. (A.1)).

Solving the optimization problem given in Eq. (7.8) has the following challenges:

1. Does there exist a converse Lyapunov function $W : \mathbb{R}^n \to \mathbb{R}$ such that $ROA_f = \{x \in \mathbb{R}^n : W(x) < 1\}$?
2. Can the constraint, $W(x) \leq J(x)$ for all $x \in \Omega$, be tightened to a convex constraint without necessarily having an analytical formula for W?
3. Does the solution, P_d, tend towards W with respect to the L^1 norm as $d \to \infty$?

In the next section we tackle the first of these challenges. We propose a converse Lyapunov function, W, whose 1-sublevel set is equal to ROA_f. Then, in Sec. 7.5 we tackle the second challenge; we propose a sufficient condition, in the form of a linear partial differential inequality, that when satisfied by a function J implies $W(x) \leq J(x)$ for all $x \in \Omega$. Finally, in Section 7.6, we tackle the third challenge of showing that there exists a sequence of d-degree polynomials, feasible to Opt. (7.8) for $d \in \mathbb{N}$, that converges to W with respect to the L^1 norm. For implementation purposes Opt. (7.8) is then tightened to an SOS optimization problem, given in Eq. (7.69), that can be efficiently numerically solved. The main result of the chapter is then given in Theorem 7.2, showing that our proposed family of d-degree SOS Optimization Problems (7.69) yields a sequence of sets that converge to the region of attraction of a given locally exponentially stable ODE with respect to the volume metric as $d \to \infty$.

7.4 A Globally Lipschitz Continuous Converse Lyapunov Function That Characterizes the ROA

In Vannelli and Vidyasagar (1985) a converse Lyapunov function, called the maximal Lyapunov function, was proposed. It was shown that for any given asymptotically stable ODE there exists a maximal Lyapunov function whose ∞-sublevel set is equal to the region of attraction of the ODE. However, since by definition any maximal Lyapunov function is unbounded outside of the region of attraction it cannot be approximated arbitrarily well (with respect to any norm) by a polynomial over any compact set that contains points outside of the region of attraction (since polynomials are bounded over compact sets). Thus, it is not possible to design an SOS based algorithm that can approximate maximal Lyapunov functions arbitrarily well. To overcome this challenge we propose a new converse Lyapunov function (found in Eq. (7.11)) whose 1-sublevel set is equal to ROA_f, is globally bounded, and is globally Lipschitz continuous. Before introducing our new converse Lyapunov function let us recall the definition of Lipschitz continuity.

Definition 7.3. *Consider sets $\Theta_1 \subset \mathbb{R}^n$ and $\Theta_2 \subset \mathbb{R}^m$. We say the function $F : \Theta_1 \to \Theta_2$ is* ***locally Lipschitz continuous*** *on Θ_1 and Θ_2, denoted $F \in LocLip(\Theta_1, \Theta_2)$, if for every compact set $X \subseteq \Theta_1$ there exists $K > 0$ (that may depend on X) such that for all $x, y \in X$*

$$||F(x) - F(y)||_2 \leq K||x - y||_2. \tag{7.9}$$

If there exists a single $K > 0$ such that Eq. (7.9) holds for all $x, y \in \Theta_1$ we say F is ***globally Lipschitz continuous****, denoted $F \in Lip(\Theta_1, \Theta_2)$.*

We consider two different types of converse Lyapunov functions. The first converse Lyapunov function (given in Eq. (7.10)) is a special case of those first found in Massera (1949) that have the form $V_1(x) := \int_0^\infty G(||\phi_f(x,t)||_2)dt$ for some class K function, $G : [0, \infty) \to [0, \infty)$ (class K is the class of functions which monotonically approach zero at zero). In Vannelli and Vidyasagar (1985) it was shown that for a locally stable ODE, the ∞-sublevel set of V_1 is equal to the region of attraction of the ODE; this Lyapunov function was named the maximal Lyapunov function. In this chapter we only consider locally exponentially stable systems and hence may restrict ourselves to the special case when $G(y) = y^{2\beta}$ for some $\beta \in \mathbb{N}$.

The second converse Lyapunov function we consider (found in Eq. (7.11)) can be thought of as a nonlinear transformation of the first converse Lyapunov function. A function of a similar structure was previously considered in Zubov (1964) and took the form $V_2(x) := \exp\left(-\int_0^\infty G(||\phi_f(x,t)||_2)dt\right) - 1$. Although Zubov (1964) used V_2 to certify the stability of a system, V_2 is not a Lyapunov function in the classical sense since it is not positive everywhere (unlike our proposed converse Lyapunov function in Eq. (7.11)). We note that Zubov (1964) did establish the globally continuity of V_2 but did not show the stronger result that V_2 is Lipschitz continuous.

Now, consider $f \in LocLip(\mathbb{R}^n, \mathbb{R}^n)$, $\lambda > 0$ and $\beta \in \mathbb{N}$. Let us denote the functions $V_\beta : ROA_f \to \mathbb{R}$ and $W_{\lambda,\beta} : \mathbb{R}^n \to \mathbb{R}$ where

$$V_\beta(x) := \int_0^\infty ||\phi_f(x,t)||_2^{2\beta} dt, \tag{7.10}$$

$$W_{\lambda,\beta}(x) := \begin{cases} 1 - \exp(-\lambda \int_0^\infty ||\phi_f(x,t)||_2^{2\beta} dt) \text{ when } x \in ROA_f \\ 1 \text{ otherwise.} \end{cases} \tag{7.11}$$

7.4.1 Converse Lyapunov Functions that Characterize the ROA

The function, V_β, given in Eq. (7.10) is a special case of a class of Lyapunov functions called maximal Lyapunov functions found in Vannelli and Vidyasagar (1985). In the following lemma we will show that V_β tends to infinity for sequences of points approaching the boundary of the region of attraction and is finite inside the region of attraction.

Lemma 7.4. *Consider $f \in LocLip(\mathbb{R}^n, \mathbb{R})$, $\beta \in \mathbb{N}$ and V_β given in Eq.* (7.10)*. Suppose there exists $R, \eta > 0$ such that $ROA_f \subset B_R(0)$ and $B_\eta(0)$ is an exponentially stable set (Defn. 7.1) of the ODE* (7.1)*. Then the following holds.*

1. *For any sequence* $\{x_k\}_{k \in \mathbb{N}} \subset ROA_f$ *such that* $\lim_{k \to \infty} x_k \in \partial ROA_f$ *we have that*
$$\lim_{k \to \infty} V_\beta(x_k) = \infty. \tag{7.12}$$

2. *We have that*
$$x \in ROA_f \text{ if and only if } V_\beta(x) < \infty. \tag{7.13}$$

Proof. We first show Statement 1) in Lem. 7.4 by showing Eq. (7.12) holds. Suppose $\{x_k\}_{k \in \mathbb{N}} \subset ROA_f$ is such that $x^* := \lim_{k \to \infty} x_k \in \partial ROA_f$. Let $0 < \eta_1 < \eta$ and consider $T_k := F_{\eta_1}(x_k)$ (where $F_\eta(x)$ is given in Eq. (7.5)). Since $x_k \in ROA_f$ it follows $T_k < \infty$ for all $k \in \mathbb{N}$. Moreover, it is clear that $||\phi_f(x_k, t)||_2 \geq \eta_1$ for all $t \in [0, T_k)$.

Now,
$$V_\beta(x_k) = \int_0^{T_k} ||\phi_f(x_k,t)||_2^{2\beta} dt + \int_{T_k}^\infty ||\phi_f(x_k,t)||_2^{2\beta} dt \geq \int_0^{T_k} ||\phi_f(x_k,t)||_2^{2\beta} dt \geq \eta_1^{2\beta} T_k. \tag{7.14}$$

We will now show $T_k \to \infty$ as $k \to \infty$ and thus Eq. (7.14) shows Eq. (7.12). For contradiction suppose $\lim_{k \to \infty} T_k \neq 0$, then there exists a bounded subsequence $\{T_{k_n}\}_{n \in \mathbb{N}} \subset \{T_k\}_{k \in \mathbb{N}}$. Now by Theorem C.1 there exists a subsequence of the subsequence $\{T_{k_n}\}_{n \in \mathbb{N}}$, we denote by $\{T_i\}_{i \in \mathbb{N}}$, that converges to a finite limit $T^* := \lim_{i \to \infty} T_i < \infty$. Let us denote the corresponding subsequence of $\{x_k\}_{k \in \mathbb{N}}$ by $\{x_i\}_{i \in \mathbb{N}}$. Since $\lim_{k \to \infty} x_k \to x^*$ and every subsequence of a convergent sequence must converge to the same limit we have $\lim_{i \to \infty} x_i = x^*$.

Since $\phi_f \in C(\mathbb{R}^n \times [0, \infty), \mathbb{R}^n)$ (by Lemma 7.3) we have that
$$||\phi_f(x^*, T^*)||_2 = \lim_{i \to \infty} ||\phi_f(x_i, T_i)||_2 \leq \eta_1 < \eta,$$
and since $B_\eta(0)$ is an exponentially stable set we have that
$$||\phi_f(x^*, T^* + t)||_2^2 = ||\phi_f(\phi_f(x^*, T^*), t)||_2^2 \leq \mu^2 e^{-2\delta t} ||\phi_f(x^*, T^*)||_2^2 \leq \mu^2 \eta^2 e^{-2\delta t}. \tag{7.15}$$

Therefore, Eq. (7.15) implies that

$$\lim_{t\to\infty}||\phi_f(x^*,t)||_2 = \lim_{t\to\infty}||\phi_f(x^*,T^*+t)||_2 = \lim_{t\to\infty}\mu\eta e^{-\delta t} = 0,$$

thus showing $x^* \in ROA_f$. Now ROA_f is an open set (by Lemma 7.1). Therefore if $x^* \in ROA_f$ then $x^* \notin \partial ROA_f$, providing a contradiction that $x^* \in \partial ROA_f$. Hence, Eq. (7.12) holds.

We now Statement 2) in Lem. 7.4 by showing Eq. (7.13) holds. First suppose $x \in ROA_f$. We will now show $V_\beta(x) < \infty$. Since $x \in ROA_f$ we have that $\lim_{t\to\infty}||\phi_f(x,t)||_2 = 0$ and thus it follows there exists $T < \infty$ such that $||\phi_f(x,t)||_2 < \eta$ for all $t \geq T$ implying $F_\eta(x) \leq T < \infty$. Moreover, by properties of the set entry time we have that $||\phi_f(x,F_\eta(x))||_2 \leq \eta$ and since $B_\eta(0)$ is an exponentially stable set we have that,

$$||\phi_f(x,t)||_2 = ||\phi_f(\phi_f(x,F_\eta(x)),t-F_\eta(x))||_2 \leq \mu\eta e^{-\delta(t-F_\eta(x))} \text{ for all } t > F_\eta(x). \tag{7.16}$$

Therefore, using the fact that $ROA_f \subset B_R(0)$ together with Eq. (7.16) we get that,

$$\begin{aligned} V_\beta(x) &= \int_0^{F_\eta(x)} ||\phi_f(x,t)||_2^{2\beta}dt + \int_{F_\eta(x)}^\infty ||\phi_f(x,t)||_2^{2\beta}dt \\ &\leq F_\eta(x)R^{2\beta} + \mu^{2\beta}\eta^{2\beta}\int_{F_\eta(x)}^\infty e^{-2\delta\beta(t-F_\eta(x))}dt \\ &= F_\eta(x)R^{2\beta} + \frac{\mu^{2\beta}\eta^{2\beta}}{2\delta\beta} < \infty. \end{aligned}$$

Now, on the other hand let us now suppose $x \in \mathbb{R}^n$ is such that $V_\beta(x) < \infty$. We will show $x \in ROA_f$. For contradiction suppose $x \notin ROA_f$. Then $\lim_{t\to\infty}||\phi_f(x,t)||_2 \neq 0$. Therefore, there exists $\varepsilon > 0$ such that $||\phi_f(x,t)||_2 > \varepsilon$ for all $t \geq 0$. Thus

$$V_\beta(x) = \int_0^\infty ||\phi_f(x,t)||_2^{2\beta}dt \geq \int_0^\infty \varepsilon^{2\beta}dt = \infty,$$

providing a contradiction that $V_\beta(x) < \infty$. Hence, Eq. (7.13) holds. □

As we will show next, the function, $W_{\lambda,\beta}$, given in Eq. (7.11), can also characterize ROA_f as its 1-sublevel set.

Corollary 7.1. *Consider $f \in LocLip(\mathbb{R}^n, \mathbb{R})$, $\beta \in \mathbb{N}$, $\lambda > 0$ and $W_{\lambda,\beta}$ given in Eq.* (7.11)*. Suppose there exists $R, \eta > 0$ such that $ROA_f \subset B_R(0)$ and $B_\eta(0)$ is an exponentially stable set (Defn. 7.1) to the ODE* (7.1)*. Then the following holds.*

1. *For any sequence $\{x_k\}_{k\in\mathbb{N}} \subset ROA_f$ such that $\lim_{k\to\infty} x_k \in \partial ROA_f$ we have that*
$$\lim_{k\to\infty} W_{\lambda,\beta}(x_k) = 1. \tag{7.17}$$

2. *We have that*
$$ROA_f = \{x \in \mathbb{R}^n : W_{\lambda,\beta}(x) < 1\}. \tag{7.18}$$

Proof. We first show Statement 1) in Cor. 7.1 by showing Eq. (7.17) holds. For $x \in ROA_f$ we have that $W_{\lambda,\beta}(x) = 1 - e^{-\lambda V_\beta(x)}$. Moreover, e^x is a continuous function of $x \in \mathbb{R}$. Therefore, by Lemma 7.4, for $\{x_k\}_{k\in\mathbb{N}} \subset ROA_f$ we have that
$$\lim_{k\to\infty} W_{\lambda,\beta}(x_k) = 1 - \exp\left(-\lambda \lim_{k\to\infty} V_\beta(x_k)\right) = 1.$$

We next show Statement 2) in Cor. 7.1 by showing Eq. (7.18) holds. If $x \in ROA_f$ then by Lemma 7.4 we have that $V_\beta(x) < \infty$ and thus $e^{-\lambda V_\beta(x)} > 0$ implying $W_{\lambda,\beta}(x) = 1 - e^{-\lambda V_\beta(x)} < 1$. Therefore, $ROA_f \subseteq \{x \in \mathbb{R}^n : W_{\lambda,\beta}(x) < 1\}$. On the other hand if $y \in \{x \in \mathbb{R}^n : W_{\lambda,\beta}(x) < 1\}$ then $a := 1 - W_{\lambda,\beta}(y) > 0$. Thus, $V_\beta(y) = -\frac{1}{\lambda}\ln(a) < \infty$. Lemma 7.4 shows if $V_\beta(y) < \infty$ then $y \in ROA_f$. Hence, $\{x \in \mathbb{R}^n : W_{\lambda,\beta}(x) < 1\} \subseteq ROA_f$. □

7.4.2 A Globally Lipschitz Continuous Lyapunov Function

The function V_β is only defined over the set ROA_f and is unbounded. Such properties make approximating V_β by polynomials challenging. On the other hand

$W_{\lambda,\beta}$ is defined over the whole of $\mathbb{R}^n$ and is bounded by 1. What is more, we next show in Prop. 7.1 that $W_{\lambda,\beta}$ is globally Lipschitz continuous. One may intuit this continuity property by considering the similarity in structure between $W_{\lambda,\beta}$ and the standard mollifier given in Eq. (B.1); a function known to be infinitely differentiable.

Proposition 7.1. *Consider $f \in C^2(\mathbb{R}^n,\mathbb{R})$ and $W_{\lambda,\beta}$ as in Eq. (7.11) where $\lambda > 0$ and $\beta \in \mathbb{N}$. Suppose there exists $\theta,\eta,R > 0$ such that $||D^\alpha f(x)||_2 < \theta$ for all $x \in B_R(0)$ and $||\alpha||_1 \leq 2$, $B_\eta(0)$ is an exponentially stable set (Defn. 7.1) to the ODE (7.1), and $ROA_f \subset B_R(0)$. Then if $\lambda > \theta\eta^{-2\beta}$ and $\beta > \frac{\theta}{2\delta} + \frac{1}{2}$ we have that $W_{\lambda,\beta} \in Lip(\mathbb{R}^n,\mathbb{R})$. Moreover, the Lipschitz constant of $W_{\lambda,\beta}$ is less than or equal to $K > 0$, where*

$$K := 2\lambda \max\left\{\frac{2\beta R^{2\beta-1}}{\theta}, \frac{2\beta(\mu\eta)^{2\beta-1}}{\delta(2\beta-1)-\theta}\right\}. \tag{7.19}$$

Proof. To prove $W_{\lambda,\beta} \in Lip(\mathbb{R}^n,\mathbb{R})$ we will now show

$$|W_{\lambda,\beta}(x) - W_{\lambda,\beta}(y)| < K||x-y||_2 \text{ for all } x,y \in \mathbb{R}^n, \tag{7.20}$$

where $K > 0$ is given in Eq. (7.19).

Case 1: $x, y \in ROA_f$**.** Since $B_\eta(0)$ is an exponentially stable set of the ODE (7.1) and by applying a similar argument in the derivation of Eq. (7.16), it follows that there exists $\delta,\mu > 0$ such that

$$\begin{aligned} ||\phi_f(x,t)||_2 &\leq \mu\eta e^{-\delta(t-F_\eta(x))} \text{ for all } t > F_\eta(x), \qquad (7.21)\\ ||\phi_f(y,t)||_2 &\leq \mu\eta e^{-\delta(t-F_\eta(x))} \text{ for all } t > F_\eta(y). \end{aligned}$$

Without loss of generality we will assume $F_\eta(x) \geq F_\eta(y)$ (otherwise we can relabel x and y).

Now,

$$
\begin{aligned}
&|W_{\lambda,\beta}(x) - W_{\lambda,\beta}(y)| && (7.22)\\
&= \left| \exp\left(-\lambda \int_0^\infty ||\phi_f(x,t)||_2^{2\beta} dt\right) - \exp\left(-\lambda \int_0^\infty ||\phi_f(y,t)||_2^{2\beta} dt\right) \right| \\
&= \exp\left(-\lambda \int_0^\infty ||\phi_f(x,t)||_2^{2\beta} dt\right) \left| 1 - \exp\left(-\lambda \int_0^\infty (||\phi_f(y,t)||_2^{2\beta} - ||\phi_f(x,t)||_2^{2\beta}) dt\right) \right| \\
&\leq \exp\left(-\lambda \int_0^\infty ||\phi_f(x,t)||_2^{2\beta} dt\right) \left| \lambda \int_0^\infty (||\phi_f(y,t)||_2^{2\beta} - ||\phi_f(x,t)||_2^{2\beta}) dt \right| \\
&= \lambda \exp\left(-\lambda V_\beta(x)\right) \left| V_\beta(x) - V_\beta(y) \right|,
\end{aligned}
$$

where the inequality in Eq. (7.22) follows by the exponential inequality given in Eq. (C.3) in Lemma C.1 and the function V_β is as in Eq. (7.10).

We first derive a bound for $|V_\beta(x) - V_\beta(y)|$.

$$
\begin{aligned}
&|V_\beta(x) - V_\beta(y)| = \left| \int_0^\infty ||\phi_f(x,t)||_2^{2\beta} - ||\phi_f(y,t)||_2^{2\beta} dt \right| && (7.23)\\
&\leq \int_0^\infty \Big| ||\phi_f(x,t)||_2 - ||\phi_f(y,t)||_2 \Big| \left(\sum_{i=0}^{2\beta-1} ||\phi_f(x,t)||_2^i ||\phi_f(y,t)||_2^{2\beta-1-i} \right) dt \\
&\leq \int_0^{F_\eta(x)} \Big| ||\phi_f(x,t) - \phi_f(y,t)||_2 \Big| \left(\sum_{i=0}^{2\beta-1} ||\phi_f(x,t)||_2^i ||\phi_f(y,t)||_2^{2\beta-1-i} \right) dt \\
&\qquad + \int_{F_\eta(x)}^\infty \Big| ||\phi_f(x,t) - \phi_f(y,t)||_2 \Big| \left(\sum_{i=0}^{2\beta-1} ||\phi_f(x,t)||_2^i ||\phi_f(y,t)||_2^{2\beta-1-i} \right) dt.
\end{aligned}
$$

We now derive a bound for the two terms that appear in the right hand side of Eq. (7.23). Using the fact $||\phi_f(x,t)||_2 < R$ and $||\phi_f(y,t)||_2 < R$ since $ROA_f \subset B_R(0)$, and using Lemma 7.2, we get,

$$
\begin{aligned}
&\int_0^{F_\eta(x)} ||\phi_f(x,t) - \phi_f(y,t)||_2 \left(\sum_{i=0}^{2\beta-1} ||\phi_f(x,t)||_2^i ||\phi_f(y,t)||_2^{2\beta-1-i} \right) dt && (7.24)\\
&\leq 2\beta R^{2\beta-1} \int_0^{F_\eta(x)} ||\phi_f(x,t) - \phi_f(y,t)||_2 dt \\
&\leq 2\beta R^{2\beta-1} ||x-y||_2 \int_0^{F_\eta(x)} e^{\theta t} dt \\
&= \frac{2\beta R^{2\beta-1}}{\theta} \left(e^{\theta F_\eta(x)} - 1 \right) ||x-y||_2.
\end{aligned}
$$

Moreover, since $\beta > \frac{\theta}{2\delta} + \frac{1}{2}$ it also follows using Eq. (7.21), and Lemma 7.2, that

$$\begin{aligned} &\int_{F_\eta(x)}^{\infty} ||\phi_f(x,t) - \phi_f(y,t)||_2 s \left(\sum_{i=0}^{2\beta-1} ||\phi_f(x,t)||_2^i ||\phi_f(y,t)||_2^{2\beta-1-i} \right) dt \qquad (7.25) \\ &\leq 2\beta(\mu\eta)^{2\beta-1} e^{\delta(2\beta-1)F_\eta(x)} ||x-y||_2 \int_{F_\eta(x)}^{\infty} e^{\theta t - \delta(2\beta-1)t} dt \\ &= \frac{2\beta(\mu\eta)^{2\beta-1}}{\delta(2\beta-1)-\theta} e^{\theta F_\eta(x)} ||x-y||_2. \end{aligned}$$

Now, combining Eqs. (7.23), (7.24) and (7.25) we get,

$$|V_\beta(x) - V_\beta(y)| \max \left\{ \frac{2\beta R^{2\beta-1}}{\theta}, \frac{2\beta(\mu\eta)^{2\beta-1}}{\delta(2\beta-1)-\theta} \right\} e^{\theta F_\eta(x)} ||x-y||_2. \qquad (7.26)$$

We next derive a bound for the $\exp(-\lambda V_\beta(x))$ term in Eq. (7.22).

$$\begin{aligned} \exp(-\lambda V_\beta(x)) &= \exp\left(-\lambda \int_0^\infty ||\phi_f(x,t)||_2^{2\beta} dt \right) \qquad (7.27) \\ &\leq \exp\left(-\lambda \int_0^{F_\eta(x)} ||\phi_f(x,t)||_2^{2\beta} dt \right) \leq e^{-\lambda F_\eta(x) \eta^{2\beta}}. \end{aligned}$$

Finally combining Eqs. (7.22), (7.26), and (7.27), and using the fact $\lambda > \theta\eta^{-2\beta}$, we get

$$\begin{aligned} &|W_{\lambda,\beta}(x) - W_{\lambda,\beta}(y)| \\ &\leq \lambda \max \left\{ \frac{2\beta R^{2\beta-1}}{\theta}, \frac{2\beta(\mu\eta)^{2\beta-1}}{\delta(2\beta-1)-\theta} \right\} e^{-(\lambda\eta^{2\beta}-\theta)F_\eta(x)} ||x-y||_2 \\ &\leq \frac{K}{2} ||x-y||_2, \end{aligned}$$

showing Eq. (7.20) holds when $x, y \in ROA_f$.

Case 2: $x \in ROA_f$ **and** $y \notin ROA_f$**.** Let us consider the set $\{z_\beta\}_{\beta \in [0,1]} \subset \mathbb{R}^n$ where for $\beta \in [0,1]$ we have that $z_\beta := (1-\beta)x + \beta y$. Now, since $x \in ROA_f$ and ROA_f is open (by Lemma 7.1) it follows there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subset ROA_f$. Therefore, since $||z_\beta - x|| = |\beta| ||x-y||_2$, it follows $z_\beta \in ROA_f$ for all $\beta \in [0, \varepsilon/||x-y||_2)$. Thus $\sigma := \sup\{\beta : z_\beta \in ROA_f\} \geq \varepsilon/||x-y||_2 > 0$. Moreover, $\sigma \leq 1$ as $z_1 = y \notin ROA_f$.

Consider $a_n := \sigma(1-1/n)$ and denote the sequence of points $w_n := z_{a_n}$. It follows $\{w_n\}_{n\in\mathbb{N}} \subset ROA_f$ and $w^* := \lim_{n\to\infty} w_n \in \partial ROA_f$. By Lemma 7.4 we have that $\lim_{n\to\infty} V_\beta(w_n) = \infty$. Therefore there exists $N \in \mathbb{N}$ such that

$$\exp(-\lambda V_\beta(w_n)) < \frac{K}{2}||x-y||_2 \text{ for all } n > N. \tag{7.28}$$

Moreover, since $y \notin ROA_f$ we have that $W_{\lambda,\beta}(y) = 1$. Thus by Eq. (7.28) we have that

$$\begin{aligned}|W_{\lambda,\beta}(w_n) - W_{\lambda,\beta}(y)| &= |1 - \exp(-\lambda V_\beta(w_n)) - 1| \\ &= \exp(-\lambda V_\beta(w_n)) \leq \frac{K}{2}||x-y||_2 \text{ for all } n > N.\end{aligned} \tag{7.29}$$

Furthermore, for any $n > N$ we have that $w_n \in ROA_f$ and $x \in ROA_f$ and thus Case 1 shows that

$$|W_{\lambda,\beta}(x) - W_{\lambda,\beta}(w_n)| < \frac{K}{2}||x-w_n||_2. \tag{7.30}$$

Thus, by Eqs. (7.29) and (7.30) and selecting any $n > N$, it now follows that

$$\begin{aligned}&|W_{\lambda,\beta}(x) - W_{\lambda,\beta}(y)| \\ &\qquad = |W_{\lambda,\beta}(x) - W_{\lambda,\beta}(w_n)| + |W_{\lambda,\beta}(w_n) - W_{\lambda,\beta}(y)| \\ &\qquad \leq \frac{K}{2}||x-w_n||_2 + \exp(-\lambda V(w_n)) \\ &\qquad \leq \frac{K}{2}\sigma\left(1-\frac{1}{n}\right)||x-y||_2 + \frac{K}{2}||x-y||_2 \\ &\qquad \leq K||x-y||_2,\end{aligned}$$

where the third inequality follows since $\sigma(1-\frac{1}{n}) < 1$ for all $n \in \mathbb{N}$. Therefore, Eq. (7.20) holds when $x \in ROA_f$ and $y \notin ROA_f$.

Case 3: $y \in ROA_f$ and $x \notin ROA_f$. It follows by a similar argument to Case 2 that

$$|W_{\lambda,\beta}(x) - W_{\lambda,\beta}(y)| \leq K||x-y||_2,$$

and thus Eq. (7.20) holds when $y \in ROA_f$ and $x \notin ROA_f$.

Case 4: $x, y \notin ROA_f$. We have that $W_{\lambda,\beta}(x) = W_{\lambda,\beta}(y) = 1$ for all $x, y \notin ROA_f$ and thus,

$$|W_{\lambda,\beta}(x) - W_{\lambda,\beta}(y)| = 0 \leq K||x - y||_2,$$

and thus Eq. (7.20) holds when $x, y \notin ROA_f$. □

7.4.3 The Converse Lyapunov Function Satisfies a PDE

Proposition 7.1 shows $W_{\lambda,\beta}$ is a Lipschitz continuous function when $\lambda > 0$ and $\beta \in \mathbb{N}$ are sufficiently large. Rademacher's Theorem (Theorem C.4 found in Appendix C) shows that Lipschitz continuous functions are differentiable almost everywhere. Therefore, $W_{\lambda,\beta}$ must satisfy some Partial Differential Equation (PDE) almost everywhere. We next derive this PDE by showing $W_{\lambda,\beta}$ satisfies Eq. (7.31).

Proposition 7.2. *Consider $f \in C^2(\mathbb{R}^n, \mathbb{R})$ and W as in Eq. (7.11). Suppose there exists $\theta, \eta, R > 0$ such that $||D^\alpha f(x)||_2 < \theta$ for all $x \in B_R(x)$ and $||\alpha||_1 \leq 2$, $B_\eta(0)$ is an exponentially stable set (Defn. 7.1) of the ODE (7.1), and $ROA_f \subset B_R(0)$. If $\lambda > \theta\eta^{-2\beta}$ and $\beta > \frac{\theta}{2\delta} + \frac{1}{2}$ then*

$$\nabla W_{\lambda,\beta}(x)^T f(x) = -\lambda||x||_2^{2\beta}(1 - W_{\lambda,\beta}(x)) \text{ for almost every } x \in \mathbb{R}^n. \tag{7.31}$$

Proof. By Prop. 7.1 we have that $W_{\lambda,\beta} \in Lip(\mathbb{R}^n, \mathbb{R})$. Therefore by Rademacher's Theorem (stated in Theorem C.4 and found in Appendix C) $W_{\lambda,\beta}$ is differentiable almost everywhere. Moreover, ϕ_f is differentiable by Lemma 7.3. Since the composition of differentiable functions is itself differentiable it follows by the chain rule that,

$$\begin{aligned} \frac{d}{dt} W_{\lambda,\beta}(\phi_f(x,t))\bigg|_{t=0} &= \nabla W_{\lambda,\beta}(x)^T \frac{\partial}{\partial t}\phi_f(x,t)\bigg|_{t=0} \qquad (7.32) \\ &= \nabla W_{\lambda,\beta}(x)^T f(x) \text{ for almost every } x \in \mathbb{R}^n. \end{aligned}$$

On the other hand, if $x \in ROA_f$ it follows $\phi_f(x,t) \in ROA_f$ for all $t \geq 0$ and thus,

$$W_{\lambda,\beta}(\phi_f(x,t)) = 1 - \exp\left(-\lambda \int_t^\infty ||\phi_f(x,s)||_2^{2\beta} ds\right). \tag{7.33}$$

By the fundamental theorem of calculus and the fact $\phi_f(x,0) = x$ for all $x \in \mathbb{R}^n$ we have that,

$$\begin{aligned} &\frac{d}{dt} W_{\lambda,\beta}(\phi_f(x,t))\bigg|_{t=0} \\ &= -\lambda ||\phi_f(x,t)||_2^{2\beta} \exp\left(-\lambda \int_t^\infty ||\phi_f(x,s)||_2^{2\beta} ds\right)\bigg|_{t=0} \\ &= -\lambda ||x||_2^{2\beta}(1 - W_{\lambda,\beta}(x)) \text{ for } x \in ROA_f. \end{aligned} \tag{7.34}$$

If $x \notin ROA_f$ then clearly $\phi_f(x,t) \notin ROA_f$ for all $t \geq 0$. Thus $W(\phi_f(x,t)) = 1$ for all $x \notin ROA_f$ and $t \geq 0$. Therefore,

$$\begin{aligned} \frac{d}{dt} W_{\lambda,\beta}(\phi_f(x,t))\bigg|_{t=0} &= \frac{d}{dt} 1\bigg|_{t=0} = 0 = -\lambda ||x||_2^{2\beta}(1-1) \\ &= -\lambda ||x||_2^{2\beta}(1 - W_{\lambda,\beta}(x)) \text{ for } x \notin ROA_f. \end{aligned} \tag{7.35}$$

Hence, Eqs. (7.32), (7.34) and (7.35) prove that the PDE given in Eq. (7.31) holds. □

7.5 A Convex Optimization Problem for Approximating the Converse Lyapunov Function

We have reduced the problem of approximating the region of attraction to solving the optimization problem given in Eq. (7.8), where $W = W_{\lambda,\beta}$ is given in Eq. (7.11). Unfortunately, no analytical formula for $W_{\lambda,\beta}$ is known. Therefore, the optimization problem given in Eq. (7.8) cannot be solved in its current form.

Fortunately, the unknown function $W_{\lambda,\beta}$ can be removed from the objective function of Opt. (7.8). To see this note that if $J(x) \geq W_{\lambda,\beta}(x)$ for all $x \in \Lambda \subseteq \Omega$, then minimizing $\int_\Lambda |J(x) - W_{\lambda,\beta}(x)| dx$ is equivalent to minimizing $\int_\Lambda J(x) dx$. Thus, Opt. (7.8) is equivalent to the following optimization problem,

$$P_d \in \arg \min_{J \in \mathcal{P}_d(\mathbb{R}^n, \mathbb{R})} \int_\Lambda J(x) dx \tag{7.36}$$
$$\text{such that } J(x) \geq W_{\lambda,\beta}(x) \text{ for all } x \in \Omega.$$

Unfortunately, the constraint of Opt. (7.36) still involves the unknown function $W_{\lambda,\beta}$. In the absence of an analytical formula for $W_{\lambda,\beta}$ we propose in Prop. 7.3 conditions, in the form of the linear partial differential inequalities given in Eqs. (7.37), (7.38) and (7.39), that when satisfied by some function $J \in C^1(\mathbb{R}^n, \mathbb{R})$ implies that $W_{\lambda,\beta}(x) \leq J(x)$. Thus, any J satisfying Eqs. (7.37), (7.38) and (7.39) is feasible to Opt. (7.36).

7.5.1 Bounding The Converse Lyapunov Function From Above

Proposition 7.3. *Consider $f \in C^1(\mathbb{R}^n, \mathbb{R})$, $\beta \in \mathbb{N}$ and $\lambda > 0$. Suppose there exists $J \in C^1(\Omega, \mathbb{R})$ that satisfies*

$$\nabla J(x)^T f(x) \leq -\lambda ||x||_2^{2\beta}(1 - J(x)) \text{ for all } x \in \Omega, \tag{7.37}$$
$$J(x) \geq 1 \text{ for all } x \in \partial\Omega, \tag{7.38}$$
$$J(0) \geq 0, \tag{7.39}$$

where $\Omega \subset \mathbb{R}^n$ is a compact set. Then $W_{\lambda,\beta}(x) \leq J(x)$ for all $x \in \Omega$, where $W_{\lambda,\beta}$ is as in Eq. (7.11).

Proof. Consider $x \in \Omega$. Let us consider the time the solution map exits the set $\Omega \subset \mathbb{R}^n$, denoted by $T_x := \sup\{t \geq 0 : \phi_f(x,t) \in \Omega\}$. Furthermore, let us denote $u(t) := J(\phi_f(x,t)) - 1$ and $\alpha(t) := \lambda||\phi_f(x,t)||_2^{2\beta}$. It follows from Eq. (7.37) that

$$\frac{d}{dt}u(t) \leq \alpha(t)u(t) \text{ for all } t \in [0, T_x].$$

Therefore by Lemma C.2 it follows that

$$u(t) \leq u(0) \exp\left(\int_0^t \alpha(s) ds\right) \text{ for all } t \in [0, T_x],$$

and thus selecting $t = T_x$ we have that

$$J(\phi_f(x,T_x)) - 1 \leq (J(x) - 1) \exp\left(\lambda \int_0^{T_x} ||\phi_f(x,s)||_2^{2\beta} ds\right). \tag{7.40}$$

By rearranging Eq. (7.40) we get that,

$$J(x) \geq 1 - (1 - J(\phi_f(x,T_x))) \exp\left(-\lambda \int_0^{T_x} ||\phi_f(x,s)||_2^{2\beta} ds\right). \tag{7.41}$$

Case 1: $T_x < \infty$. In this case the solution map exits the set Ω in some finite time. Since $\phi_f \in C(\mathbb{R}^n \times [0,\infty), \mathbb{R}^n)$ (by Lemma 7.3) it is clear that $\phi_f(x,T_x) \in \partial\Omega$. Therefore by Eq. (7.38) we have that $J(\phi_f(x,T_x)) \geq 1$.
Hence, $(1 - J(\phi_f(x,T_x))) \exp\left(-\lambda \int_0^{T_x} ||\phi_f(x,s)||_2^{2\beta} ds\right) \leq 0$. Thus, by Eq. (7.41) we have that,

$$\begin{aligned} J(x) &\geq 1 - (1 - J(\phi_f(x,T_x))) \exp\left(-\lambda \int_0^{T_x} ||\phi_f(x,s)||_2^{2\beta} ds\right) \\ &\geq 1 \geq W_{\lambda,\beta}(x), \end{aligned}$$

since $W_{\lambda,\beta}(x) \leq 1$.

Case 2a: $T_x = \infty$ and $x \in ROA_f$. In this case we have $\lim_{t\to\infty} ||\phi_f(x,t)||_2 = ||\phi_f(x,T_x)||_2 = 0$ since $x \in ROA_f$. Moreover, since $J(\phi_f(x,T_x)) = J(0) \geq 0$ (by Eq. (7.39)) and $\exp(x) \geq 0$ for all $x \in \mathbb{R}$ it follows from Eq. (7.41) that

$$\begin{aligned} J(x) &\geq 1 - (1 - J(0)) \exp\left(-\lambda \int_0^{\infty} ||\phi_f(x,s)||_2^{2\beta} ds\right) \\ &\geq 1 - \exp\left(-\lambda \int_0^{\infty} ||\phi_f(x,s)||_2^{2\beta} ds\right) = W_{\lambda,\beta}(x). \end{aligned}$$

Case 2b: $T_x = \infty$ and $x \in \Omega/ROA_f$. If $x \in \Omega/ROA_f$ we have that $W(x) = 1$. Moreover, if $T_x = \infty$ then the solution map never exits the set Ω, that is $\phi_f(x,t) \in \Omega$ for all $t \geq 0$. Since J is differentiable and Ω is compact we have that J is bounded, that is, there exists $M > 0$ such that $|J(\phi_f(x,t))| < M$ for all $t \geq 0$. Since $x \notin ROA_f$ we have that $\phi_f(x,t) \notin ROA_f$ for all $t \geq 0$. This there exists $\varepsilon > 0$ such that

$||\phi_f(x,t)||_2^{2\beta} \geq \varepsilon^{2\beta}$ for all $t \geq 0$. Thus, since $|J(\phi_f(x,t))| < M$ for all $t \geq 0$, we have that

$$\begin{aligned}
&\left|(1 - J(\phi_f(x,T_x))) \exp\left(-\lambda \int_0^{T_x} ||\phi_f(x,s)||_2^{2\beta} ds\right)\right| \\
&= \lim_{T\to\infty} \left|\left(1 - J(\phi_f(x,T))\right) \exp\left(-\lambda \int_0^{T} ||\phi_f(x,s)||_2^{2\beta} ds\right)\right| \\
&= \lim_{T\to\infty} |1 - J(\phi_f(x,T))| \exp\left(-\lambda \int_0^{T} ||\phi_f(x,s)||_2^{2\beta} ds\right) \\
&\leq \lim_{T\to\infty} \left\{(M+1) \exp\left(-T\lambda\varepsilon^{2\beta}\right)\right\} = 0,
\end{aligned}$$

implying $(1 - J(\phi_f(x,T_x))) \exp\left(-\lambda \int_0^{T_x} ||\phi_f(x,s)||_2^{2\beta} ds\right) = 0$.

It is now clear by Eq. (7.41) that

$$\begin{aligned}
J(x) &\geq 1 - (1 - J(\phi_f(x,T_x))) \exp\left(-\lambda \int_0^{T_x} ||\phi_f(x,s)||_2^{2\beta} ds\right) \\
&\geq 1 = W_{\lambda,\beta}(x).
\end{aligned}$$

□

Corollary 7.2. *Consider $f \in C^1(\mathbb{R}^n, \mathbb{R})$, $\beta \in \mathbb{N}$ and $\lambda > 0$. Suppose there exists $J \in C^1(\Omega, \mathbb{R})$ that satisfies Eqs.* (7.37), (7.38) *and* (7.39) *for some compact set Ω. Then $J(x) \geq 0$ for all $x \in \Omega$.*

Proof. By Prop. 7.3 we have that $J(x) \geq W_{\lambda,\beta}(x) \geq 0$, where $W_{\lambda,\beta}$ is as in Eq. (7.11).

□

7.5.2 Tightening The Problem of Approximating Our Proposed Converse Lyapunov Function

Using Prop. 7.3 we now tighten the optimization problem given in Eq. (7.36). For given $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$, $\lambda > 0$, $\beta \in \mathbb{N}$, $R > 0$ and $\Lambda \subseteq \Omega \subset \mathbb{R}^n$ consider the following optimization problem,

$$P_d \in \arg \min_{J \in \mathcal{P}_d(\mathbb{R}^n, \mathbb{R})} \int_\Lambda J(x) dx \tag{7.42}$$

$$\text{such that } J \text{ satisfies } (7.37), (7.38), \text{ and } (7.39).$$

Clearly the Opt. (7.42) is a tightening of the Opt. (7.8) since if J is feasible to Opt. (7.42) then by Prop. 7.3 we have that J is also feasible to Opt. (7.8). Moreover, Opt. (7.42) is a convex optimization problem since it is linear in its decision variable, J, in both the constraints and objective function. In the next section we further tighten Opt. (7.42) to an SOS Optimization Problem (7.69) that can be tractably solved. For implementation purposes we select $\Omega = B_R(0)$, where $R > 0$, and $\Lambda \subseteq \Omega$ as some rectangular set (of form $[a_1, b_1] \times ... \times [a_1, b_2] \subset \mathbb{R}^n$).

7.6 Our Proposed Converse Lyapunov Function can be Approximated Arbitrarily well by a Polynomial Function

In the previous section we have proposed an optimization problem, given in Eq. (7.42), for approximating our proposed converse Lyapunov function (given in Eq. (7.11)). In this section we show, later in Theorem 7.1, that there exists a polynomial function arbitrarily "close" to the converse Lyapunov function $W_{\lambda,\beta}$ and also a feasible solution to some $d \in \mathbb{N}$ instantiation of the family of optimization problems given in Eq. (7.42). Later in Section 7.7 we will tighten the family of d-degree optimization problems given in Eq. (7.42) to a family of d-degree SOS optimization problem. We will prove that the sequence of solutions to the d-degree SOS optimization problem converges locally in the L^1-norm to our proposed converse Lyapunov function. Theorem 7.1, stated and proved in this chapter, is a key component to the convergence proof of our d-degree SOS optimization problem. In order to prove Theorem 7.1 we take the following steps:

(A) In Lemma 7.5 we take the mollification of $W_{\lambda,\beta}$ to show there exists an infinitely differentiable function that satisfies Eqs. (7.43), (7.44) and (7.45).

(B) In Prop. 7.4 we use Lemma 7.5 together with partitions of unity (Theorem C.6) to show there exists an infinitely differentiable function that satisfies Eqs. (7.48), (7.49) and (7.50).

(C) In Theorem 7.1 we use Prop. 7.4 together with the polynomial approximation results in Theorem C.3 to show there exists a polynomial function that satisfies Eqs. (7.60), (7.61) and (7.62).

Lemma 7.5. *Consider $f \in C^2(\mathbb{R}^n, \mathbb{R})$ and W as in Eq.* (7.11)*. Suppose there exists $\theta, \eta, R > 0$ such that $||D^\alpha f(x)||_2 < \theta$ for all $x \in B_R(x)$ and $||\alpha||_1 \leq 2$, $B_\eta(0)$ is an exponentially stable set (Defn. 7.1) of the ODE* (7.1)*, and $ROA_f \subset B_R(0)$. If $\lambda > \theta \eta^{-2\beta}$ and $\beta > \frac{\theta}{2\delta} + \frac{1}{2}$ then for any $\varepsilon > 0$ and $R_1 > R$ there exists $J \in C^\infty(B_{R_1}(0), \mathbb{R})$ such that*

$$\sup_{x \in B_{R_1}(0)} |J(x) - W_{\lambda,\beta}(x)| < \varepsilon, \tag{7.43}$$

$$\nabla J(x)^T f(x) < -\lambda(1 - J(x))||x||_2^{2\beta} + \varepsilon \text{ for all } x \in B_{R_1}(0), \tag{7.44}$$

$$J(x) = 1 \text{ for all } x \in \partial B_R(0) \text{ and } J(0) \geq 0. \tag{7.45}$$

Proof. Let $\varepsilon > 0$ and $R_2 > R_1 > R$. Since $W_{\lambda,\beta} \in Lip(\mathbb{R}^n, \mathbb{R})$ (by Prop. 7.1) we know that by Theorem C.4 that $W_{\lambda,\beta} \in W^{1,\infty}(\mathbb{R}^n, \mathbb{R})$.

For $\sigma > 0$ let us denote the σ-mollification of $W_{\lambda,\beta}$ by $J_\sigma(x) := [W_{\lambda,\beta}]_\sigma(x)$. We note that the domain of $W_{\lambda,\beta}$ is $\mathbb{R}^n$. However, for mollification purposes we consider $W_{\lambda,\beta}$ over the restricted domain $B_{R_2}(0) \subset \mathbb{R}^n$.

Let $\sigma_1 := \frac{R_2 - R_1}{2}$. It is clear that $B_{R_1}(0) \subset < B_{R_2}(0) >_\sigma$ for all $0 < \sigma < \sigma_1$. Therefore, by Prop. B.1 we have that $J_\sigma \in C^\infty(< B_{R_2}(0) >_\sigma, \mathbb{R}) \subset C^\infty(B_{R_1}(0), \mathbb{R})$ for all $0 < \sigma < \sigma_1$.

We will now show there exists $\sigma > 0$ such that Eqs. (7.43), (7.44) and (7.45) hold.

First we show Eq. (7.43) holds. By Prop. B.1 we know that there exists $\sigma_2 > 0$ such that for all $0 < \sigma < \sigma_2$ we have that

$$\sup_{x \in B_{R_1}(0)} |J_\sigma(x) - W_{\lambda,\beta}(x)| < \varepsilon.$$

We now show Eq. (7.44) holds. Let us denote $r(x) := ||x||_2^{2\beta}$. It is clear using the triangle inequality and the fact that $||x - z|| < 2R_1$ for all $x, z \in B_{R_1}(0)$ that

$$\begin{aligned} r(x) - r(x-z) &= (||x||_2 - ||x-z||_2) \sum_{k=0}^{2\beta-1} ||x||_2^{2\beta-1-k} ||x-z||_2^k \\ &\leq \left({R_1}^{2\beta-1} \sum_{k=0}^{2\beta-1} 2^k \right) ||z||_2 \text{ for all } x, z \in B_{R_1}(0). \end{aligned} \tag{7.46}$$

Let $\sigma_3 := \frac{\varepsilon}{KL_f + \lambda\left({R_1}^{2\beta-1} \sum_{k=0}^{2\beta-1} 2^k\right)}$ where K (given in Eq. (7.19)) is the Lipschitz constants of $W_{\lambda,\beta}$ and L_f is the Lipschitz constants of f. For $0 < \sigma < \sigma_3$, using Prop. B.1

and the fact $W_{\lambda,\beta}$ satisfies Eq. (7.31), we have that

$$\begin{aligned}
&\nabla J_\sigma(x)^T f(x) + \lambda(1 - J_\sigma(x))||x||_2^{2\beta} \qquad (7.47)\\
&= \nabla[W_{\lambda,\beta}]_\sigma(x)^T f(x) + \lambda(1 - [W_{\lambda,\beta}]_\sigma(x))r(x)\\
&= [\nabla W_{\lambda,\beta}]_\sigma(x)^T f(x) + \lambda(1 - [W_{\lambda,\beta}]_\sigma(x))r(x)\\
&= ([\nabla W_{\lambda,\beta}^T f]_\sigma(x) + \lambda[r]_\sigma(x) - \lambda[W_{\lambda,\beta} r]_\sigma(x)) + [\nabla W_{\lambda,\beta}]_\sigma(x)^T f(x) - [\nabla W_{\lambda,\beta}^T f]_\sigma(x)\\
&\qquad + \lambda r(x) - \lambda[r]_\sigma(x) + \lambda[W_{\lambda,\beta} r]_\sigma(x) - \lambda[W_{\lambda,\beta}]_\sigma(x) r(x)\\
&= [\nabla W_{\lambda,\beta}^T f + \lambda(1 - W_{\lambda,\beta})r]_\sigma(x) + [\nabla W_{\lambda,\beta}]_\sigma(x)^T f(x) - [\nabla W_{\lambda,\beta}^T f]_\sigma(x)\\
&\qquad + \lambda(1 - [W_{\lambda,\beta}]_\sigma) r(x) - \lambda[(1 - W_{\lambda,\beta})r]_\sigma(x)\\
&= \int_{B_\sigma(0)} \eta_\sigma(z) \nabla W_{\lambda,\beta}(x-z)^T (f(x) - f(x-z))dz\\
&\qquad + \lambda \int_{B_\sigma(0)} \eta_\sigma(z)(1 - W_{\lambda,\beta}(x-z))(r(x) - r(x-z))dz\\
&\leq \operatorname*{ess\,sup}_{x\in\mathbb{R}^n}\{||\nabla W_{\lambda,\beta}(x)||_2\} \int_{B_\sigma(0)} \eta_\sigma(z)||f(x) - f(x-z)||_2 dz\\
&\qquad + \lambda \int_{B_\sigma(0)} \eta_\sigma(z)|r(x) - r(x-z)|dz\\
&\leq \left(KL_f + \lambda {R_1}^{2\beta-1} \sum_{k=0}^{2\beta-1} 2^k \right) \int_{B_\sigma(0)} \eta_\sigma(z)||z||_2 dz\\
&\leq \left(KL_f + \lambda {R_1}^{2\beta-1} \sum_{k=0}^{2\beta-1} 2^k \right) \sigma < \varepsilon \text{ for all } x \in B_{R_1}(0).
\end{aligned}$$

Where the first inequality in Eq. (7.47) follows by the Cauchy Swartz inequality and the second inequality follows by the fact $\operatorname{ess\,sup}_{x\in\mathbb{R}^n}\{||\nabla W_{\lambda,\beta}(x)||_2\} \leq K$ (By Rademacher's theorem stated in Theorem C.4) and Eq. (7.46).

We now show Eq. (7.45) holds. Since $ROA_f \subset B_R(0)$ and ROA_f is an open set (by Lemma 7.1) it follows that for all $x \in \partial B_R(0)$ we have $x \notin ROA_f$ and thus $W_{\lambda,\beta}(x) = 1$ for all $x \in \partial B_R(0)$. Now, there exists a sufficiently small $\sigma_4 > 0$ such that $B_{\sigma_4}(x) \cap ROA_f = \emptyset$ for all $x \in \partial B_R(0)$. Thus for $0 < \sigma < \sigma_4$

$$J_\sigma(x) = \int_{B_\sigma(0)} \eta_\sigma(z) W(x-z) dz = \int_{B_\sigma(0)} \eta_\sigma(z) dz = 1,$$

for all $x \in \partial B_R(0)$.

Moreover, $\eta_\sigma(x) \geq 0$ and $W_{\lambda,\beta}(x) \geq 0$ for all $\sigma > 0$ and $x \in \mathbb{R}^n$ so therefore $J_\sigma(x) \geq 0$ for all $\sigma > 0$ and $x \in \mathbb{R}^n$. Thus $J_\sigma(0) \geq 0$ for all $\sigma > 0$.

In conclusion for $\sigma < \min\{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$ we have that J_σ satisfies Eqs. (7.43), (7.44) and (7.45). □

Proposition 7.4. *Consider $f \in C^2(\mathbb{R}^n, \mathbb{R})$ and W as in Eq.* (7.11). *Suppose there exists $\theta, \eta, R > 0$ such that $||D^\alpha f(x)||_2 < \theta$ for all $x \in B_R(x)$ and $||\alpha||_1 \leq 2$, $B_\eta(0)$ is an exponentially stable set (Defn. 7.1) of the ODE* (7.1), *and $ROA_f \subset B_R(0)$. If $\lambda > \theta \eta^{-2\beta}$ and $\beta > \frac{\theta}{2\delta} + \frac{1}{2}$ then for any $\varepsilon > 0$ and $R_1 > R$ there exists $J \in C^\infty(B_{R_1}(0), \mathbb{R})$ such that*

$$\sup_{x \in B_{R_1}(0)} |J(x) - W_{\lambda,\beta}(x)| < \varepsilon, \tag{7.48}$$

$$\nabla J(x)^T f(x) \leq -\lambda(1 - J(x))||x||_2^{2\beta} + \varepsilon ||x||_2^{2\beta} \text{ for } x \in B_{R_1}(0), \tag{7.49}$$

$$J(x) = 1 \text{ for all } x \in \partial B_R(0) \text{ and } J(0) = 0. \tag{7.50}$$

Proof. Consider the sets $U_m = B_{R_1}(0)/(B_{1/m}(0))^{cl}$ for $m \in \mathbb{N}$. It is clear $\{U_m\}_{m \in \mathbb{N}}$ form an open cover (Defn. C.1) of $B_{R_1}(0)/\{0\}$, that is $\cup_{m \in \mathbb{N}} U_m = B_{R_1}(0)/\{0\}$. By Theorem C.6 (found in Appendix C) there exists a partition of unity, we denote by $\{\psi_m\}_{m \in \mathbb{N}} \subset C^\infty(\mathbb{R}^n, \mathbb{R})$, subordinate to the open cover $\{U_m\}_{m \in \mathbb{N}}$.

Let $\varepsilon > 0$. For each $m \in \mathbb{N}$ it was shown in Lemma 7.5 that there exists $J_m \in$

$C^\infty(B_{R_1}(0),\mathbb{R})$ such that

$$\sup_{x\in B_{R_1}(0)} |J_m(x) - W_{\lambda,\beta}(x)| \tag{7.51}$$

$$< \frac{\varepsilon}{2^{m+1}(\sup_{x\in U_m}\{|\nabla\psi_m(x)^T f(x)|\}+1)m^{2\beta}},$$

$$\nabla J_m(x)^T f(x) < -\lambda(1-J_m(x))||x||_2^{2\beta} + \frac{\varepsilon}{2m^{2\beta}}$$

$$\text{for all } x\in B_{R_1}(0), \tag{7.52}$$

$$J_m(x) = 1 \text{ for all } x\in\partial B_R(0) \text{ and } J_m(0)\geq 0. \tag{7.53}$$

Note, $\sup_{x\in U_m}\{|\nabla\psi_m(x)^T f(x)|\} < \infty$ for each $m\in\mathbb{N}$ since U_m is bounded and the function $\psi_m(x)^T f(x)$ is continuous in x.

We now consider the function $J(x) := \sum_{m=1}^\infty \psi_m(x)J_m(x)$. We first note that $J\in C^\infty(B_{R_1}(0),\mathbb{R})$. This is due to the fact that $J_m\in C^\infty(B_{R_1}(0),\mathbb{R})$ and $\psi_m\in C^\infty(\mathbb{R}^n,\mathbb{R})$ for all $m\in\mathbb{N}$. Moreover, for any $x\in\mathbb{R}^n$ Theorem C.6 (found in Section C) shows that there is an open set $S\subset\mathbb{R}^n$ containing $x\in\mathbb{R}^n$ such that only finitely many ψ_m's are non-zero over S. Thus J is a finite sum of $C^\infty(B_{R_1}(0),\mathbb{R})$ functions over S and thus differentiable at x. Since $x\in\mathbb{R}^n$ was arbitrarily chosen it follows $J\in C^\infty(B_{R_1}(0),\mathbb{R})$.

We now show J satisfies Eq. (7.48). Using the fact $\sum_{m=1}^\infty \psi_m(x) = 1$ for all $x\in B_{R_1}(0)/\{0\}$ and $\sum_{m=1}^\infty \psi_m(0)=0$ together with Eq. (7.51) we have that

$$|J(x) - W_{\lambda,\beta}(x)| = \left|\sum_{m=1}^\infty \psi_m(x)J_m(x) - W_{\lambda,\beta}(x)\right|$$

$$\leq \sum_{m=1}^\infty \psi_m(x)|J_m(x) - W_{\lambda,\beta}(x)| \leq \sum_{m=1}^\infty \frac{\psi_m(x)\varepsilon}{2} < \varepsilon \text{ for } x\in B_{R_1}(0).$$

We now show J satisfies Eq. (7.49). Before doing so we note that $\sum_{m=1}^\infty \psi_m(x) = 1$ for all $x\in B_{R_1}(0)/\{0\}$. Since only finitely many ψ_m's are non-zero for each $x\in B_{R_1}(0)/\{0\}$ it follows $\sum_{m=1}^\infty \psi_m(x)$ is a finite sum of infinitely differentiable functions. Therefore, we can interchange the derivative and the summation to show $0 = \frac{\partial}{\partial x_i}1 =$

$\frac{\partial}{\partial x_i} \sum_{m=1}^\infty \psi_m(x) = \sum_{m=1}^\infty \frac{\partial}{\partial x_i} \psi_m(x)$ for all $x \in B_{R_1}(0)/\{0\}$ and $i \in \{1, ..., n\}$. Thus it follows $\sum_{m=1}^\infty \nabla \psi_m(x) = [0, ..., 0]^T \in \mathbb{R}^n$ for all $x \in B_{R_1}(0)/\{0\}$. Hence,

$$W_{\lambda\beta}(x) \sum_{m=1}^\infty \nabla \psi_m(x)^T f(x) = 0 \text{ for all } x \in B_{R_1}(0)/\{0\}. \tag{7.54}$$

For $x \in B_{R_1}(0)/\{0\}$ let us denote $I_x := \{m \in \mathbb{N} : x \in U_m\}$. Note, $\{U_m\}_{m\in\mathbb{N}}$ forms an open cover for $B_{R_1}(0)/\{0\}$ so $I_x \neq \emptyset$ for all $x \in B_{R_1}(0)/\{0\}$.

It is clear that for $x \in B_{R_1}(0)/\{0\}$ and $m \in I_x$ that $x \in U_m = B_{R_1}(0)/B_{\frac{1}{m}}(0)$ and so $||x||_2 \geq \frac{1}{m}$ implying $\frac{1}{m^{2\beta}} \leq ||x||_2^{2\beta}$. Therefore,

$$\sup_{m \in I_x} \left\{ \frac{1}{m^{2\beta}} \right\} \leq ||x||_2^{2\beta} \text{ for all } x \in B_{R_1}(0)/\{0\}. \tag{7.55}$$

Moreover, for $x \in B_{R_1}(0)/\{0\}$ and $m \notin I_x$ we have that $x \notin U_m$. Thus, since $\{x \in \mathbb{R}^n : \psi_m(x) \neq 0\} \subset U_m$ (by Theorem C.6 found in Appendix C) we have that

$$\psi_m(x) = 0 \text{ for all } x \in B_{R_1}(0)/\{0\} \text{ and } m \notin I_x. \tag{7.56}$$

Now, using Eqs. (7.51), (7.52), (7.54), (7.55) and (7.56), and the fact $\sum_{m=1}^\infty \frac{1}{2^m} = 1$

we have that,

$$\begin{aligned}
&\nabla J(x)^T f(x) + \lambda(1-J(x))||x||_2^{2\beta} \qquad (7.57)\\
&= \sum_{m=1}^{\infty} \psi_m(x)\left(\nabla J_m(x)^T f(x) + \lambda(1-J_m(x))||x||_2^{2\beta}\right)\\
&\qquad\qquad + \sum_{m=1}^{\infty} J_m(x)\nabla\psi_m(x)^T f(x) - W_{\lambda\beta}(x)\sum_{m=1}^{\infty}\nabla\psi_m(x)^T f(x)\\
&= \sum_{m\in I_x} \psi_m(x)\left(\nabla J_m(x)^T f(x) + \lambda(1-J_m(x))||x||_2^{2\beta}\right)\\
&\qquad\qquad + \sum_{m\in I_x}(J_m(x) - W_{\lambda\beta}(x))\nabla\psi_m(x)^T f(x)\\
&\le \sum_{m\in I_x}\psi_m(x)\frac{\varepsilon}{2m^{2\beta}} + \sum_{m\in I_x}\frac{\varepsilon}{2^{m+1}m^{2\beta}}\\
&\le \varepsilon \sup_{m\in I_x}\left\{\frac{1}{m^{2\beta}}\right\}\left(\frac{1}{2}\sum_{m\in I_x}\psi_m(x) + \frac{1}{2}\sum_{m\in I_x}\frac{1}{2^m}\right)\\
&\le \varepsilon \sup_{m\in I_x}\left\{\frac{1}{m^{2\beta}}\right\} \le \varepsilon||x||_2^{2\beta} \text{ for all } x\in B_{R_1}(0)/\{0\}.
\end{aligned}$$

Eq. (7.57) shows J satisfies Eq. (7.49) for $x \in B_{R_1}(0)/\{0\}$. We still need to show J satisfies Eq. (7.49) for $x=0$. Let us denote the function $F(x) := \nabla J(x)^T f(x) + \lambda(1-J(x))||x||_2^{2\beta}$. To show J satisfies Eq. (7.49) at $x=0$ we must show $F(0)\le 0$. We first note that $F \in C^2(B_{R_1}(0),\mathbb{R})$ since $J\in C^\infty(B_{R_1}(0),\mathbb{R})$, $f\in C^2(\mathbb{R}^n,\mathbb{R})$ and $||x||_2^{2\beta}\in C^2(\mathbb{R}^n,\mathbb{R})$. Thus $F\in LocLip(\mathbb{R}^n,\mathbb{R})$. Therefore,

$$|F(0)-F(x)| \le L_F||x||_2 \text{ for all } x\in B_{R_1}(0), \qquad (7.58)$$

where L_F is the Lipschitz constant of F. Then, Eq. (7.57) together with Eq. (7.58) implies that

$$F(0) \le L_F||x||_2 + F(x) \le L_F||x||_2 + \varepsilon||x||_2^{2\beta} \text{ for all } x\in B_{R_1}(0)/\{0\}. \qquad (7.59)$$

Now, for contradiction suppose the negation of $F(0)\le 0$, that is there exists $a>0$ such that $F(0)\ge a$. Considering $x = \min\{\frac{a}{3(L_F+1)\sqrt{n}}, \frac{1}{\sqrt{n}}(\frac{a}{3\varepsilon})^{1/\beta}, \frac{R_1}{\sqrt{n}}\}[1,...,1]^T \in$

$B_{R_1}(0)/\{0\} \subset \mathbb{R}^n$ and using Eq. (7.59) we have that

$$a \leq F(0) \leq \frac{2}{3}a,$$

providing a contradiction. Therefore, $F(0) \leq 0$ and so J satisfies Eq. (7.49) for all $x \in B_{R_1}(0)$.

We now show J satisfies Eq. (7.50). Let $x \in \partial B_R(0)$. By Eq. (7.53) we have that $J_m(x) = 1$ for all $m \in \mathbb{N}$. Therefore, using the fact $\sum_{m=1}^{\infty} \psi_m(x) = 1$ for all $x \in B_{R_1}(0)/\{0\}$ and $\partial B_R(0) \subset B_{R_1}(0)/\{0\}$ since $R_1 > R$, we have that

$$J(x) = \sum_{m=1}^{\infty} \psi_m(x) J_m(x) = \sum_{m=1}^{\infty} \psi_m(x) = 1.$$

Moreover, $0 \notin B_{R_1}(0)/\{0\}$ so $\psi_m(0) = 0$ for all $m \in \mathbb{N}$. Hence,
$J(0) = \sum_{m=1}^{\infty} \psi_m(x) J_m(x) = 0.$ □

Theorem 7.1. *Consider $f \in C^2(\mathbb{R}^n, \mathbb{R})$ and W as in Eq. (7.11). Suppose there exists $\theta, \eta, R > 0$ such that $||D^\alpha f(x)||_2 < \theta$ for all $x \in B_R(x)$ and $||\alpha||_1 \leq 2$, $B_\eta(0)$ is an exponentially stable set (Defn. 7.1) of the ODE (7.1), and $ROA_f \subset B_R(0)$. If $\lambda > \theta \eta^{-2\beta}$ and $\beta > \frac{\theta}{2\delta} + \frac{1}{2}$ then for any $\varepsilon > 0$ there exists $P \in \mathcal{P}(\mathbb{R}^n, \mathbb{R})$ such that*

$$\sup_{x \in B_R(0)} |P(x) - W_{\lambda,\beta}(x)| < \varepsilon, \tag{7.60}$$

$$\nabla P(x)^T f(x) < -\lambda(1 - P(x))||x||_2^{2\beta} \text{ for all } x \in B_R(0), \tag{7.61}$$

$$P(x) > 1 \text{ for all } x \in \partial B_R(0) \text{ and } P(0) > 0. \tag{7.62}$$

Proof. Let $\varepsilon > 0$ and $R_1 > R$. By Prop. 7.4 there exists $J \in C^\infty(B_{R_1}(0), \mathbb{R})$ that satisfies

$$\sup_{x \in B_{R_1}(0)} |J(x) - W_{\lambda,\beta}(x)| < \frac{\varepsilon}{a}, \tag{7.63}$$

$$\nabla J(x)^T f(x) \leq -\lambda(1 - J(x))||x||_2^{2\beta} + \frac{\varepsilon}{a}||x||_2^{2\beta} \text{ for } x \in B_{R_1}(0), \tag{7.64}$$

$$J(x) = 1 \text{ for all } x \in \partial B_R(0) \text{ and } J(0) = 0, \tag{7.65}$$

where

$$a := \max\left\{3, \frac{\sup_{x \in B_R(0)} ||f(x)||_2}{\lambda R} + R^{-2\beta} + \frac{1}{\lambda} + 2\right\}. \tag{7.66}$$

Now, Theorem C.3, found in Appendix C, shows there exists $\tilde{P} \in \mathcal{P}(\mathbb{R}^n, \mathbb{R})$ such that

$$|J(x) - \tilde{P}(x)| < \frac{\varepsilon}{aR^{2\beta}} ||x||_2^{2\beta} \text{ for all } x \in (B_R(0))^{cl}, \tag{7.67}$$

$$||\nabla J(x) - \nabla \tilde{P}(x)||_2 < \frac{\varepsilon}{aR^{2\beta}} ||x||_2^{2\beta} \text{ for all } x \in (B_R(0))^{cl}. \tag{7.68}$$

Let $P(x) := \tilde{P}(x) + \frac{a-2}{a}\varepsilon \in \mathcal{P}(\mathbb{R}^n, \mathbb{R})$. We will now show P satisfies Eqs. (7.60), (7.61) and (7.62).

We first show Eq. (7.60) holds. Using the triangle inequality along with Eqs. (7.63) and (7.67) we have that

$$\begin{aligned}
|P(x) - W_{\lambda,\beta}(x)| &\leq |\tilde{P}(x) - W_{\lambda,\beta}(x)| + \frac{a-2}{a}\varepsilon \\
&\leq |\tilde{P}(x) - J(x)| + |J(x) - W_{\lambda,\beta}(x)| + \frac{a-2}{a}\varepsilon \\
&\leq \frac{\varepsilon}{a} + \frac{\varepsilon}{a} + \frac{a-2}{a}\varepsilon = \varepsilon.
\end{aligned}$$

We now show Eq. (7.61) holds. Using Eqs. (7.64), (7.66), (7.67), and (7.68) we have that

$$\begin{aligned}
&\nabla P(x)^T f(x) + \lambda(1 - P(x))||x||_2^{2\beta}\\
&\le \nabla \tilde{P}(x)^T f(x) + \lambda(1 - \tilde{P}(x))||x||_2^{2\beta} - \lambda\frac{\varepsilon(a-2)}{a}||x||_2^{2\beta}\\
&\qquad - \nabla J(x)^T f(x) - \lambda(1 - J(x))||x||_2^{2\beta} + \frac{\varepsilon}{a}||x||_2^{2\beta}\\
&= (\nabla \tilde{P}(x) - \nabla J(x))^T f(x) + \lambda(J(x) - \tilde{P}(x))||x||_2^{2\beta} + \frac{\varepsilon}{a}\bigg(1 - \lambda(a-2)\bigg)||x||_2^{2\beta}\\
&\le ||\nabla \tilde{P}(x) - \nabla J(x)||_2 ||f(x)||_2 + \lambda R^{-2\beta}\frac{\varepsilon}{a}||x||_2^{2\beta}\\
&\qquad + \frac{\varepsilon}{a}(1 - \lambda(a-2))||x||_2^{2\beta}\\
&\le \left(\frac{\sup_{x\in B_R(0)} ||f(x)||_2}{R} + \lambda R^{-2\beta} + 1 - \lambda(a-2)\right)\frac{\varepsilon}{a}||x||_2^{2\beta}\\
&\le 0.
\end{aligned}$$

We now show Eq. (7.62) holds. From Eq. (7.67) we have that $\tilde{P}(x) > J(x) - \frac{\varepsilon}{aR^{2\beta}}||x||_2^{2\beta}$ for all $x \in (B_R(0))^{cl}$. Moreover, Eq. (7.65) we have that $J(x) = 1$ for all $x \in \partial B_R(0)$. Therefore $P(x) = \tilde{P}(x) + \frac{a-2}{a}\varepsilon > 1 + \frac{a-2}{a}\varepsilon - \frac{\varepsilon}{aR^{2\beta}}||x||_2^{2\beta} > 1 + \frac{a-3}{a}\varepsilon > 1$. Also from Eq. (7.67) we have that $\tilde{P}(0) = J(0)$. From Eq. (7.65) we have that $J(0) = 0$. Therefore $P(0) = \tilde{P}(0) + \frac{a-2}{a}\varepsilon > 0$. □

7.7 An SOS Optimization Problem For ROA Approximation

For a given ODE (7.1) we next propose a sequence of convex Sum-of-Squares (SOS) optimization problems, indexed by $d \in \mathbb{N}$. We show that the sequence of solutions, $\{P_d\}_{d\in\mathbb{N}}$, yields a sequence of sublevel sets which are contained inside the region of attraction of the ODE and which converge to the region of attraction of the ODE with respect to the volume metric as $d \to \infty$.

For given $f \in \mathcal{P}(\mathbb{R}^n, \mathbb{R}^n)$, $\lambda > 0$, $\beta \in \mathbb{N}$, $R > 0$ and integration region $\Lambda \subset \mathbb{R}^n$

consider the following sequence of SOS optimization problems indexed by $d \in \mathbb{N}$:

$$\begin{aligned}
&P_d \in \arg \min_{J \in \mathcal{P}_d(\mathbb{R}^n,\mathbb{R})} c^T \alpha && (7.69)\\
&J(x) = c^T Z_d(x),\\
&k_1, k_2, s \in \sum_{SOS}^{d} \text{ and } p \in \mathcal{P}_d(\mathbb{R}^n, \mathbb{R})\\
&J(0) \geq 0,\\
&k_1(x) = -\nabla J^T(x) f(x) - \lambda(1 - J(x))||x||_2^{2\beta} - s(x)(R^2 - ||x||_2^2),\\
&k_2(x) = (J(x) - 1) - p(x)(R^2 - ||x||_2^2),
\end{aligned}$$

where $\alpha_i = \int_\Lambda Z_{d,i}(x)dx$, recalling $Z_d : \mathbb{R}^n \to \mathbb{R}^{\mathcal{N}_d}$ is the vector of monomials of degree $d \in \mathbb{N}$ and $\mathcal{N}_d = \binom{d+n}{d}$.

We will show next, in Cor. 7.3, that the family of SOS optimization problems given in Eq. (7.69) yields an inner approximation of ROA_f for each $d \in \mathbb{N}$ (an approximation certifiably contained inside of ROA_f).

Corollary 7.3. *Consider $f \in \mathcal{P}(\mathbb{R}^n, \mathbb{R})$, $\lambda > 0$, $\beta \in \mathbb{N}$, $R > 0$ and $\Lambda \subset \mathbb{R}^n$. Suppose $ROA_f \subseteq B_R(0)$ and there exists $\eta > 0$ such that $B_\eta(0)$ is an exponentially stable set. Then we have that*

$$\{x \in B_R(0) : P_d(x) < 1\} \subseteq ROA_f \textit{ for all } d \in \mathbb{N}, \tag{7.70}$$

where P_d is any solution to the SOS Problem (7.69) for $d \in \mathbb{N}$.

Proof. Suppose P_d is any solution to the SOS Problem (7.69) for $d \in \mathbb{N}$. Then P_d satisfies the constraints of the SOS Problem (7.69) and thus satisfies Eqs. (7.37), (7.38), and (7.39) for $\Omega = B_R(0)$. Therefore, $W_{\lambda,\beta}(x) \leq P_d(x)$ for all $x \in B_R(0)$ by Prop. 7.3. Hence, it is clear that

$$\{x \in B_R(0) : P_d(x) < 1\} \subseteq \{x \in B_R(0) : W_{\lambda,\beta}(x) < 1\}. \tag{7.71}$$

Moreover, Cor. 7.1 shows $\{x \in B_R(0) : W_{\lambda,\beta}(x) < 1\} = ROA_f$ and thus Eq. (7.70) holds. □

Cor. 7.3 implies that solution maps initialized inside our ROA_f approximation asymptotically coverage to the origin. That is for any $d \in \mathbb{N}$ and for all $y \in \{x \in B_R(0) : P_d(x) < 1\}$ we have that $\lim_{t\to\infty} ||\phi_f(y,t)||_2 = 0$, where P_d is any solution to the SOS Problem (7.69) for $d \in \mathbb{N}$ (note this does not rule out the possibility that $\{x \in B_R(0) : P_d(x) < 1\} = \emptyset$).

Further to Cor. 7.3, we will next show, in Theorem 7.2, that for sufficiently large $\lambda > 0$ and $\beta \in \mathbb{N}$ the sequence of SOS optimization problems given in Eq. (7.69) yields a sequence of sets that tend to ROA_f with respect to the volume metric as $d \to \infty$. We first recall the volume metric (defined in Appendix A). For sets $A, B \subset \mathbb{R}^n$, we denote the volume metric as $D_V(A,B)$, where

$$D_V(A,B) := \mu((A/B) \cup (B/A)).$$

We note that D_V is a metric (Defn. A.1), as shown in Lem. A.1 (found in Appendix A).

Theorem 7.2. *Consider $f \in \mathcal{P}(\mathbb{R}^n, \mathbb{R})$ and integration region $\Lambda \subset \mathbb{R}^n$. Suppose there exists $\theta, \eta, R > 0$ such that $||D^\alpha f(x)||_2 < \theta$ for all $x \in B_R(0)$ and $||\alpha||_1 \le 2$, $B_\eta(0)$ is an exponentially stable set (Defn. 7.1) of the ODE (7.1), and $ROA_f \subset B_R(0)$. Then if $ROA_f \subseteq \Lambda \subset B_R(0)$, $\lambda > \theta\eta^{-2\beta}$ and $\beta > \frac{\theta}{2\delta} + \frac{1}{2}$ we have that*

$$\lim_{d\to\infty} D_V\bigg(ROA_f, \{x \in \Lambda : P_d(x) < 1\}\bigg) = 0, \tag{7.72}$$

where P_d is any solution to Problem (7.69) for $d \in \mathbb{N}$.

Proof. By Cor. 7.1 we have that $ROA_f = \{x \in \mathbb{R}^n : W_{\lambda,\beta}(x) < 1\}$. Moreover, since P_d satisfies the constraints of the SOS Problem (7.69) it follows that P_d satisfies Eqs. (7.37), (7.38), and (7.39) for $\Omega = B_R(0)$. Therefore, $W_{\lambda,\beta}(x) \le P_d(x)$ for all $x \in B_R(0)$ by Prop. 7.3. Thus, by Cor. A.1 (found in Appendix A) it follows that Eq. (7.72) holds if $\lim_{d \to \infty} ||P_d - W_{\lambda,\beta}||_{L^1(\Lambda,\mathbb{R})} = 0$. To show $\lim_{d\to\infty} ||P_d - W_{\lambda,\beta}||_{L^1(\Lambda,\mathbb{R})} = 0$ we must show for all $\varepsilon > 0$ there exists $D \in \mathbb{N}$ such that

$$\int_\Lambda |P_d(x) - W_{\lambda,\beta}(x)| dx < \varepsilon \text{ for all } d > D. \tag{7.73}$$

Now, let $\varepsilon > 0$. Then Theorem 7.1 shows there exists $J \in \mathcal{P}(\mathbb{R}^n, \mathbb{R})$ such that

$$\sup_{x \in B_R(0)} |J(x) - W_{\lambda,\beta}(x)| < \frac{\varepsilon}{\mu(\Lambda) + 1}, \tag{7.74}$$
$$\nabla J(x)^T f(x) < -\lambda(1 - J(x))||x||_2^{2\beta} \text{ for all } x \in B_R(0),$$
$$J(x) > 1 \text{ for all } x \in \partial B_R(0) \text{ and } J(0) > 0.$$

Since $B_R(0) = \{x \in \mathbb{R}^n : R^2 - ||x||_2^2 \geq 0\}$ and $\partial B_R(0) = \{x \in \mathbb{R}^n : R^2 - ||x||_2^2 \geq 0, ||x||_2^2 - R^2 \geq 0\}$ we have that by Putinar's Positivstellesatz (Theorem C.5 given in Appendix C) there exists $s_i \in \sum_{SOS}$ for $i \in \{1, ..., 5\}$ such that

$$- \nabla J(x)^T f(x) - \lambda(1 - J(x))||x||_2^{2\beta} - s_1(x)(R^2 - ||x||_2^2) = s_2(x), \text{ for all } x \in \mathbb{R}^n. \tag{7.75}$$

$$J(x) - 1 - (s_3(x) - s_4(x))(R^2 - ||x||_2^2) = s_5(x), \text{ for all } x \in \mathbb{R}^n. \tag{7.76}$$

Let $D := \max\{\max_{i=1,..,5}\{deg(s_i)\}, deg(J)\}$. Then from Eqs. (7.75) and (7.76) and since $J(0) > 0$ it follows that J is feasible to the SOS Problem (7.69) for any $d > D$. Since, P_d is the optimal solution to the SOS Problem (7.69) it follows that the objective function of the SOS Problem (7.69) evaluated at P_d is less than or equal to the objective function evaluated at the feasible solution J for $d > D$. That is by writing P_d and J with respect to the monomial vector, $P_d(x) = c_d^T Z_d(x)$ and $J(x) = b^T Z_{deg(J)}(x)$, we have that

$$\int_\Lambda P_d(x) dx = c_d^T \alpha \leq b^T \gamma = \int_\Lambda J(x) dx \text{ for all } d > D, \tag{7.77}$$

where $\alpha_i = \int_\Lambda Z_{d,i}(x)dx$, and $\gamma_i = \int_\Lambda Z_{deg(J),i}(x)dx$.

We now show Eq. (7.73). Using the fact $W_{\lambda,\beta}(x) \leq P_d(x)$ for all $x \in \Lambda$ together with Eqs. (7.74) and (7.77) we get that,

$$\int_\Lambda |P_d(x) - W_{\lambda,\beta}(x)|dx = \int_\Lambda P_d(x)dx - \int_\Lambda W_{\lambda,\beta}(x)dx$$
$$\leq \int_\Lambda J(x)dx - \int_\Lambda W_{\lambda,\beta}(x)dx$$
$$\leq \mu(\Lambda) \sup_{x\in\Lambda}\{|J(x) - W_{\lambda,\beta}(x)|\} < \varepsilon \text{ for all } d > D.$$

Hence by Cor. A.1 (found in Appendix A) it follows that Eq. (7.72) holds. □

7.8 Numerical Examples

We now present several numerical examples that demonstrate that by solving the SOS problem, given in Eq. (7.69), we are able to approximate the region of attraction of a nonlinear system. Note that for numerical implementation it is best to choose $\lambda > 0$ as small as possible. This is because the Lipschitz constant (given in Eq. (7.19)) of $W_{\lambda,\beta}$ (given in Eq. (7.10)) grows as $\lambda > 0$ increases. For these numerical examples, we solve Opt. (7.69) using SOSTOOLS, see Prajna *et al.* (2002a), to reformulate the problem as a Semi-Definite Programming (SDP) problem that is then solved by Sedumi, see Sturm (1999).

Example 7.1. *Consider the Van der Pol oscillator defined by the ODE:*

$$\dot{x}_1(t) = -x_2(t) \tag{7.78}$$
$$\dot{x}_2(t) = x_1(t) - x_2(t)(1 - x_1^2(t)).$$

In Fig. 7.1 we have plotted our estimation of the region of attraction of the ODE (7.78)*. Our estimation is given by the* 1*-sublevel set of the solution to the SOS optimization problem given in Eq.* (7.69) *for* $d = 12$*,* $\lambda = 0.05$*,* $\beta = 2$*,* $R = \sqrt{2^2 + 2.7^2} \approxeq 3.36$*,* $\Lambda = [-2, 2] \times [-2.7, 2.7]$*, and* $f = [-x_2, x_1 + x_2(x_1^2)]^T$*.*

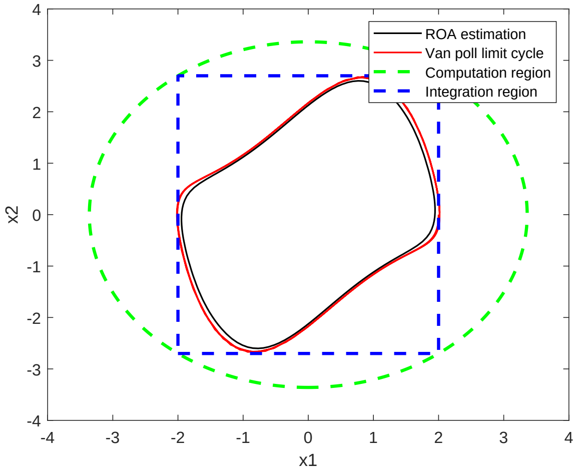

Figure 7.1: Graph showing an estimation of the region of attraction of the Van der Pol oscillator (Example 7.1) found by solving the SOS Problem (7.69). The black line is the 1-sublevel set of a solution to the SOS Problem (7.69). The red line is the boundary of the region of attraction found by simulating a reverse time trajectory using Matlab's `ODE45` function. The dotted blue line is the integration region, $\Lambda = (-2,2) \times (-2.7,2.7)$. The dotted green line is the computation region, $B_R(0)$ where $R = 3.36$.

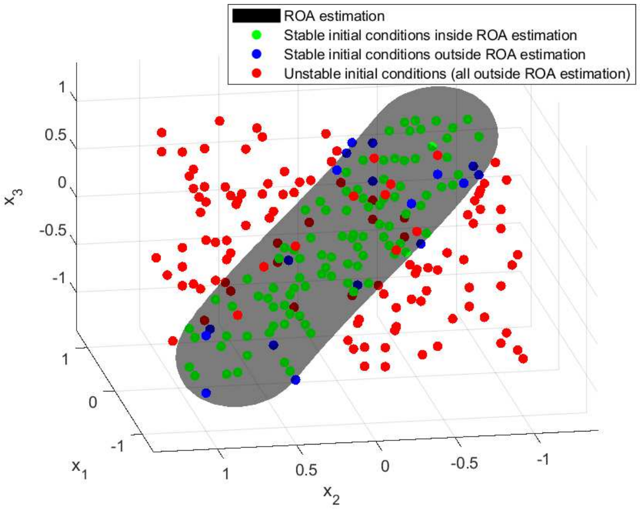

Figure 7.2: Graph showing an estimation of the region of attraction of servomechanism with multiplicative feedback control (Example 7.2). The estimation of the region of attraction is given by the transparent black sublevel set that is the 1-sublevel set of a solution to the SOS Problem (7.69). The scattered points are randomly generated initial conditions with associated trajectories (found using Matlab's `ODE45` function) that tend towards the origin (blue and green points) or away from the origin (red points).

Example 7.2. *Consider the third order servomechanism with multiplicative feedback control found in Ku and Chen (1958) given by the following ODE:*

$$T\frac{d^2y}{dt^2} + \frac{dy}{dt} + K_2(1 - K_3y^2)\frac{dy}{dt} + K_1y = 0, \tag{7.79}$$

where $T \in \mathbb{R}$ is a time constant and $K_1, K_2, K_2 \in \mathbb{R}$ are gain constants. We consider the case $T = K_2 = 1$ and $K_1 = K_3 = 1$. The ODE (7.79) *can be represented in the form $\dot{x}(t) = f(x(t))$ with*

$$f(x) = [x_2, x_3, (1/T)(-x_3 - K_2(1 - K_3x_1^2)x_2 - K_1x_1)]^T. \tag{7.80}$$

Through numerical experiments using Matlab's `ODE45` *function it was found that the ODE with vector field given in Eq.* (7.80) *appeared to have unbounded region of attraction. Therefore, for this system, Theorem 7.2 does not show that the sequence of sublevel sets to the solution to the SOS problems given in Eq.* (7.69) *converges to the region of attraction as $d \to \infty$. Nevertheless, in Fig. 7.2 we have plotted the* 1*-sublevel set of the solution to the SOS optimization problem given in Eq.* (7.69) *for $d = 10$, $\lambda = 0.5$, $\beta = 2$, $R = \sqrt{3}$, $\Lambda = [-1, 1]^3$ and f given in Eq.* (7.80)*. Fig. 7.2 indicates that even for systems with unbounded regions of attraction, our proposed SOS algorithm can provide arbitrarily good inner estimations of $ROA_f \cap \Lambda$, where $\Lambda \subset \mathbb{R}^n$ is some compact set. Through Monte Carlo simulation the volume of $ROA_f \cap \Lambda$ was estimated to be 0.3372 whereas the volume of our ROA approximation was found to be 0.2806, an error of 0.0566.*

7.9 Conclusion

For a given locally exponentially stable dynamical system, described by an ODE, we have proposed a family SOS optimization problems that yields a sequence of sublevel sets that converge to the region of attraction of the ODE with respect to the volume metric. In order to facilitate this result we proposed a new converse Lyapunov function that was shown to be globally Lipschitz continuous. We have provided several numerical examples of practical interest showing how our proposed family of SOS problems can provide arbitrarily good approximations of regions of attraction. In future work we aim extend this work to systems with weaker forms of stability and investigate systems with unbounded regions of attraction.

Chapter 8

A CONVERSE SUM OF SQUARES LYAPUNOV FUNCTION FOR OUTER APPROXIMATION OF MINIMAL ATTRACTOR SETS OF NONLINEAR SYSTEMS

> For those systems with bounded solutions, it is found that nonperiodic solutions are ordinarily unstable with respect to small modifications, so that slightly differing initial states can evolve into considerably different states.
>
> Edward Lorenz

8.1 Background and Motivation

Many dynamical systems described by nonlinear ODEs are unstable. Their associated solutions do not converge towards an equilibrium point, but rather converge towards some invariant subset of the state space called an attractor set. For a given ODE, in general, the existence, shape and structure of the attractor sets of the ODE are unknown. In this chapter we propose a new method for computing attractor sets. Similarly to Chapter 7, in this chapter we consider nonlinear Ordinary Differential Equations (ODEs) of the form

$$\dot{x}(t) = f(x(t)), \quad x(0) = x_0. \tag{8.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is the vector field and $x_0 \in \mathbb{R}^n$ is the initial condition. We denote the solution map (which exists and is continuous on $x \in X \subset \mathbb{R}^n$ when f is Lipschitz continuous and X is compact and invariant under f) of the ODE (8.1) by $\phi_f : X \times [0, \infty) \to \mathbb{R}^n$ where

$$\frac{d}{dt}\phi_f(x,t) = f(\phi_f(x,t)) \text{ for all } x \in X \text{ and } t \geq 0,$$
$$\phi_f(x,0) = x \text{ for all } x \in X.$$

An ODE is asymptotically stable about some equilibrium point, x_0, if there exists some neighborhood of the equilibrium, $\mathcal{N}(x_0)$, such that $\lim_{t\to\infty} \phi_f(x,t) = x_0$ for any $x \in \mathcal{N}(x_0)$. Attractor sets generalize the notion of asymptotic stability to where solutions tend towards a compact invariant subset of $\mathbb{R}^n$ (rather than being restricted to tend towards a single equilibrium point). Specifically, a compact set $A \subset \mathbb{R}^n$ is said to be an attractor set of the ODE (8.1) if for all $x \in A$ there exists $\varepsilon > 0$ such that $\lim_{t\to\infty} \inf_{y\in A} ||y - \phi_f(z,t)||_2 = 0$ for all $z \in \{y \in \mathbb{R}^n : ||x-y||_2 < \varepsilon\}$, and $x \in A$ implies $\phi_f(x,t) \in A$ for all $t \geq 0$. An attractor set is said to be minimal if there does not exists any other attractor sets contained within it.

Attractor sets provide information about the long term behavior of dynamical systems. The computation of attractor is used for design of secure private communications, see Cuomo *et al.* (1993); Zhao *et al.* (2018), the computation of Unstable Periodic Orbits (UPOs), see Lakshmi *et al.* (2020), and risk quantification of financial systems, see Gao *et al.* (2018). Furthermore, identification of minimal attractor sets can be used to bound the domain of strange attractors and "non-determinism" in chaos theory, see Lee (2016).

It is well known that the sublevel sets of Lyapunov functions yield attractor sets, see Lin *et al.* (1996). A Lyapunov function of an ODE is any function that is positive and decreases along the solution map of the ODE. In Li *et al.* (2005); Yu and Liao (2005) quadratic Lyapunov functions were used to estimate bounds for Lorenz attractor. In Goluskin (2020) attractor sets are indirectly approximated by searching for Sum-of-Squares (SOS) Lyapunov functions that provide bounds for $\sup_{(x,t)\in\Omega\times[0,\infty)} \Phi(\phi_f(x,t))$, where $\Omega \subset \mathbb{R}^n$, $\Phi : \mathbb{R}^n \to \mathbb{R}$, and ϕ_f is the solution map to some ODE (8.1). In Jones and Peet (2019c) attractor sets approximated by using SOS to search for Lyapunov functions outside some handpicked set $D \subset \mathbb{R}^n$ that is known to contain the attractor set. In Schlosser and Korda (2020); Wang *et al.* (2012b) an alternative SOS based method was proposed for attractor set approximation. Impressively, the method proposed in Schlosser and Korda (2020) was shown to provide an arbitrarily close approximation of an attractor set with respect to the Lebesgue measure. However, the methods in Schlosser and Korda (2020); Wang *et al.* (2012b) do not yield Lyapunov functions and hence any approximation found cannot be shown to also be an attractor set.

The problem of computing attractor sets is related to the problem of certifying the asymptotic stability of equilibrium points of an ODE (8.1); since certifying $A^* = \{0\}$ is an attractor set of an ODE (8.1) is equivalent to showing the asymptotic stability of the ODE (8.1) about $0 \in \mathbb{R}^n$. The use of SOS Lyapunov functions to certify the asymptotic stability of equilibrium points of an ODE (8.1) has been well treated in the literature, see Zheng *et al.* (2018); Anderson and Papachristodoulou (2015); Cunis *et al.* (2020); Valmorbida and Anderson (2017); Awrejcewicz *et al.* (2021); Jones and Peet (2021a); Ahbe (2020).

SOS programming provides a computationally tractable method for searching for SOS Lyapunov functions and hence computing attractor sets of ODEs. However, it is currently unknown how conservative it is to restrict the search Lyapunov functions to SOS polynomials. The goal of this chapter is then to: 1) Propose a Lyapunov characterization of attractor sets that is well suited to the problem of approximating the minimal attractor. 2) Show that for a given ODE with, sufficiently smooth vector field, there exists a sequence of SOS Lyapunov functions that yield optimal outer set approximations of attractor sets of the ODE. Note that, an optimal outer set approximation of a set $A^* \subset \mathbb{R}^n$ is any set $A \subset \mathbb{R}^n$ such that $A^* \subseteq A$ and $D(A^*, A)$

is minimal, where D is some set metric.

Specifically, given an ODE (8.1), we propose a new Lyapunov characterization of attractor sets. We show that if V satisfies,

$$\nabla V(x)^T f(x) \leq -(V(x) - 1) \text{ for all } x \in \Omega, \tag{8.2}$$

$$\{x \in \Omega : V(x) \leq 1\} \subseteq \Omega^\circ, \tag{8.3}$$

$$\{x \in \Omega : V(x) \leq 1\} \neq \emptyset, \tag{8.4}$$

where $\Omega \subset \mathbb{R}^n$ is some compact set and Ω° is the interior of Ω, then the 1-sublevel set of V is an attractor set of the ODE (8.1). To approximate the minimal attractor set of an ODE we then propose a sequence of d-degree optimization problem, each solved by a d-degree Sum-of-Square (SOS) polynomial function that satisfies Eqs. (8.2), (8.3) and (8.4), and has minimal 1-sublevel set. We show in Corollary 8.3 that the sequence of d-degree solutions to the optimization problem yield a sequence of 1-sublevel sets that each contain the minimal attractor of the ODE (8.1), are themselves attractor sets, and converge to the minimal attractor of the ODE (8.1) with respect to the volume metric.

Our proposed optimization problem for optimal outer set approximations of minimal attractors is solved by finding the SOS polynomial Lyapunov function with minimal 1-sublevel set volume. Unfortunately, there is no known closed expression for the volume of a sublevel set of a polynomial, see Lasserre (2019); making our optimization problem hard to solve. For SOS polynomials, $V = Z_d(x)^T P Z_d(x)$ where $P > 0$, rather than minimizing the sublevel set volume of V directly there exist several heuristics based on maximizing the eigenvalues of P. For instance in Dabbene *et al.* (2017) an optimization problem was proposed with Trace(P) objective function. Alternatively, $\log\det(P)$ functions have also been used as a metric for volume of $\{x \in \mathbb{R}^n : Z_d(x)^T P Z_d(x) \leq 1\}$, first being proposed in Magnani *et al.* (2005) and subsequently being used in the works of Ahmadi *et al.* (2017); Jones and Peet (2019c). In this chapter we also take a similar determinant maximizing approach and maximize $(\det(P))^{\frac{1}{n}}$ which is equivalent to maximizing $\log\det(P)$ but can be implemented on a larger array of SDP solvers, see Lofberg (2004).

In order to establish the convergence of our proposed method for optimal outer approximations of minimal attractor sets we propose a new converse Lyapunov theorem. Specifically, given an attractor set we show that there exists a sequence of SOS Lyapunov functions each satisfying Eqs. (8.2), (8.3), and (8.4), and whose 1-sublevel sets converge to the attractor set with respect to the volume metric.

Other important converse Lyapunov results concerning smooth Lyapunov functions include Lin *et al.* (1996); Teel and Praly (2000); where it is shown that asymptotically stable nonlinear systems with sufficiently smooth vector fields admit smooth (but not necessarily SOS) Lyapunov functions that can certify the stability of the systems. In terms of SOS converse Lyapunov theory we mention Peet and Papachristodoulou (2010) that showed that if the system's solutions converge locally to an equilibrium point at an exponential rate then there always exists a SOS Lyapunov function that can certify this local exponential stability. However, for asymptotically stable systems whose solutions converge to an equilibrium point at a sub-exponential rate there may not exist SOS Lyapunov functions that can certify this stability, as shown by the counterexample presented in Ahmadi and El Khadir (2018).

Before proceeding, we note that there is no contradiction with the counterexample found in Ahmadi and El Khadir (2018) and our proposed converse Lyapunov theorem (stated in Theorem 8.2). Although SOS Lyapunov functions cannot be used to certify the stability of equilibrium points in general (as proven by the counterexample from Ahmadi and El Khadir (2018)), Theorem 8.2 shows that SOS Lyapunov functions can be used to certify that arbitrarily small neighborhoods of equilibrium points are attractor sets. Hence SOS Lyapunov functions can certify the “stability” of arbitrarily small neighborhoods of equilibrium points.

The rest of the chapter is organized as follows. Attractor sets are defined in terms of solution maps of ODEs in Section 8.2. A Lyapunov type theorem is proposed in Section 8.3 that provides sufficient conditions for a set to be an attractor set. In Section 8.4, given an ODE, it is shown that there exists a sequence of SOS Lyapunov functions that yield a sequence of sublevel sets that converge to the minimal attractor set of the ODE. An SOS based algorithm for minimal attractor set approximation is then proposed in Section 8.5 and numerical examples are shown in Section 8.6. Finally our conclusion is given in Section 8.7.

8.2 Attractor Sets are Defined Using Solution Maps of Nonlinear ODEs

Consider a nonlinear Ordinary Differential Equation (ODE) of the form

$$\dot{x}(t) = f(x(t)), \quad x(0) = x_0 \in \mathbb{R}^n, \quad t \in [0,\infty), \tag{8.5}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is the vector field and $x_0 \in \mathbb{R}^n$ is the initial condition.

We now recall from Chapter 7 the definition of a solution map of an ODE. Given $X \subset \mathbb{R}^n$, $I \subset [0,\infty)$, and an ODE (8.5) we say any function $\phi_f : X \times I \to \mathbb{R}^n$ satisfying

$$\begin{aligned} &\frac{\partial \phi_f(x,t)}{\partial t} = f(\phi_f(x,t)) \text{ for } (x,t) \in X \times I, && (8.6)\\ &\phi_f(x,0) = x \text{ for } x \in X,\\ &\phi_f(\phi_f(x,t),s) = \phi_f(x,t+s) \text{ for } x \in X \ t,s \in I \text{ with } t+s \in I, \end{aligned}$$

is a solution map of the ODE (8.5) over $X \times I$. For simplicity throughout this chapter we will assume there exists a unique solution map to the ODE (8.5) over all $(x,t) \in \mathbb{R}^n \times [0,\infty)$. Note that the uniqueness and existence of a solution map sufficient for the purposes of this chapter, such as for initial conditions inside some invariant set (like the Basin of Attraction of an attractor set given in Eq. (8.9)) and for all $t \geq 0$, can be shown to hold under minor smoothness assumption on f, see Khalil (1996).

An important property of solution maps, we next recall in Lem. 8.1, is that they inherit the smoothness of their associated vector field. This smoothness property of solution maps is used in the proof of Prop. 8.1.

Lemma 8.1 (Smoothness of the solution map. Page 149 in Hirsch *et al.* (2004))**.** *Consider $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. Then if ϕ_f is a solution map (satisfying Eq.* (8.6)*) then $\phi_f \in C^1(\mathbb{R}^n \times \mathbb{R}, \mathbb{R})$.*

8.2.1 Attractor Sets of Nonlinear ODEs

A compact attractor set of the ODE (8.5) is defined as follows.

Definition 8.1. *We say that $A \subset \mathbb{R}^n$ is a compact* ***attractor set*** *of the ODE* (8.5)*, defined by $f : \mathbb{R}^n \to \mathbb{R}^n$, if*

1. *A is compact and nonempty ($A \neq \emptyset$).*

2. *A is a forward invariant set. That is if ϕ_f is a solution map of the ODE* (8.5) *we have that,*

$$\phi_f(x,t) \in A \text{ for all } x \in A \text{ and } t \geq 0. \tag{8.7}$$

3. *For each element of A there is a neighborhood of initial conditions for which the solution map asymptotically tends towards A. That is, for all $x \in A$ there exists $\delta > 0$ such that for any $\varepsilon > 0$ there exists $T \geq 0$ for which*

$$D(\phi_f(y,t), A) < \varepsilon \text{ for all } y \in B_\delta(x) \text{ and } t \geq T. \tag{8.8}$$

*Furthermore, we say A is a **minimal attractor set** if there does not exist any other attractor set, B, such that $B \subset A$, that is there exists $x \in A$ such that $x \notin B$ (B is strictly contained in A).*

For simplicity we will often refer to compact attractor sets as attractor sets (leaving out the word compact).

Note, in the case where $A \subset \mathbb{R}^n$ is a single point, that is $A = \{x_0\}$, the condition given in Eq. (8.8) reduces to the classical condition of asymptotic stability of the equilibrium point $x_0 \in \mathbb{R}^n$. That is, the condition given in Eq. (8.8) reduces to requiring the existence of $\delta > 0$ such that $\lim_{t \to \infty} ||\phi_f(x,t) - x_0||_2 = 0$ for all $x \in B_\delta(x_0)$.

Each attractor set of the ODE (8.5) has an associated set of initial conditions for which solution maps initialized at these initial conditions converge towards the attractor set as $t \to \infty$. We call this set the basin of attraction of the attractor set and define it next.

Definition 8.2. *Given an attractor set $A \subset \mathbb{R}^n$ of the ODE* (8.5) *(defined by $f : \mathbb{R}^n \to \mathbb{R}^n$) we define the **basin of attraction** of A as*

$$BOA_f(A) := \{x \in \mathbb{R}^n : \lim_{t \to \infty} D(A, \phi_f(x,t)) = 0\}. \tag{8.9}$$

In the special case when the minimal attractor set is a single point the attractor set is commonly referred to as an equilibrium point and its associated basin of attraction is referred to as the region of attraction. However, although this special case is important for stability analysis, in general attractor sets can take more complicated structures such as limit cycles and in dimensions three and above (chaotic) ”strange attractors”.

8.3 A Lyapunov Approach to Finding and Certifying Minimal Attractor Sets

In this section, we propose a new Lyapunov characterization of attractor sets. To explain the motivation for this new characterization, consider a typical Lyapunov characterization of attractor sets, as given in Lin *et al.* (1996).

Theorem 8.1 (Lin *et al.* (1996))**.** *Consider* $f \in LocLip(\mathbb{R}^n, \mathbb{R}^n)$. *Let* $A \subset \mathbb{R}^n$ *be a nonempty, compact and forward invariant set. Then* A *is an attractor set of the ODE* (8.5) *defined by* f *with* $BOA_f(A) = \mathbb{R}^n$ *if and only if there exists* $V \in C^\infty(\mathbb{R}^n, [0,\infty))$ *such that*

$$\kappa_1(D(A,x)) \leq V(x) \leq \kappa_2(D(A,x)) \text{ for all } x \in \mathbb{R}^n, \tag{8.10}$$

$$\nabla V(x)^T f(x) \leq -\kappa_3(D(A,x)) \text{ for all } x \in \mathbb{R}^n/A, \tag{8.11}$$

where κ_1 *and* κ_2 *are class* K_∞ *functions (recalling the definition of class* K_∞ *from Section 2) and* κ_3 *is a continuous positive definite function.*

Theorem 8.1 defines a method for certifying that a set $A \subset \mathbb{R}^n$ is an attractor set by searching for a Lyapunov function valid for A – an optimization problem with decision variable V. However, this formulation is not well-suited to the problem of *finding* **minimal** attractor sets - a bilinear problem wherein both the attractor set A and Lyapunov function V are (unknown) decision variables. To resolve this problem, we propose Prop. 8.1, wherein the proposed attractor set is defined as the 1-sublevel set of V and hence there is only a single decision variable. In Section 8.5, we will show that this formulation allows us to combine the problems of certification and volume minimization of the attractor set using SOS programming and determinant maximization.

Proposition 8.1. *Consider* $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. *Suppose there exists* $V \in C^1(\mathbb{R}^n, [0,\infty))$

such that

$$\nabla V(x)^T f(x) \leq -(V(x)-1) \text{ for all } x \in \Omega, \tag{8.12}$$

$$\{x \in \Omega : V(x) \leq 1\} \subseteq \Omega^\circ, \tag{8.13}$$

$$\{x \in \Omega : V(x) \leq 1\} \neq \emptyset, \tag{8.14}$$

where $\Omega \subset \mathbb{R}^n$ is a compact set. Then $\{x \in \Omega : V(x) \leq 1\}$ is an attractor set (Defn. 8.1) to the ODE (8.5) *defined by f.*

Note that the Lyapunov function V in Prop. 8.1 is not required to be positive semidefinite. However, later in Section 8.5 we will include a positivity constraint on V – allowing us to minimize the volume of the 1-sublevel set.

Proof of Proposition 8.1. Throughout this proof we will use the following notation: $S_a := \{x \in \Omega : V(x) \leq 1 + a\}$ where $a \geq 0$.

In order to prove S_0 is an attractor set we will split the remainder of the proof into the following parts showing:

1. S_0 is a compact set.
2. If $S_a \subseteq \Omega^\circ$, where $a \geq 0$, then S_a is an invariant set.
3. There exists $a > 0$ such that $S_a \subseteq \Omega^\circ$.
4. For any $x \in S_0$ and for all $\varepsilon > 0$ there exists $\delta > 0$ and $T \geq 0$ such that $D(\phi_f(y,T), S_0) < \varepsilon$ for all $t \geq T$ and $y \in B_\delta(x)$.
5. S_0 is an attractor set.

Proof S_0 is a compact set: Since V is continuous it follows that $S_0 = \{x \in \Omega : V(x) \leq 1\}$ is closed by Lemma C.3. Moreover, S_0 is bounded since $S_0 \subseteq \Omega^\circ$ and Ω is bounded. Since $S_0 \subset \mathbb{R}^n$ is closed and bounded it follows that S_0 is a compact set.

Proof $S_a \subset \Omega^\circ$ is an invariant set: We now prove that if $S_a \subseteq \Omega^\circ$, where $a \geq 0$, then S_a is an invariant set. To see this, suppose for contradiction that there exists $y \in S_a$ and $T \geq 0$ such that $\phi_f(y,T) \notin S_a$. That is $V(\phi_f(y,0)) \leq 1+a$ and $V(\phi_f(y,T)) > 1+a$. Now, since $V(\phi_f(y,\cdot))$ is continuous (since V is continuous, ϕ_f is continuous by Lem. 8.1, and the composition of continuous functions is continuous) it follows by the intermediate value theorem that there exists $0 \leq s_1 < s_2 \leq T$ such that $V(\phi_f(y,s_1)) = 1+a$ and $V(\phi_f(y,t)) > 1+a$ for all $t \in (s_1,s_2]$. Thus $\phi_f(y,s_1) \in S_a \subseteq \Omega^\circ$ but $\phi_f(y,t) \notin S_a$ for all $t \in (s_1,s_2]$. Since Ω° is open and $\phi_f(y,s_1) \in S_a \subseteq \Omega^\circ$ there exists $\varepsilon > 0$ such that $B_\varepsilon(\phi_f(y,s_1)) \subset \Omega^\circ$. Again, using the continuity of $V(\phi_f(y,\cdot))$ there exists $\delta > 0$ such that $\phi_f(y,s_1+s) \in B_\varepsilon(\phi_f(y,s_1)) \subseteq \Omega^\circ$ for all $s \in [0,\delta]$. Therefore, $V(\phi_f(y,t)) > 1+a$ and $\phi_f(y,t) \in \Omega^\circ$ for all $t \in (s_1,s_3]$, where $s_3 := \min\{s_2, s_1+\delta\}$. Applying the mean value theorem there exists $s_1 < c < s_3$ such that

$$\frac{d}{dt}V(\phi_f(y,c)) = \frac{V(\phi_f(y,s_3)) - V(\phi_f(y,s_1))}{s_3 - s_1} > \frac{1+a-1-a}{s_3-s_1} = 0. \tag{8.15}$$

On the other hand since $\phi_f(y,t) \in \Omega^\circ$ for all $t \in (s_1,s_3]$ it follows that $\phi_f(y,c) \in \Omega^\circ$ and therefore Eq. (8.12) can be applied to give

$$\frac{d}{dt}V(\phi_f(y,c)) \leq 1 - V(\phi_f(y,c)) < 1-1-a = -a, \tag{8.16}$$

using the fact that $c \in (s_1,s_3)$ and $V(\phi_f(y,t)) > 1+a$ for all $t \in (s_1,s_3]$.

Now Eqs. (8.15) and (8.16) contradict each other proving S_a is invariant.

Proof there exists $a > 0$ such that $S_a \subseteq \Omega^\circ$: Let $\alpha := \inf_{z\in\partial\Omega}\{V(z)-1\}$. We first claim that $\alpha > 0$. Since Ω is compact it follows that $\partial\Omega$ is compact. Then, by the extreme value theorem (using the fact V is continuous) there exists $z^* \in \partial\Omega$ such that $\alpha = V(z^*)-1$. Since $S_0 \subseteq \Omega^\circ$ it follows $z^* \notin S_0$ implying $V(z^*) > 1$. Therefore, $\alpha = V(z^*) - 1 > 1 - 1 = 0$.

Let $a \in (0, \alpha)$. We next claim $S_a \subseteq \Omega^\circ$. To see this suppose for contradiction that $S_a \nsubseteq \Omega^\circ$. Then there exists $y \in S_a$ and $y \in \partial\Omega$. Hence $V(y) \le a + 1 < \alpha + 1$ (since $y \in S_a$), but $\alpha + 1 \le V(z)$ for all $z \in \partial\Omega$ (since $y \in \partial\Omega$ and $\alpha := \inf_{z \in \partial\Omega}\{V(z)\}$) implying $V(y) \ge \alpha + 1$ providing a contradiction.

<u>Proof S_0 has an attracting neighborhood:</u> We now prove that for any $x \in S_0$ and for all $\varepsilon > 0$ there exists $\delta > 0$ and $T \ge 0$ such that $D(\phi_f(y, T), S_0) < \varepsilon$ for all $t \ge T$ and $y \in B_\delta(x)$; that is S_0 satisfies Eq. (8.8).

First note that, by Part 2 of the proof we know that there exists $a > 0$ such that $S_a \subseteq \Omega^\circ$ and by Part 1 of the proof we know that S_a is an invariant set.

Now, let $x \in S_0$ and $\varepsilon > 0$. Since V is continuous there exists $\delta > 0$ such that for all $y \in B_\delta(x)$ we have $|V(y) - V(x)| < \frac{a}{2}$, implying $V(y) < \frac{a}{2} + V(x) \le \frac{a}{2} + 1 < a + 1$ for all $y \in B_\delta(x)$. Therefore, for all $y \in B_\delta(x)$ we have $y \in S_a \subseteq \Omega^\circ$. Since, $S_a \subset \Omega^\circ$ is invariant it follows that for any $y \in B_\delta(x)$ we have $\phi_f(y, t) \in \Omega^\circ$ for all $t \ge 0$. Thus, by Eq. (8.12) we have that

$$\frac{d}{dt} V(\phi_f(y, t)) \le -(V(\phi_f(y, t)) - 1) \text{ for all } (y, t) \in B_\delta(x) \times [0, \infty).$$

Now using Gronwall's inequality (Lem. C.2) and the fact $y \in S_a$ we have that

$$V(\phi_f(y, t)) - 1 \le e^{-t}(V(y) - 1) \le a e^{-t} \text{ for all } (y, t) \in B_\delta(x) \times [0, \infty).$$

Therefore, it now follows for any $\eta > 0$ that

$$\phi_f(y, t) \in S_\eta \text{ for all } y \in B_\delta(x) \text{ and } t \ge \ln\left(\frac{a}{\eta}\right). \tag{8.17}$$

We now construct $\eta > 0$ such that $S_\eta \subseteq B_\varepsilon(S_0)$ (recalling the notation $B_\varepsilon(S_0)$ is defined in Sec 2) implying that if $\phi_f(y, t) \in S_\eta$ for all $t \ge T$ then $D(S_0, \phi_f(y, t)) < \varepsilon$ for all $t \ge T$; therefore proving S_0 has an attracting neighborhood ie S_0 satisfies Eq. (8.8).

First note that if $\Omega / B_\varepsilon(S_0) = \emptyset$ then $\Omega \subseteq B_\varepsilon(S_0)$ and we can trivially take $\eta = a$. Then by Eq. (8.17) we have that $\phi_f(y,t) \in S_a \subseteq \Omega^\circ \subseteq B_\varepsilon(S_0)$ for all $y \in B_\delta(x)$ and $t \geq 0$. Thus, $D(\phi_f(y,t), S_0) < \varepsilon$ for all $y \in B_\delta(x)$ and $t \geq 0$. Hence, Eq. (8.8) is satisfied.

Let us now consider the case $\Omega / B_\varepsilon(S_0) \neq \emptyset$. Let $\eta \in (0, b)$ where $b = \min\{\inf_{z \in \Omega / B_\varepsilon(S_0)} V(z) - 1, \frac{a}{2}\}$, where $\inf_{z \in \Omega / B_\varepsilon(S_0)} V(z)$ exists since $\Omega / B_\varepsilon(S_0)$ is compact and V is continuous. Note that $b > 0$ since $a > 0$ and $\inf_{z \in \Omega / B_\varepsilon(S_0)} V(z) - 1 > 0$ (because $\Omega / B_\varepsilon(S_0)$ is compact so by the extreme value theorem there exists $z^* \in \Omega / B_\varepsilon(S_0)$ such that $V(z^*) = \inf_{z \in \Omega / B_\varepsilon(S_0)} V(z)$ and since $z^* \notin S_0$ it follows that $V(z^*) > 1$).

We now claim that $S_\eta \subseteq B_\varepsilon(S_0)$. First we note that $S_\eta \subseteq \Omega^\circ$ since $S_\eta \subset S_a$ and $S_a \subset \Omega^\circ$. Now suppose for contradiction that $S_\eta \not\subseteq B_\varepsilon(S_0)$. Then there exists $w \in S_\eta \subseteq \Omega$ such that $w \notin B_\varepsilon(S_0)$ implying $w \in \Omega / B_\varepsilon(S_0)$. Now, $V(w) \leq \eta + 1 < \inf_{z \in \Omega / B_\varepsilon(S_0)}\{V(z)\} \leq V(w)$ implying $0 < 0$, providing a contradiction.

Proof S_0 is an attractor set: Now since we have shown S_0 is a compact set, is non-empty (Eq. (8.14)), and satisfies Eqs. (8.7) and (8.8) it follows that S_0 is an attractor set. □

If V and Ω satisfy Eqs. (8.12), (8.13), and (8.14) (as in Prop. 8.1) and $\{x \in \Omega : V(x) \leq 1 + a\} \subseteq \Omega^\circ$ for some $a \geq 0$, then we next show that $\{x \in \Omega : V(x) \leq 1 + a\}$ is a subset of the basin of attraction of the attractor set $\{x \in \Omega : V(x) \leq 1\}$.

Corollary 8.1. *Consider $f \in LocLip(\mathbb{R}^n, \mathbb{R}^n)$. Suppose there exists $V \in C^1(\mathbb{R}^n, [0, \infty))$ and a compact set $\Omega \subset \mathbb{R}^n$ satisfying Eqs.* (8.12), (8.13), *and* (8.14) *(as in Prop. 8.1). Then, for any $a > 0$ such that $\{x \in \Omega : V(x) \leq 1 + a\} \subseteq \Omega^\circ$ it follows that $\{x \in \Omega : V(x) \leq 1 + a\} \subseteq BOA_f(\{x \in \Omega : V(x) \leq 1\})$.*

Proof. Follows by a similar argument to the proof of Prop. 8.1 □

In Prop. 8.1 we have shown that if a function V satisfies Eqs. (8.12), (8.13) and (8.14) then the 1-sublevel set of V is an attractor set of the ODE defined by f. In the next section we now prove that these Lyapunov characterizations of attractor sets are not conservative, even when V is restricted to be an SOS polynomial.

8.4 Converse Lyapunov Functions for Attractor Set Characterization

In the previous section, we have shown that if there exists a function V which satisfies Eqs. (8.12), (8.13) and (8.14), then the set $\{x \in \Omega : V(x) \leq 1\}$ is an attractor set of the ODE defined by $f \in LocLip(\mathbb{R}^n, \mathbb{R}^n)$. In this section, we show that for **any** attractor set $A \subset \mathbb{R}^n$ and any $\epsilon > 0$, there exists an SOS function V which satisfies Eq. (8.12) and for which $A \subset \{x \in \Omega : V(x) \leq 1\}$ and $D_V(A, \{x \in \Omega : V(x) \leq 1\}) \leq \epsilon$. This implies that the Lyapunov characterization of attractor sets in Section 8.3 is not conservative and furthermore, these conditions remain tight even when the Lyapunov functions are constrained to be SOS. In Section 8.5, we will use this result to propose a sequence of SOS programming problems whose limit yields an attractor set which is arbitrarily close to the minimal attractor set.

To begin, we quote a result on existence of smooth converse Lyapunov function from Teel and Praly (2000), which was based on a Yoshizawa type Lyapunov function, found in Yoshizawa (1966), of form

$$W(x) := \sup_{t \geq 0} \{e^t \kappa(D(A, \phi_f(x,t)))\},$$

where $\kappa \in K_\infty$.

Corollary 8.2 (Cor. 2 in Teel and Praly (2000)). *Consider $f \in LocLip(\mathbb{R}^n, \mathbb{R}^n)$. The set $A \subset \mathbb{R}^n$ is an attractor set to the ODE* (8.5) *if and only if there exists $V \in C^\infty(BOA_f(A), \mathbb{R})$ such that*

1. *$V(x) \geq 0$ for all $x \in BOA_f(A)$ and $V(x) = 0$ if and only if $x \in A$.*

2. $\nabla V(x)^T f(x) \le -V(x)$ *for all* $x \in BOA_f(A)$*.*

In Thm. 8.2 we consider polynomial approximations of $\sqrt{V(x)+\gamma}$, for some $\gamma > 0$ and with V as defined in Cor. 8.2, to show that for any given attractor set $A \subset \mathbb{R}^n$ there exists a sequence of Sum-of-Squares polynomials, each satisfying Eq. (8.12), each of whose 1-sublevel sets contain A, and whose 1-sublevel sets converge to A (with respect to the volume metric).

Theorem 8.2. *For* $f \in LocLip(\mathbb{R}^n, \mathbb{R}^n)$*, suppose* $A \subset \mathbb{R}^n$ *is an attractor set of the ODE* (8.5) *defined by* f *and let* Ω *be any compact set such that* $A \subseteq \Omega^\circ$ *and* $\Omega \subset BOA_f(A)$*. Then there exists* $\alpha > 0$*,* $N \in \mathbb{N}$*, and a sequence,* $\{P_d\}_{d \in \mathbb{N}} \subset \sum_{SOS}(\mathbb{R}^n, \mathbb{R})$ *with* $P_d \in \sum_{SOS}^d(\mathbb{R}^n, \mathbb{R})$ *for all* $d \in \mathbb{N}$*, such that*

1. $\nabla P_d(x)^T f(x) < -(P_d(x) - 1)$ *for all* $x \in \Omega$ *and* $d \ge N$*.*

2. $P_d(x) > 1 + \alpha$ *for all* $x \in \partial\Omega$ *and* $d \ge N$*.*

3. $A \subseteq \{x \in \Omega : P_d(x) \le 1\}$ *for all* $d \ge N$*.*

4. $\lim_{d \to \infty} D_V(A, \{x \in \Omega : P_d(x) \le 1\}) = 0$ *(recalling* D_V *denotes the volume metric defined in Appendix A).*

Proof. Let us suppose $A \subset \mathbb{R}^n$ is an attractor set to the ODE (8.5). By Cor. 8.2 there exists $W \in C^\infty(BOA_f(A), \mathbb{R})$ such that

1. $W(x) \ge 0$ for all $x \in BOA_f(A)$ and $W(x) = 0$ if and only if $x \in A$.

2. $\nabla W(x)^T f(x) \le -W(x)$ for all $x \in BOA_f(A)$.

For $\gamma > 0$ let $J(x) := W(x) + \gamma$. It trivially follows that since $A \subseteq \Omega \subset BOA_f(A)$ we have

$$\nabla J(x)^T f(x) \le -(J(x) - \gamma) \text{ for all } x \in BOA_f(A). \tag{8.18}$$

$$A = \{x \in \Omega : J(x) \le \gamma\}. \tag{8.19}$$

Now, to prove Theorem 8.2 we first show that there exists $\gamma > 0$, $\alpha > 0$, $N \in \mathbb{N}$, and $\{G_d\}_{d \in \mathbb{N}} \subset \sum_{SOS}$ such that

$$\nabla G_d(x)^T f(x) < -(G_d(x) - \gamma) \text{ for all } x \in \Omega \text{ and } d > N. \tag{8.20}$$

$$G_d(x) \leq J(x) \text{ for all } x \in \Omega \text{ and } d \geq N. \tag{8.21}$$

$$G_d(x) \geq \gamma(1 + \alpha) \text{ for all } x \in \partial\Omega \text{ and } d \geq N \tag{8.22}$$

$$\lim_{d \to \infty} ||G_d - J||_{L^1(\Omega, \mathbb{R})} = 0. \tag{8.23}$$

To show that there exists $\{G_d\}_{d \in \mathbb{N}} \subset \sum_{SOS}$ that satisfies Eqs. (8.20), (8.21), (8.22) and (8.23) we take the square route of J and approximate this by a polynomial. We next argue that the square route of J is sufficiently smooth for polynomial approximation.

Since $W(x) \geq 0$ it follows that $J(x) \geq \gamma > 0$. Therefore $H(x) := +\sqrt{J(x)}$ is differentiable, that is $H \in C^1(\mathbb{R}^n, \mathbb{R})$, since the function $g(x) := \sqrt{x}$ is differentiable over $(0, \infty)$ and J maps onto $(\gamma, \infty) \subset (0, \infty)$. Using the fact H is differentiable and applying the chain rule we find that

$$\begin{aligned} ||\nabla H(x)||_2 = ||\nabla \sqrt{W(x) + \gamma}||_2 &= \frac{1}{2\sqrt{W(x) + \gamma}} ||\nabla W(x)||_2 \\ &\leq \frac{1}{2\sqrt{\gamma}} ||\nabla W(x)||_2 \leq \frac{C}{2\sqrt{\gamma}} \text{ for all } x \in \Omega, \end{aligned} \tag{8.24}$$

where $C := \sup_{x \in \Omega} ||\nabla W(x)||_2$. Note that the first inequality in Eq. (8.24) follows since $W(x) \geq 0$ for all $x \in \Omega$ implies $\frac{1}{\sqrt{W(x)+\gamma}} \leq \frac{1}{\sqrt{\gamma}}$ for all $x \in \Omega$.

Moreover, applying the chain rule, the inequality in Eq. (8.18), and the fact that $\frac{1}{2\sqrt{J(x)}} \geq 0$ for all $x \in \Omega$ we find that,

$$\begin{aligned} \nabla H(x)^T f(x) = \nabla \sqrt{J(x)}^T f(x) &= \frac{1}{2\sqrt{J(x)}} \nabla J(x)^T f(x) \\ &\leq \frac{-1}{2\sqrt{J(x)}} (J(x) - \gamma) \text{ for all } x \in \Omega. \end{aligned}$$

Then it follows that H satisfies

$$2H(x)\nabla H(x)^T f(x) \le -(H^2(x) - \gamma) \text{ for all } x \in \Omega. \tag{8.25}$$

To show that there exists $\{G_d\}_{d\in\mathbb{N}} \subset \sum_{SOS}$ that satisfies Eq. (8.23) we must show that that there exists $\{G_d\}_{d\in\mathbb{N}} \subset \sum_{SOS}$ such that for any $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$||G_d - J||_{L^1(\Omega,\mathbb{R})} < \varepsilon \text{ for all } d \ge N. \tag{8.26}$$

In order to do this we first approximate H by a polynomial. Let,

$$\gamma > \frac{M_1 C}{2} > 0, \tag{8.27}$$

where $M_1 := \sup_{x\in\Omega} ||f(x)||_2$ and recalling $C := \sup_{x\in\Omega} ||\nabla W(x)||_2$. Note that γ from Eq. (8.27) is a constant that only depends on the problems data (f and Ω).

Also let,

$$\begin{aligned}
&\varepsilon > 0, && (8.28)\\
&0 < \alpha < \frac{1}{\gamma} \min_{x\in\partial\Omega} W(x) && (8.29)\\
&0 < \theta < \min\left\{\varepsilon, (\mu(\Omega)+1)(\min_{x\in\partial\Omega} W(x) - \gamma\alpha)\right\} && (8.30)\\
&0 < \delta < \min\left\{\frac{\sqrt{\gamma} - M_1M_3}{M_1M_2 + M_1M_3 + M_2}, \frac{\sqrt{\gamma}}{M_2}\right\}, && (8.31)\\
&0 < \sigma < \min\left\{\frac{2(\sqrt{\gamma} - (M_1M_2 + M_1M_3 + M_2)\delta - M_1M_3)}{(2M_1+1)\delta^2 + 2(1+M_1)\delta + 1},\right. && (8.32)\\
&\left.\frac{2(\sqrt{\gamma} - M_2\delta)}{(\delta+1)^2}, \frac{\sqrt{\theta}}{\sqrt{2(\mu(\Omega)+1)(\delta+1)}}, \frac{\theta}{4M_2(\delta+1)(\mu(\Omega)+1)}\right\},
\end{aligned}$$

recalling $M_1 := \sup_{x\in\Omega} ||f(x)||_2 \ge 0$ and where $M_2 := \sup_{x\in\Omega} |H(x)| \ge 0$, and $M_3 := \sup_{x\in\Omega} ||\nabla H(x)||_2 \ge 0$. Note that $\alpha > 0$ since $\gamma > 0$ and $\min_{\delta\in\partial\Omega} W(x) > 0$ (since $A \subseteq \Omega^\circ$ implies $A \cap \partial\Omega = \emptyset$ and $W(x) = 0$ iff $x \in A$). Also note that $\theta > 0$ since $\varepsilon > 0$ and $\min_{x\in\partial\Omega} W(x) - \gamma\alpha$ by Eq. (8.29). Moreover, $\delta > 0$ since $\gamma > \frac{M_1 C}{2}$ (by Eq. (8.27)) and $M_3 \le \frac{C}{2\sqrt{\gamma}}$ (by Eq. (8.24)) implying that $\sqrt{\gamma} - M_1M_3 > 0$. Furthermore, $\sigma > 0$

since $\delta < \frac{\sqrt{\gamma} - M_1 M_3}{M_1 M_2 + M_1 M_3 + M_2}$ implying $2(\sqrt{\gamma} - (M_1 M_2 + M_1 M_3 + M_2)\delta - M_1 M_3) > 0$ and $\delta < \frac{\sqrt{\gamma}}{M_2}$ implying $2(\sqrt{\gamma} - M_2 \delta) > 0$.

Now, by Theorem C.2 there exists polynomials $\{R_d\}_{d \in \mathbb{N}} \subset \mathcal{P}(\mathbb{R}^n, \mathbb{R})$ and $N \in \mathbb{N}$ such that

$$|H(x) - R_d(x)| < \delta\sigma \text{ for all } d \geq N. \tag{8.33}$$

$$||\nabla H(x) - \nabla R_d(x)||_2 < \delta\sigma \text{ for all } d \geq N. \tag{8.34}$$

Since $R_d \in \mathcal{P}(\mathbb{R}^n, \mathbb{R})$ for all $d \in \mathbb{N}$ it follows that $G_d(x) := (R_d(x) - \sigma)^2$ is a SOS polynomial, that is $G_d \in \sum_{SOS}$ for all $d \in \mathbb{N}$. We first show that G_d satisfies Eq. (8.20). Recalling $\gamma > \frac{M_1 C}{2}$ (by Eq. (8.27)), $M_1 := \sup_{x \in \Omega} ||f(x)||_2$ and $C := \sup_{x \in \Omega} ||\nabla W(x)||_2$, it follows that

$$
\begin{aligned}
&\nabla G_d(x)^T f(x) + (G_d(x) - \gamma) \\
&= (\nabla (R_d(x) - \sigma)^2)^T f(x) + ((R_d(x) - \sigma)^2 - \gamma) \\
&= 2(R_d(x) - \sigma)\nabla R_d(x)^T f(x) + (R_d^2(x) - \gamma) - 2\sigma R_d(x) + \sigma^2 \\
&\leq 2R_d(x)\nabla R_d(x)^T f(x) - 2H(x)\nabla H(x)^T f(x) + (R_d^2(x) - H^2(x)) \\
&\qquad\qquad - 2\sigma R_d(x) + \sigma^2 - 2\sigma \nabla R_d(x)^T f(x) \\
&= 2(R_d(x) - H(x))\nabla R_d(x)^T f(x) + 2H(x)\nabla(R_d - H)(x)^T f(x) \\
&\qquad\qquad + (R_d(x) - H(x))(R_d(x) + H(x)) + 2\sigma(H(x) - R_d(x)) \\
&\qquad\qquad - 2\sigma\nabla(R_d - H)(x)^T f(x) - 2\sigma\nabla H(x)^T f(x) \\
&\qquad\qquad - 2\sigma H(x) + \sigma^2 \\
&\leq 2|R_d(x) - H(x)| ||\nabla R_d(x)||_2 ||f(x)||_2 \\
&\qquad\qquad + 2H(x)||\nabla R_d(x) - \nabla H(x)||_2 ||f(x)||_2 \\
&\qquad\qquad + |R_d(x) - H(x)|(|R_d(x)| + H(x)) + 2\sigma|H(x) - R_d(x)| \\
&\qquad\qquad + 2\sigma||\nabla(R_d - H)(x)||_2 ||f(x)||_2 + 2\sigma||\nabla H(x)||_2 ||f(x)||_2 \\
&\qquad\qquad - 2\sigma\sqrt{\gamma} + \sigma^2 \\
&\leq 2\delta\sigma M_1(||\nabla(R_d - H)(x)||_2 + ||\nabla H(x)||_2) + 2M_1M_2\delta\sigma \\
&\qquad\qquad + \delta\sigma(|R_d(x) - H(x)| + H(x) + M_2) + 2\delta\sigma^2 + 2M_1\delta\sigma^2 \\
&\qquad\qquad + 2M_1M_3\sigma - 2\sigma\sqrt{\gamma} + \sigma^2 \\
&\leq 2M_1\delta^2\sigma^2 + 2M_1M_3\delta\sigma + 2M_1M_2\delta\sigma + \delta^2\sigma^2 + 2M_2\delta\sigma \\
&\qquad\qquad + 2\delta\sigma^2 + 2M_1\delta\sigma^2 + 2M_1M_3\sigma - 2\sqrt{\gamma}\sigma + \sigma^2 \\
&= \sigma\Bigg(((2M_1 + 1)\delta^2 + 2(1 + M_1)\delta + 1)\sigma \\
&\qquad\qquad + 2((M_1M_2 + M_1M_3 + M_2)\delta + M_1M_3 - \sqrt{\gamma})\Bigg) \\
&< 0 \text{ for all } x \in \Omega \text{ and } d \geq N. \qquad (8.35)
\end{aligned}
$$

Where all the equalities in Eq. (8.35) follow from rearranging terms or adding and subtracting terms. The first inequality in Eq. (8.35) follows by applying the inequality in Eq. (8.25). The second inequality in Eq. (8.35) follows by the triangle inequality and the Cauchy Swarz in inequality. The third and fourth inequalities in Eq. (8.35) follows by Eqs. (8.33) and (8.34). Finally, the last inequality (the fifth inequality) in Eq. (8.35) follows by Eq. (8.32).

We now show G_d satisfies Eq. (8.21).

$$\begin{aligned}
&G_d(x) - J(x) = (R_d(x) - \sigma)^2 - H(x)^2 \\
&= R_d(x)^2 - 2\sigma R_d(x) + \sigma^2 - H(x)^2 \\
&= (R_d(x) - H(x))(R_d(x) + H(x)) + 2\sigma(H(x) - R_d(x)) - 2\sigma H(x) + \sigma^2 \\
&\leq \delta\sigma(\delta\sigma + 2M_2) + 2\delta\sigma^2 - 2\sigma\sqrt{\gamma} + \sigma^2 \\
&= \sigma\Bigg((\delta^2 + 2\delta + 1)\sigma + 2M_2\delta - 2\sqrt{\gamma}\Bigg) \qquad (8.36) \\
&< 0 \text{ for all } x \in \Omega \text{ and } d \geq N.
\end{aligned}$$

Where the first inequality in Eq. (8.36) follows using Eq. (8.33) and the fact $H(x) \geq \sqrt{\gamma}$ for all $x \in \Omega$. The second inequality in Eq. (8.36) follows by Eq. (8.32) ($\sigma < \frac{2\sqrt{\gamma}-2M_2\delta}{(\delta+1)^2}$).

We now show G_d satisfies Eq. (8.22).

$$\begin{aligned} J(x) - G_d(x) &= J(x) - (R_d(x) - \sigma)^2 \\ &= H(x)^2 - R_d(x)^2 + 2\sigma R_d(x) - \sigma^2 \\ &= (H(x) - R_d(x))(H(x) + R_d(x)) + 2\sigma(R_d(x) - H(x)) + 2\sigma H(x) - \sigma^2 \\ &\le \delta\sigma(2M_2 + \delta\sigma) + 2\delta\sigma^2 + 2\sigma M_2 - \sigma^2 \\ &= (\delta^2 + 2\delta - 1)\sigma^2 + 2M_2(\delta + 1)\sigma \\ &\le (\delta^2 + 2\delta + 1)\sigma^2 + 2M_2(\delta + 1)\sigma \\ &< \frac{\theta}{\mu(\Omega) + 1} \text{ for all } x \in \Omega \text{ and } d \ge N. \end{aligned} \tag{8.37}$$

Where the first inequality in Eq. (8.37) follows using Eq. (8.33). The second inequality in Eq. (8.37) follows by $\sigma > 0$ and $-1 < 1$. The third inequality in Eq. (8.37) follows by Eq. (8.32) ($\sigma < \frac{\sqrt{\theta}}{\sqrt{2(\mu(\Omega)+1)(\delta+1)}}$ and $\sigma < \frac{\theta}{4M_2(\delta+1)(\mu(\Omega)+1)}$). Now by rearranging Eq. (8.37) and using the fact that $J(x) := W(x) + \gamma$ we have that,

$$\begin{aligned} G_d(x) &> J(x) - \frac{\theta}{\mu(\Omega) + 1} \\ &= W(x) + \gamma - \frac{\theta}{\mu(\Omega) + 1} \\ &= \gamma + \frac{(\mu(\Omega) + 1)(\min_{x \in \partial\Omega} W(x)) - \theta}{\mu(\Omega) + 1} \\ &> \gamma(1 + \alpha) \text{ for all } x \in \partial\Omega \text{ and } d \ge N. \end{aligned} \tag{8.38}$$

Where the first inequality in Eq. (8.38) follows by Eq. (8.37) and the second inequality follows by Eq. (8.30). Hence Eq. (8.38) shows Eq. (8.22) holds.

We now show G_d satisfies Eq. (8.23) by showing G_d satisfies Eq. (8.26). By Eqs. (8.36) and (8.37) and the fact that $\theta < \varepsilon$ (Eq. (8.30)), it follows that

$$|G_d(x) - J(x)| < \frac{\varepsilon}{\mu(\Omega) + 1} \text{ for all } x \in \Omega \text{ and } d \ge N, \tag{8.39}$$

and thus

$$||G_d - J||_{L^1(\Omega,\mathbb{R})} \leq \sup_{x\in\Omega} |G_d(x) - J(x)|\mu(\Omega) < \varepsilon \text{ and } d \geq N.$$

Therefore Eq. (8.23) holds.

Now, set $P_d(x) := \frac{G_d(x)}{\gamma}$. Recall that $\gamma > 0$ from Eq. (8.27) is a constant that only depends on the problems data (f and Ω) and not $d \in \mathbb{N}$. Therefore, $\lim_{d\to\infty} P_d = \frac{1}{\gamma}\lim_{d\to\infty} G_d$. Moreover, it follows that $\{P_d\}_{d\in\mathbb{N}} \subset \sum_{SOS}$ since $\{G_d\}_{d\in\mathbb{N}} \subset \sum_{SOS}$ and $\gamma > 0$ (by Eq. (8.27)). Furthermore, it follows by Eqs. (8.20), (8.21), (8.22), and (8.23) that

$$\nabla P_d(x)^T f(x) < -(P_d(x) - 1) \text{ for all } x \in \Omega \text{ and } d \geq N, \tag{8.40}$$

$$P_d(x) \leq \tilde{J}(x) \text{ for all } x \in \Omega \text{ and } d \geq N, \tag{8.41}$$

$$P_d(x) > 1 + \alpha \text{ for all } x \in \partial\Omega \text{ and } d \geq N \tag{8.42}$$

$$\lim_{d\to\infty} ||P_d - \tilde{J}||_{L^1(\Omega,\mathbb{R})} = 0, \tag{8.43}$$

where $\tilde{J}(x) = \frac{J(x)}{\gamma}$.

We now argue that Theorem 8.2 is proven. Eq. (8.19) implies that $A = \{x \in \Omega : \tilde{J}(x) \leq 1\}$. Then Eq. (8.41) implies that $A \subseteq \{x \in \Omega : P_d(x) \leq 1\}$ for all $d \geq N$. Moreover, Eqs. (8.41) and (8.43) together with Theorem A.1 (found in Appendix A) imply that $\lim_{d\to\infty} D_V(\{x \in \Omega : \tilde{J}(x) \leq 1\}, \{x \in \Omega : P_d(x) \leq 1\}) = 0$, implying $\lim_{d\to\infty} D_V(A, \{x \in \Omega : P_d(x) \leq 1\}) = 0$ (since $A = \{x \in \Omega : \tilde{J}(x) \leq 1\}$). □

Remark 8.1. *Theorem 8.2 shows that for any attractor set, A, there exists an SOS polynomial, P, that satisfies the Lyapunov conditions (Eqs.* (8.12), (8.13) *and* (8.14)*) of Prop. 8.1 that has a* 1*-sublevel set arbitrarily close to the attractor set with respect to the volume metric. Note that P satisfies Eq.* (8.12) *directly from the statement of Theorem 8.2. Also note that P satisfies Eq.* (8.13) *since by Theorem 8.2 we have that*

$P(x) > 1$ for all $\delta\Omega$. Also note that Eq. (8.14) *since by Theorem 8.2 we have that $A \subseteq \{x \in \Omega : P(x) \leq 1\}$ and since $A \neq \emptyset$ it follows that $\{x \in \Omega : P(x) \leq 1\} \neq \emptyset$.*

Theorem 8.2 shows that the Lyapunov characterization of attractor sets proposed in Section 8.3 is not conservative and that this non conservatism is retained even if the Lyapunov functions are constrained to be SOS. However, in order to apply the results of Sections 8.3 and 8.4 to compute outer approximations of minimal attractors, we require an algorithm which can enforce the Lyapunov inequality conditions of Prop. 8.1 while minimizing the volume of the 1-sublevel set of the Lyapunov function. In the following section propose such an algorithm based on convex optimization and SOS programming.

8.5 A Family of SOS Problems for Minimal Attractor Set Approximation

In Section 8.3, we proposed a Lyapunov characterization of attractor sets for a given ODE defined by a vector field f. In Section 8.4, we showed that this characterization is not conservative even if the Lyapunov functions are constrained to be SOS. Given these two results, we may now formulate a polynomial optimization characterization of the minimal attractor set $A^* \subset \Omega$ of a given ODE defined by a vector field, f. The following optimization problem enforces the Lyapunov conditions of Prop.8.1 while minimizing the distance between the minimal attractor A^* and the 1-sublevel set of the Lyapunov function:

$$\begin{aligned} &\inf_{J \in \mathcal{F}} D_V(A^*, \{x \in \Omega : J(x) \leq 1\}) \\ &\text{such that } \nabla J(x)^T f(x) \leq -(J(x) - 1) \text{ for all } x \in \Omega, \\ &\qquad \{x \in \Omega : J(x) \leq 1\} \subseteq \Omega^\circ, \\ &\qquad \{x \in \Omega : J(x) \leq 1\} \neq \emptyset, \end{aligned} \tag{8.44}$$

where $\mathcal{F}$ is some set of functions which we may take to be the set of SOS polynomials.

In Subsection 8.5.2, we will propose a SOS programming approach to solving Optimization Problem (8.44). Specifically, in Subsection 8.5.2, we propose a sequence of quasi-SOS programming problems, each involving volume minimization, and whose limit yields the minimal attractor set of the ODE defined by f. However, the SOS constraints in Subsection 8.5.2 do not enforce $\{x \in \Omega : J(x) \leq 1\} \neq \emptyset$ - thus reducing the computational complexity of the algorithm. We show that it is not necesary to enforce this constraint because as we will show next in Subsection 8.5.1, by selecting Ω sufficiently large and enforcing $\nabla J(x)^T f(x) \leq -(J(x)-1)$ for all $x \in \Omega$ it follows that J automatically satisfies $\{x \in \Omega : J(x) \leq 1\} \neq \emptyset$. Moreover, unlike Opt. (8.44), the objective function of our proposed quasi-SOS programming problem will not involve the unknown set A^*. This is because, as we will next show in Subsection 8.5.1, for sufficiently large Ω and J such that $\nabla J(x)^T f(x) \leq -(J(x)-1)$ for all $x \in \Omega$ it follows $A^* \subseteq \{x \in \Omega : J(x) \leq 1\}$ (the 1-sublevel set of J contains the minimal attractor set). We later use result in Subsection 8.5.2 to eliminate A^* from the objective function.

Note in addition, we will show in Subsection 8.5.4 that if sublevel set volume is minimized and Ω is sufficiently large, then we may likewise eliminate the constraint $\{x \in \Omega : J(x) \leq 1\} \subseteq \Omega^\circ -$ thus further reducing computational complexity of the SOS programming problem.

8.5.1 A Reduced Form of Optimization Problem (8.44)

In Prop. 8.1 we have proposed a Lyapunov characterization of attractor sets. We have shown that if V satisfies Eqs. (8.12), (8.13) and (8.14) then the 1-sublevel set of V is an attractor set of the ODE defined by the vector field f. In Eq. (8.44) we have proposed an optimization problem that searches over functions J that satisfy Eqs. (8.12), (8.13) and (8.14) while minimizing the distance between the 1-sublevel set of J and the minimal attractor set of the ODE defined by f.

Later, in Subsection 8.5.2 we will propose an SOS programming problem for solving Opt. (8.44) that searches for a J that satisfies Eqs. (8.12) and (8.13) while minimizing the volume of the 1-sublevel set of J, but does not directly enforce Eq. (8.14) - instead choosing Ω to be sufficiently large. Fortunately, as we will show next in Lemma 8.2 that if Ω is chosen sufficiently large such that $A \subseteq \Omega$, for some attractor set A of the ODE, then any continuous V satisfying Eq. (8.12) automatically satisfies Eq. (8.14). Lemma 8.2 then shows that if Ω contains the minimal attractor and V satisfies Eqs. (8.12) and (8.13), then the 1-sublevel set of V is an attractor set.

Lemma 8.2. *Consider $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. Suppose there exists an attractor set (Defn. 8.1) $A \subset \mathbb{R}^n$ of the ODE* (8.5) *defined by f, $V \in C^1(\mathbb{R}^n, [0,\infty))$, and a compact set $\Omega \subset \mathbb{R}^n$ such that*

$$\nabla V(x)^T f(x) \leq -(V(x)-1) \text{ for all } x \in \Omega, \tag{8.45}$$

$$A \subseteq \Omega, \tag{8.46}$$

then $\{x \in \Omega : V(x) \leq 1\} \neq \emptyset$.

Proof. In order to prove $\{x \in \Omega : V(x) \leq 1\} \neq \emptyset$ we show that $A \cap \{x \in \Omega : V(x) \leq 1\} \neq \emptyset$.

Suppose for contradiction that $A \cap \{x \in \Omega : V(x) \leq 1\} = \emptyset$. Then $V(y) > 1$ for all $y \in A$. Since $A \subseteq \Omega$ (by Eq. (8.46)) is an attractor set it is an invariant set. Therefore, $\phi_f(y,t) \in A \subseteq \Omega$ for all $t \geq 0$ and thus by Eq. (8.45) it follows that

$$\frac{d}{dt} V(\phi_f(y,t)) \leq -(V(\phi_f(y,t)) - 1) \text{ for all } (y,t) \in A \times [0,\infty).$$

Then, using Gronwall's inequality (Lem. C.2) we have that

$$V(\phi_f(y,t)) - 1 \leq e^{-t}(V(y)-1) \text{ for all } (y,t) \in A \times [0,\infty). \tag{8.47}$$

Let $c := \inf_{t \ge 0}\{V(\phi_f(y,t)) - 1\}$. We will now argue that $c > 0$. Using the fact that $\phi_f(y,t) \in A$ for all $t \ge 0$ it follows that $c = \inf_{t \ge 0}\{V(\phi_f(y,t)) - 1\} \ge \inf_{z \in A}\{V(z) - 1\}$. Then, since V is continuous and A is compact it follows by the extreme value theorem that there exists $z^* \in A$ such that $V(z^*) - 1 = \inf_{z \in A}\{V(z) - 1\} \ge c$. Since we have assumed $A \cap \{x \in \Omega : V(x) \le 1\} = \emptyset$ it follows that if $z^* \in A$ then $z^* \notin \{x \in \Omega : V(x) \le 1\}$ and hence $c \ge V(z^*) - 1 > 0$.

Now, by Eq. (8.47) and since $c > 0$ it follows that $0 < ce^t \le V(y) - 1$ for all $t \ge 0$ and $y \in A$ implying that V is unbounded over A, contradicting the continuity of V. Therefore it follows that $A \cap \{x \in \Omega : V(x) \le 1\} \ne \emptyset$ and hence $\{x \in \Omega : V(x) \le 1\} \ne \emptyset$. □

Later in Subsection 8.5.2 we will propose an optimization problem that has an objective function independent of the unknown set A^* (unlike Opt. 8.44). In order to formulate this optimization problem we require $A^* \subseteq \{x \in \Omega : V(x) \le 1\}$. Next, we show that if Ω contains a neighborhood of the minimal attractor, A^*, and V satisfies Eqs. (8.12), then the 1-sublevel set of V contains the minimal attractor set.

Lemma 8.3. *Consider $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. Suppose $A^* \subset \mathbb{R}^n$ is the minimal attractor set (Defn. 8.1) of the ODE* (8.5) *defined by f, $V \in C^1(\mathbb{R}^n, [0,\infty))$, $\sigma > 0$, and a compact set $\Omega \subset \mathbb{R}^n$ such that*

$$\nabla V(x)^T f(x) \le -(V(x) - 1) \textit{ for all } x \in \Omega, \tag{8.48}$$

$$B_\sigma(A^*) \subseteq \Omega, \tag{8.49}$$

then $A^* \subseteq \{x \in \Omega : V(x) \le 1\}$.

Proof. To show $A^* \subseteq \{x \in \Omega : V(x) \le 1\}$ we will show $A^* \cap \{x \in \Omega : V(x) \le 1\}$ is an attractor set. Then if $A^* \nsubseteq \{x \in \Omega : V(x) \le 1\}$ it follows that $A^* \cap \{x \in \Omega :$

$V(x) \leq 1\} \subset A^*$, that is there exists an attractor set that is a strict subset of A^*, contradicting the fact that A^* is the minimal attractor set.

To show $A^* \cap \{x \in \Omega : V(x) \leq 1\}$ is an attractor set we will split the remainder of the proof into three parts, showing $A^* \cap \{x \in \Omega : V(x) \leq 1\}$ satisfies the three properties of attractor sets in Defn. 8.1.

Proof $A^* \cap \{x \in \Omega : V(x) \leq 1\}$ is nonempty and compact: By the proof of Lemma 8.2 it follows that $A^* \cap \{x \in \Omega : V(x) \leq 1\} \neq \emptyset$. Moreover, since A^* is compact and Ω is compact, implying $\{x \in \Omega : V(x) \leq 1\} \subseteq \Omega$ is compact, it follows $\{x \in \Omega : V(x) \leq 1\}$ is compact.

Proof $A^* \cap \{x \in \Omega : V(x) \leq 1\}$ is invariant: Let $y \in A^* \cap \{x \in \Omega : V(x) \leq 1\}$ then $y \in A^*$ and $y \in \{x \in \Omega : V(x) \leq 1\}$. Since A^* is an attractor set it is invariant and therefore $\phi_f(y,t) \in A^*$ for all $t \geq 0$. In order to prove $A^* \cap \{x \in \Omega : V(x) \leq 1\}$ is invariant we must also show $\phi_f(y,t) \in \{x \in \Omega : V(x) \leq 1\}$ for all $t \geq 0$. For contradiction suppose there exists $T > 0$ such that $\phi_f(y,t) \notin \{x \in \Omega : V(x) \leq 1\}$. That is, $V(\phi_f(y,T)) > 1$.

Using the fact A^* is invariant and applying the Granwall Bellman Lemma to Eq. (8.48) we get,

$$V(\phi_f(y,t)) - 1 \leq e^{-t}(V(y) - 1) \text{ for all } (y,t) \in A^* \times [0,\infty).$$

Hence, if $V(\phi_f(y,T)) > 1$ we get that that $0 < V(\phi_f(y,T)) - 1 \leq e^{T}(V(y) - 1)$ implying $V(y) > 1$ contradicting the fact that $y \in \{x \in \Omega : V(x) \leq 1\}$. Thus we have shown that if $y \in A^* \cap \{x \in \Omega : V(x) \leq 1\}$ then $\phi_f(x,t) \in A^*$ for all $t \geq 0$ and $\phi_f(x,t) \in \{x \in \Omega : V(x) \leq 1\}$ implying $\phi_f(x,t) \in A^* \cap \{x \in \Omega : V(x) \leq 1\}$ for all $t \geq 0$, proving $A^* \cap \{x \in \Omega : V(x) \leq 1\}$ is an invariant set.

Proof $A^* \cap \{x \in \Omega : V(x) \leq 1\}$ has an attracting neighborhood: We now show that for all $y \in A^* \cap \{x \in \Omega : V(x) \leq 1\}$ there exists $\delta > 0$ such that for any $\varepsilon > 0$

there exists $T \geq 0$ for which

$$D(\phi_f(z,t), A^* \cap \{x \in \Omega : V(x) \leq 1\}) < \varepsilon \tag{8.50}$$
$$\text{for all } z \in B_\delta(y) \text{ and } t \geq T.$$

Let $y \in A^* \cap \{x \in \Omega : V(x) \leq 1\}$ then $y \in A^*$. Since A^* is an attractor set there exists $\delta > 0$ such that for any $0 < \varepsilon < \sigma$ there exists $T_1 \geq 0$ for which

$$D(\phi_f(z,t), A^*) < \varepsilon \text{ for all } z \in B_\delta(y) \text{ and } t \geq T_1. \tag{8.51}$$

Since $0 < \varepsilon < \sigma$ and $D(\phi_f(z,t), A^*) < \varepsilon$ for all $(z,t) \in B_\delta(y) \times [T_1, \infty)$, it follows by Eq. (8.49) that

$$\phi_f(z,t) \in \Omega \text{ for all } (z,t) \in B_\delta(y) \times [T_1, \infty). \tag{8.52}$$

Next, we will consider the cases $\Omega / B_\varepsilon(\{x \in \Omega : V(x) \leq 1\}) = \emptyset$ and $\Omega / B_\varepsilon(\{x \in \Omega : V(x) \leq 1\}) \neq \emptyset$ separately showing Eq. (8.50) holds for each case.

In the case $\Omega / B_\varepsilon(\{x \in \Omega : V(x) \leq 1\}) = \emptyset$ we get that $\Omega \subseteq B_\varepsilon(\{x \in \Omega : V(x) \leq 1\})$, and hence using this fact together with Eq. (8.52) it follows that,

$$D(\phi_f(z,t), \{x \in \Omega : V(x) \leq 1\}) < \varepsilon \text{ for all } (z,t) \in B_\delta(y) \times [T_1, \infty). \tag{8.53}$$

Now, Eqs. (8.51) and (8.53) imply

$$\begin{aligned} &D(\phi_f(z,t), A^* \cap \{x \in \Omega : V(x) \leq 1\}) \\ &\leq \max\{D(\phi_f(z,t), A^*), D(\phi_f(z,t), \{x \in \Omega : V(x) \leq 1\})\} < \varepsilon \\ &\qquad \text{for all } (z,t) \in B_\delta(y) \times [T_1, \infty). \end{aligned} \tag{8.54}$$

Thus Eq. (8.54) shows Eq. (8.50) in the case $\Omega / B_\varepsilon(\{x \in \Omega : V(x) \leq 1\}) = \emptyset$.

Next let us consider the case $\Omega / B_\varepsilon(\{x \in \Omega : V(x) \leq 1\}) \neq \emptyset$. By Eqs. (8.48) and (8.52), Gronwall's inequality (Lem. C.2), and the semi-group property of solution maps (Eq. (8.6)) we have that

$$V(\phi_f(z,T_1+t))-1 \leq e^{-t}(V(\phi_f(z,T_1))-1) \leq ae^{-t} \text{ for all } (z,t) \in B_\delta(y) \times [0,\infty), \tag{8.55}$$

where $a := \sup_{z \in B_\delta(y)} |V(\phi_f(z,T_1))-1| \geq 0$. Hence, it now follows for any $\eta > 0$ that

$$\phi_f(z,t+T_1) \in \{x \in \Omega : V(x) \leq 1+\eta\} \tag{8.56}$$
$$\text{for all } z \in B_\delta(y) \text{ and } t \geq \max\left\{0, \ln\left(\frac{a}{\eta}\right)\right\}.$$

Let $T_2 := T_1 + \max\left\{0, \ln\left(\frac{a}{\eta}\right)\right\}$. We now construct $\eta > 0$ such that $\{x \in \Omega : V(x) \leq 1+\eta\} \subseteq B_\varepsilon(\{x \in \Omega : V(x) \leq 1\})$. Then since $\phi_f(y,t) \in \{x \in \Omega : V(x) \leq 1+\eta\}$ for all $t \geq T_2$ (by Eq. (8.56)), it follows that $D(\{x \in \Omega : V(x) \leq 1\}, \phi_f(y,t)) < \varepsilon$ for all $t \geq T_2$.

Let $\eta \in (0,b)$ where
$b := \inf_{z \in \Omega/B_\varepsilon(\{x \in \Omega : V(x) \leq 1\})}(V(z)-1)$, where $\inf_{z \in \Omega/B_\varepsilon(\{x \in \Omega : V(x) \leq 1\})} V(z)$ exists since $\Omega/B_\varepsilon(\{x \in \Omega : V(x) \leq 1\})$ is compact and V is continuous. Note that $b > 0$ since $\inf_{z \in \Omega/B_\varepsilon(\{x \in \Omega : V(x) \leq 1\})} V(z) - 1 > 0$ (because $\Omega/B_\varepsilon(\{x \in \Omega : V(x) \leq 1\})$ is compact so by the extreme value theorem there exists $z^* \in \Omega/B_\varepsilon(\{x \in \Omega : V(x) \leq 1\})$ such that $V(z^*) = \inf_{z \in \Omega/B_\varepsilon(\{x \in \Omega : V(x) \leq 1\})} V(z)$ and since $z^* \notin \{x \in \Omega : V(x) \leq 1\}$ it follows that $V(z^*) > 1$).

We now claim that $\{x \in \Omega : V(x) \leq 1+\eta\} \subseteq B_\varepsilon(\{x \in \Omega : V(x) \leq 1\})$. Suppose for contradiction that $\{x \in \Omega : V(x) \leq 1+\eta\} \nsubseteq B_\varepsilon(\{x \in \Omega : V(x) \leq 1\})$. Then there exists $w \in \{x \in \Omega : V(x) \leq 1+\eta\} \subseteq \Omega$ such that $w \notin B_\varepsilon(\{x \in \Omega : V(x) \leq 1\})$ implying $w \in \Omega/B_\varepsilon(\{x \in \Omega : V(x) \leq 1\})$. Now, $V(w) \leq \eta + 1 < \inf_{z \in \Omega/B_\varepsilon(\{x \in \Omega : V(x) \leq 1\})}\{V(z)\} \leq V(w)$ implying $0 < 0$, providing a contradiction.

Therefore, taking $t \geq T_2$ it follows from Eq. (8.56) that $\phi_f(z,t) \in \{x \in \Omega : V(x) \leq$

$1+\eta\} \subseteq B_\varepsilon(\{x \in \Omega : V(x) \leq 1\})$ for all $(z,t) \in B_\delta(y) \times [T_2, \infty)$ implying,

$$D(\phi_f(z,t), \{x \in \Omega : V(x) \leq 1\}) < \varepsilon \text{ for all } (z,t) \in B_\delta(y) \times [T_2,\infty). \tag{8.57}$$

Now, Eqs. (8.51) and (8.57) it follows that

$$\begin{aligned} &D(\phi_f(z,t), A^* \cap \{x \in \Omega : V(x) \leq 1\}) \\ &\leq \max\{D(\phi_f(z,t), A^*), D(\phi_f(z,t), \{x \in \Omega : V(x) \leq 1\})\} < \varepsilon \\ &\qquad \text{for all } (z,t) \in B_\delta(y) \times [T_2,\infty). \end{aligned} \tag{8.58}$$

Therefore Eqs. (8.54) and (8.58) prove Eq. (8.50).

□

We now propose an SOS optimization problem for enforcing the constraints of Optimization Problem (8.44).

8.5.2 An SOS Representation of the Lyapunov Inequality Constraint

Suppose $A^* \subset \mathbb{R}^n$ is the minimal attractor of some ODE (8.5) (defined by the vector field $f : \mathbb{R}^n \to \mathbb{R}^n$) and $\Omega \subset BOA_f(A^*)$ is some compact set such that $B_\sigma(A^*) \subseteq \Omega^\circ$, for some $\sigma > 0$. Let us consider the problem of approximating the minimal attractor A^* by some set A that can be certified as an attractor set (but not necessarily the minimal attractor set). One way to approach this problem is by solving Opt. (8.44), since any feasible solution, J, to Opt. (8.44) satisfies the Lyapunov conditions of Prop. 8.1, and hence, $A := \{x \in \Omega : J(x) \leq 1\}$ is an attractor set to the ODE defined by a vector field, f.

We now consider how to enforce the conditions of Opt. (8.44) using SOS optimization. Fortunately it is not necessary to enforce the constraint $\{x \in \Omega : J(x) \leq 1\} \neq \emptyset$ since Lem. 8.2 shows that J automatically satisfies this constraint when $A^* \subseteq \Omega$. We next propose a SOS tightening of the remaining constraints of Opt. (8.44), taking Ω to have the form $\Omega = \{x \in \mathbb{R}^n : g_\Omega(x) \geq 0\}$ so that $\partial\Omega = \{x \in \mathbb{R}^n : g_\Omega(x) = 0\}$.

For some $\alpha > 0$ we now consider the following optimization problem,

$$\begin{aligned} &\inf_{J \in \sum_{SOS}^d} D_V(A^*, \{x \in \Omega : J(x) \leq 1\}) \qquad (8.59)\\ &\text{such that } J, s_0, k_0, k_1 \in \sum_{SOS}^d, \quad p_0 \in \mathcal{P}_d(\mathbb{R}^n, \mathbb{R}),\\ &\text{where } k_0(x) = -\nabla J(x)^T f(x) - (J(x) - 1) - s_0(x) g_\Omega(x),\\ &k_1(x) = (J(x) - 1 - \alpha) - p_0(x) g_\Omega(x). \end{aligned}$$

The problem with solving Opt. (8.59) in its current form is that evaluating the objective function requires knowledge of the minimal attractor set, A^* (which is unknown). Fortunately, however, we can formulate an optimization problem which is equivalent to Opt. (8.59), but with an objective function that does not depend on the unknown minimal attractor set, A^*.

If $B_\sigma(A^*) \subseteq \Omega^\circ$, for some $\sigma > 0$ and J is feasible to Opt. (8.59), then Corollary 8.3 shows that minimizing $D_V(A^*, \{x \in \Omega : J(x) \leq 1\})$ is equivalent to minimizing $\mu(\{x \in \Omega : J(x) \leq 1\})$. Roughly speaking, if J is feasible to Opt. (8.59) then J satisfies the constraints of Opt. (8.59). Hence $\nabla J(x)^T f(x) \leq -(\nabla J(x) - 1)$ for all $x \in \Omega$. Thus if $B_\sigma(A^*) \subseteq \Omega^\circ$, where $\sigma > 0$ and A^* is the minimal attractor of the ODE defined by f, it follows by Lemma 8.3 that $A^* \subseteq \{x \in \Omega : J(x) \leq 1\}$. Hence, by Lem. A.2 we have that $D_V(A^*, \{x \in \Omega : J(x) \leq 1\}) = \mu(\{x \in \Omega : J(x) \leq 1\}) - \mu(A^*)$. Now, $\mu(A^*)$ is a constant (since A^* is not a decision variable). Therefore minimizing $D_V(A^*, \{x \in \Omega : J(x) \leq 1\})$ is equivalent to minimizing $\mu(\{x \in \Omega : J(x) \leq 1\})$.

For some $\alpha > 0$, we now consider the following family of d-degree SOS problems,

$$\begin{aligned} & J_{d,\alpha} \in \arg\inf_J \mu(\{x \in \Omega : J(x) \leq 1\}) \qquad (8.60)\\ & J, s_0, k_0, k_1 \in \sum_{SOS}^d, \quad p_0 \in \mathcal{P}_d(\mathbb{R}^n, \mathbb{R})\\ & \text{where } k_0(x) = -\nabla J(x)^T f(x) - (J(x) - 1) - s_0(x) g_\Omega(x)\\ & k_1(x) = (J(x) - 1 - \alpha) - p_0(x) g_\Omega(x). \end{aligned}$$

We now show that for sufficiently small $\alpha > 0$ and "large" Ω our quasi-SOS optimization problem proposed in Opt. (8.60) is not conservative since for sufficiently large enough degree its solution yields an arbitrarily close approximation of the minimal attractor set (in the volume metric). Moreover, each solution to Opt. (8.60) yields an attractor set.

Corollary 8.3. *Consider $f \in \mathcal{P}(\mathbb{R}^n, \mathbb{R})$. Suppose $A^* \subset \mathbb{R}^n$ is a minimal attractor set to the ODE* (8.5) *defined by f, $\sigma > 0$, and $\Omega \subset \mathbb{R}^n$ is some compact set such that $B_\sigma(A^*) \subseteq \Omega$ and $\Omega \subset BOA_f(A^*)$, $\Omega = \{x \in \mathbb{R}^n : g_\Omega(x) \geq 0\}$, and $\partial\Omega = \{x \in \mathbb{R}^n : g_\Omega(x) = 0\}$, where $g_\Omega \in \mathcal{P}(\mathbb{R}^n, \mathbb{R})$. Suppose $\{J_{d,\alpha}\}_{d \in \mathbb{N}}$ is such that $J_{d,\alpha}$ solves the d-degree optimization problems given in Eq.* (8.60) *for $\alpha > 0$, then:*

1. *$\{x \in \Omega : J_{d,\alpha}(x) \leq 1\}$ is an attractor set for each $d \in \mathbb{N}$ and $\alpha > 0$.*
2. *$A^* \subseteq \{x \in \Omega : J_{d,\alpha}(x) \leq 1\}$ for each $d \in \mathbb{N}$ and $\alpha > 0$.*
3. *There exists $\beta > 0$ such that for any $\alpha \in (0, \beta)$ we have that $\lim_{d \to \infty} D_V(A^*, \{x \in \Omega : J_{d,\alpha}(x) \leq 1\}) = 0$.*

Proof. In order to prove Cor. 8.3 we will now split the remainder of the proof into three parts showing each of the three statements of Cor. 8.3.

Proof $\{x \in \Omega : J_{d,\alpha}(x) \leq 1\}$ is an attractor set: By Prop. 8.1 it follows that $\{x \in \Omega : J_{d,\alpha}(x) \leq 1\}$ is an attractor set if $J_{d,\alpha}$ satisfies Eqs. (8.12), (8.13) and (8.14).

Since $J_{d,\alpha}$ is assumed to feasible to Opt. (8.60) it follows that $J_{d,\alpha}$ satisfies the constraints of Opt. (8.60) and hence $J_{d,\alpha}$ trivially satisfies Eq. (8.12). Moreover by the constraints of Opt. (8.60) it follows that $J_{d,\alpha}(x) \geq 1+\alpha > 1$ for all $x \in \partial\Omega$ and hence $\{x \in \Omega : J_{d,\alpha}(x) \leq 1\} \subseteq \Omega^\circ$, implying $J_{d,\alpha}$ satisfies Eq. (8.13). Finally since $A^* \subseteq \Omega$ and $J_{d,\alpha}$ satisfies Eq. (8.12) it follows by Lem. 8.2 that $J_{d,\alpha}$ satisfies Eq. (8.14).

Proof $\underline{A^* \subseteq \{x \in \Omega : J_{d,\alpha}(x) \leq 1\}}$**:** Since $J_{d,\alpha}$ satisfies the constraints of Opt. (8.60) it follows that

$$\nabla J_{d,\alpha}(x)^T f(x) \leq -(\nabla J_{d,\alpha}(x) - 1) \text{ for all } x \in \Omega.$$

Since $B_\sigma(A^*) \subset \Omega$ and A^* is the minimal attractor set of the ODE defined by f it follows from Lemma 8.3 that $A^* \subseteq \{x \in \Omega : J_{d,\alpha}(x) \leq 1\}$.

Proof $\underline{\lim_{d\to\infty} D_V(A^*, \{x \in \Omega : J_{d,\alpha}(x) \leq 1\}) = 0}$**:** We show that there exists $\beta > 0$ such that for any $\alpha \in (0,\beta)$ and $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $D_V(A^*, \{x \in \Omega : J_{d,\alpha}(x) \leq 1\}) < \varepsilon$ for all $d \geq N$.

By Theorem 8.2 it follows that there exists $\beta > 0$ such that for $\varepsilon > 0$ there exists $N_1 \in \mathbb{N}$, and $P_m \in \sum_{SOS}^m(\mathbb{R}^n, \mathbb{R})$ such that

$$\nabla P_m(x)^T f(x) < -(P_m(x) - 1) \text{ for all } x \in \Omega \text{ and } m > N_1, \tag{8.61}$$

$$P_m(x) > 1 + \beta > 1 + \alpha \text{ for all } x \in \partial\Omega, \alpha \in (0,\beta), \text{ and } m > N_1, \tag{8.62}$$

$$A^* \subseteq \{x \in \Omega : P_m(x) \leq 1\} \text{ for all } m > N_1, \tag{8.63}$$

$$D_V(A^*, \{x \in \Omega : P_m(x) \leq 1\}) < \varepsilon \text{ for all } m \geq N_1. \tag{8.64}$$

For any $\alpha \in (0,\beta)$, by Eqs. (8.61) and (8.62) and Theorem C.5 there exists $s_0, s_1, s_2, s_3, s_4, s_5 \in \sum_{SOS}$ for each $m > N_1$ such that $-\nabla P_m(x)^T f(x) - (P_m(x) - 1) - s_0(x)g_\Omega(x) = s_1(x)$, and $(P_m(x) - 1 - \alpha) - s_2(x)g_\Omega(x) = s_3(x)$, and $(P_m(x) - 1 - \alpha) + s_4(x)g_\Omega(x) = s_5(x)$. Fix $m > N_1$ and let $N_2 := \max\{m, \max_{0\leq i\leq 5} \deg(s_i)\}$. Then it follows that P_m is feasible to Opt. (8.60) for degree $d \geq N_2$ (with $p_0(x) :=$

$0.5(s_4(x) - s_2(x)))$. Since, $J_{d,\alpha}$ solves the Opt. (8.60) and P_m is feasible to Opt. (8.60) it follows that,

$$\mu(\{x \in \Omega : J_{d,\alpha}(x) \leq 1\}) \leq \mu(\{x \in \Omega : P_m(x) \leq 1\}) \text{ for all } d \geq N_2. \tag{8.65}$$

Hence, using Lemma A.2 along with the fact that $A^* \subseteq \{x \in \Omega : J_{d,\alpha}(x) \leq 1\}$ (by Lem. 8.3) and Eqs. (8.63), (8.64), and (8.65), it follows that,

$$\begin{aligned} D_V(A^*, \{x \in \Omega : J_{d,\alpha}(x) \leq 1\}) &= \mu(\{x \in \Omega : J_{d,\alpha}(x) \leq 1\}) - \mu(A^*) \\ &\leq \mu(\{x \in \Omega : P_m(x) \leq 1\}) - \mu(A^*) \\ &= D_V(A^*, \{x \in \Omega : P_m(x) \leq 1\}) \\ &< \varepsilon \text{ for all } d \geq N_2. \end{aligned}$$

□

8.5.3 Heuristic Volume Minimization of Sublevel Sets of SOS Polynomials

Unfortunately, it is still not possible for us to solve the family of d-degree optimization problems given in Eq. (8.60) since there is no known convex closed form analytical expression for the objective function (the volume of a sublevel set of an SOS polynomial). To make the problem tractable we replace the objective function in Eq. (8.60) with a convex objective function based on the determinant. We next show present two convex candidate objective functions based on the determinant.

Lemma 8.4. *The functions $f_1 : S^n_{++} \to \mathbb{R}$ and $f_2 : S^n_{++} \to \mathbb{R}$ defined as,*

$$\begin{aligned} f_1(X) &= -\log\det(P), \\ f_2(X) &= -(\det(P))^{\frac{1}{n}}, \end{aligned}$$

are convex.

Heuristically, maximizing $\det(P)$ increases the value of $V(x) = Z_d(x)^T P Z_d(x)$ for all $x \in \mathbb{R}^n$. Therefore, for larger $\det(P)$ there will be less $y \in \mathbb{R}^n$ such that $y \in \{x \in \mathbb{R}^n : V(x) \leq 1\}$. Hence we would expect $\mu(\{x \in \mathbb{R}^n : V(x) \leq 1\})$ to decrease as $\det(P)$ increases. In the 2-degree (quadratic) case this argument is not heuristic. We next show that maximizing the determinant is equivalent to minimizing the volume of the sublevel set of a quadratic polynomial.

Lemma 8.5 (Jones and Peet (2019c)). *Consider $P \in S^n_{++}$. The following holds,*

$$\mu(\{x \in \mathbb{R}^n : x^T P x \leq 1\}) = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)\sqrt{\det(P)}},$$

where Γ is the gamma function.

Lemma 8.5 shows that maximizing $\det(P)$ minimizes $\mu(\{x \in \mathbb{R}^n : x^T P x \leq 1\})$. Thus equivalently, maximizing the convex functions $\log\det(P)$ or $(\det(P))^{\frac{1}{N_d}}$ minimizes $\mu(\{x \in \mathbb{R}^n : x^T P x \leq 1\})$ (since both the functions $f_1(x) = \log(x)$ and $f_2(x) = x^{\frac{1}{n}}$ are monotonic functions for $x > 0$). We next extend this approach of maximizing the determinant to minimize the volume of a sublevel set of a SOS polynomial to higher degrees.

Rather than solving Opt. (8.60) we solve the following family d-degree SOS problems for some $\alpha > 0$,

$$\begin{aligned} & P_d \in \arg\sup_{J} (\det P)^{\frac{1}{N_d}} && (8.66)\\ & \quad J, s_0, k_0, k_1 \in \sum_{SOS}^{d}, \quad p_0 \in \mathcal{P}_d(\mathbb{R}^n, \mathbb{R}) \\ & \quad \text{where, } J = Z_d(x)^T P Z_d(x), \text{ and,} \\ & \quad P > 0, \\ & \quad k_0(x) = -\nabla J(x)^T f(x) - (J(x) - 1) - s_0(x) g_\Omega(x) \\ & \quad k_1(x) = (J(x) - 1 - \alpha) - p_0(x) g_\Omega(x). \end{aligned}$$

Note that it is equivalent to solve Opt. (8.66) with an objective function of form $(\det P)^{\frac{1}{N_d}}$ or $\log\det P$. For implementation purposes we have chosen to use an objective function of form $(\det P)^{\frac{1}{N_d}}$ since Yalmip allows this formulation of the problem to be solved by various SDP solves, see Lofberg (2004). For an objective function of form $\log\det P$ we use SOSTOOLS, see Prajna *et al.* (2002b), and SDPT3, see Tutuncu *et al.* (2002).

8.5.4 A Further Simplification of Optimization Problem (8.66)

Typically, through numerical experimentation, we find that if sublevel set volume of $\{x \in \Omega : J(x) \le 1\}$ is sufficiently minimized and Ω is sufficiently large, then $\{x \in \Omega : J(x) \le 1\} \subseteq \Omega^\circ$ is automatically satisfied. Therefore, it is often unnecessary to enforce the constraint $(J(x) - 1 - \alpha) - p_0(x)g_\Omega(x) \in \sum_{SOS}^d$ in Opt. (8.66) – thus further reducing computational complexity of the SOS programming problem.

8.6 Numerical Examples

In this section we will present the results of solving the Opt. (8.66) for several dynamical systems. For these examples Opt. (8.66) was solved using Yalmip, see Lofberg (2004). In Example 8.1 we approximate a "strange" attractor, in Example 8.2 we approximate a limit cycle, and finally in Example 8.3 we approximate an equilibrium point.

Example 8.1 (Numerical Approximation of the Lorenz Attractor)**.** *Consider the following three dimensional second order nonlinear dynamical system (known as the*

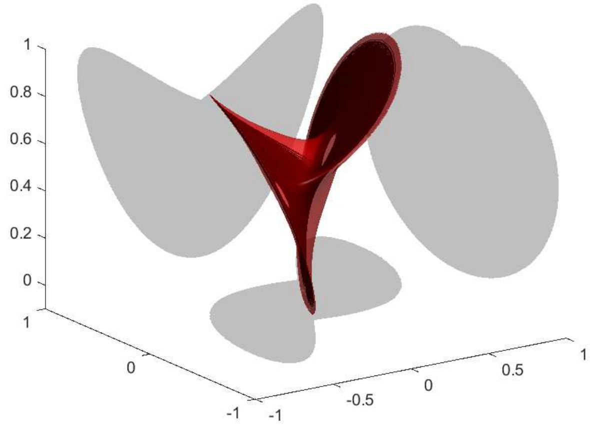

Figure 8.1: Graph showing an estimation of the Lorenz attractor (Example 8.1) given by the red transparent surface. This surface is the 1-sublevel set of a solution to the SOS Problem (8.66). The grayed shaded surfaces represent the projection of our Lorenz attractor estimation on the xy, xz, and yz axes. The black line is an approximation of the attractor found by simulating a Lorenz trajectory using Matlab's `ODE45` function.

Lorenz system):

$$\dot{x}_1(t) = \sigma(x_2(t) - x_1(t)), \tag{8.67}$$

$$\dot{x}_2(t) = \rho x_1(t) - x_2(t) - x_1(t)x_3(t),$$

$$\dot{x}_3(t) = x_1(t)x_2(t) - \beta x_3(t),$$

where $(\sigma, \rho, \beta) = (10, 28, \frac{8}{3})$*. It is well known that for such* (σ, ρ, β) *the ODE* (8.67) *exhibits a global "chaotic" attractor.*

Fig. 8.1 shows our Lorenz attractor approximation given by the 1*-sublevel of the solution to the SOS Problem* (8.66) *for* $d = 8$, $\alpha = 0.0001$, $g_\Omega(x) = R^2 - x_1^2 - x_2^2 - x_3^2$, $R = 3$, *and scaled dynamics given by the ODE* (8.67)*. For* $d \geq 10$ *the volume of our Lorenz attractor approximation becomes so small that we are unable store enough grid-points to sufficiently plot the contour of the* 1*-sublevel set of our SOS Lyapunov function.*

Example 8.2 (Numerical approximation of the Van der Poll oscillator)**.** *Consider the following two dimensional third order nonlinear dynamical system:*

$$\dot{x_1}(t) = x_2(t), \tag{8.68}$$
$$\dot{x_2}(t) = (1 - x_1^2(t))x_2(t) - x_1(t).$$

It is well known that the ODE (8.5) *possess a limit cycle called the Van der Poll oscillator. Let us denote this limit cycle by $A^* \subset \mathbb{R}^n$. The ODE also possess an unstable equilibrium point at the origin (which is not an attractor set since the solution map initialized inside neighborhoods of the origin moves away from the origin towards the limit cycle). However, $\phi_f(0,t) = 0 \in \mathbb{R}^2$ for all $t \geq 0$, where ϕ_f is the solution map of the ODE* (8.5)*. Therefore, $0 \notin BOA_f(A^*)$. In order to apply Theorem 8.2 we require $\Omega \subset BOA_f(A^*)$. Hence, we must be careful to construct $\Omega = \{x \in \mathbb{R}^n : g_\Omega(x) \geq 0\}$ such that $0 \notin \Omega$.*

Fig. 8.2 shows our Van der Poll oscillator approximation given by the 1*-sublevel of the solution to the SOS Problem* (8.66) *for $d = 12$, $\alpha = 0.0001$, $g_1(x) = -(R_1^2 - x_1^2 - x_2^2)(R_2^2 - x_2^2 - x_2^2)$, $R_1 = 0.45$, $R_2 = 1$, and scaled dynamics given by the ODE* (8.68)*.*

Example 8.3. *Consider the following two dimensional seventh order nonlinear dynamical system:*

$$\begin{aligned} \dot{x_1}(t) &= -2x_2(t)(-x_1^4(t) + 2x_1^2(t)x_2^2(t) + x_2^4(t)) \\ &\quad - 2x_1(t)(x_1^2(t) + x_2^2(t))(x_1^4(t) + 2x_1^2(t)x_2^2(t) - x_2^4(t)), \\ \dot{x_2}(t) &= 2x_1(t)(x_1^4(t) + 2x_1^2(t)x_2^2(t) - x_2^4(t)) \\ &\quad - 2x_2(t)(x_1^2(t) + x_2^2(t))(-x_1^4(t) + 2x_1^2(t)x_2^2(t) + x_2^4(t)). \end{aligned} \tag{8.69}$$

It was shown in Ahmadi and El Khadir (2018) that $A^* = \{0\}$ is a global attractor set of the ODE (8.69)*. In other words, the ODE* (8.69) *is globally asymptotically*

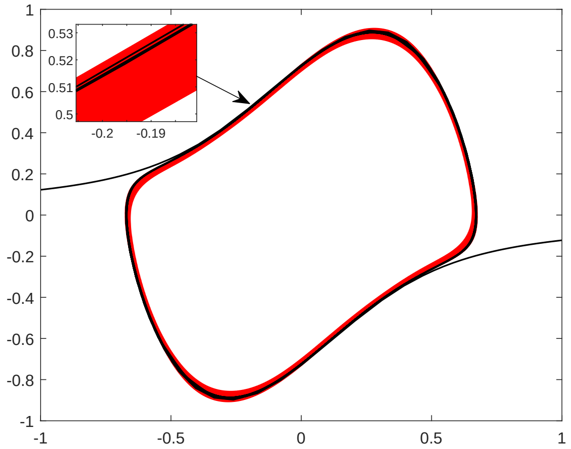

Figure 8.2: Graph showing an estimation of the attractor (given by the red area) of the ODE (8.68) in Example 8.2. This red area is the 1-sublevel set of a solution to the SOS Problem (8.66). The two black lines are simulated solution maps of the ODE (8.69) using Matlab's `ODE45` function initialized outside of the limit cycle.

stable about the origin. This stability was shown using the following non-polynomial Lyapunov function:

$$W(x) = \begin{cases} \frac{x_1^4+x_2^4}{x_1^2+x_2^2} \; if \; x \neq 0 \\ 0 \; otherwise \end{cases}.$$

Clearly, W is not a SOS polynomial (or even polynomial). It was further shown in Ahmadi and El Khadir (2018) that there exists no polynomial Lyapunov function that can certify the asymptotic stability of the origin of the ODE (8.69)*. However, Theorem 8.2 implies there does exist a SOS Lyapunov functions that can certify the stability of an arbitrarily small neighborhood of $A^* = \{0\}$ with respect the volume metric. Furthermore, we can heuristically attempt to find these Lyapunov functions by solving the SOS Opt.* (8.66)*.*

Fig. 8.3 shows our approximation of the ODE (8.69) *given by the* 1*-sublevel of the solution to the SOS Problem* (8.66) *for $d = 10$, $\alpha = 0.0001$, $g_\Omega(x) = R^2 - x_1^2 - x_2^2$, $R = 1$, and f as in the ODE* (8.69) *(scaled by a factor of 1000 to improve SDP solver*

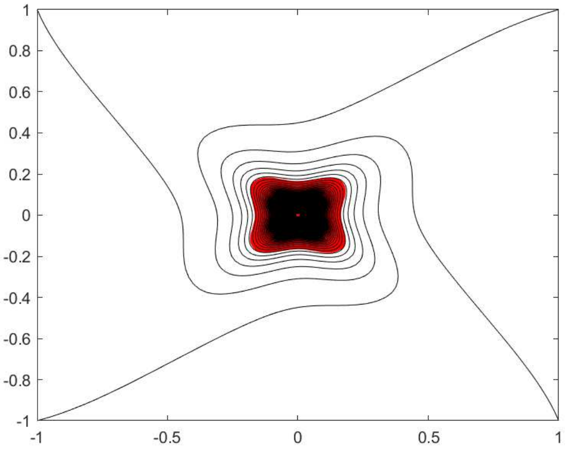

Figure 8.3: Graph showing an estimation of the attractor (given by the red area) of the ODE (8.69) in Example 8.3. This red area is the 1-sublevel set of a solution to the SOS Problem (8.66). The four black lines are simulated solution maps initialized at $(\pm 1, \pm 1)$ of the ODE (8.69) using Matlab's `ODE45` function.

performance). Unfortunately, increasing $d \in \mathbb{N}$ to a greater value than 10 *makes the SDP solver (Mosek) return a numerical error. We believe improvements in SDP solvers for large scale problems will allow us to solve the SOS Opt.* (8.66) *for larger degrees and improve our estimations of attractor sets.*

8.7 Conclusion

We have proposed a new Lyapunov characterization of attractor sets that is well suited to the problem of finding the minimal attractor set. We have shown that our proposed Lyapunov characterization of attractor sets is non-conservative even when restricted to SOS Lyapunov functions. Specifically, given an attractor set associated with some ODE we have shown that there exists a sequence of SOS Lyapunov functions that yield a sequence of sublevel sets, each containing the attractor set, each being an attractor set themselves, and converging to the attractor set in the volume metric. We have used this theoretical result to design an SOS based algorithm for minimal attractor set approximation based on determinant maximization as a proxy for sublevel set volume minimization. Several numerical examples demonstrate how our proposed SOS based algorithm can provide tight approximations of several well known attractor sets such as the Lorenz attractor and Van-der-Poll oscillator.

BIBLIOGRAPHY

Abu-Khalaf, M. and F. L. Lewis, “Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach”, Automatica **41**, 5, 779–791 (2005).

Achdou, Y., F. Camilli and I. Capuzzo Dolcetta, “Homogenization of Hamilton–Jacobi equations: numerical methods”, Mathematical models and methods in applied sciences **18**, 07, 1115–1143 (2008).

Ahbe, E., *Region of Attraction Analysis of Uncertain Equilibrium Points and Limit Cycles - with Application to Airborne Wind Energy Systems*, Ph.D. thesis, ETH Zurich, Zurich (2020).

Ahmadi, A., G. Hall, A. Makadia and V. Sindhwani, “Geometry of 3d environments and sum of squares polynomials”, arXiv (2017).

Ahmadi, A. A. and B. El Khadir, “A globally asymptotically stable polynomial vector field with rational coefficients and no local polynomial Lyapunov function”, Systems & Control Letters **121**, 50–53 (2018).

Ahmadi, A. A. and B. El Khadir, “On algebraic proofs of stability for homogeneous vector fields”, IEEE Transactions on Automatic Control **65**, 1, 325–332 (2019).

Ahmadi, A. A., M. Krstic and P. A. Parrilo, “A globally asymptotically stable polynomial vector field with no polynomial Lyapunov function”, in “Proceedings of the IEEE Conference on Decision and Control”, pp. 7579–7580 (IEEE, 2011).

Ahmadi, A. A. and A. Majumdar, “DSOS and SDSOS optimization: more tractable alternatives to sum of squares and semidefinite optimization”, SIAM Journal on Applied Algebra and Geometry **3**, 2, 193–230 (2019).

Altarovici, A., O. Bokanowski and H. Zidani, “A general Hamilton-Jacobi framework for non-linear state-constrained control problems”, ESAIM: Control, Optimisation and Calculus of Variations **19**, 2, 337–357 (2013).

Anderson, J. and A. Papachristodoulou, “Advances in computational Lyapunov analysis using Sum-of-Squares programming”, Discrete & Continuous Dynamical Systems-B **20**, 8, 2361 (2015).

Awrejcewicz, J., D. Bilichenko, A. K. Cheib, N. Losyeva and V. Puzyrov, “Estimating the region of attraction based on a polynomial Lyapunov function", Applied Mathematical Modelling **90**, 1143–1152 (2021).

Baldi, S., P. A. Ioannou and E. B. Kosmatopoulos, "A scalable iterative convex design for nonlinear systems", in "2012 American Control Conference (ACC)", pp. 979–984 (IEEE, 2012).

Baldi, S., G. Valmorbida, A. Papachristodoulou and E. B. Kosmatopoulos, "Piecewise polynomial policy iterations for synthesis of optimal control laws in input-saturated systems", in "2015 American Control Conference (ACC)", pp. 2850–2855 (IEEE, 2015).

Bäuerle, N. and U. Rieder, "More risk-sensitive markov decision processes", Mathematics of Operations Research **39**, 1, 105–120 (2013).

Bellman, R., "Dynamic programming", Science **153**, 3731, 34–37 (1966).

Bertsekas, D. P., *Dynamic programming and optimal control*, vol. 1 (Athena scientific Belmont, MA, 1995).

Bertsekas, D. P., *Abstract dynamic programming* (Athena Scientific, 2018).

Bertsekas, D. P., *Reinforcement learning and optimal control* (Athena Scientific Belmont, MA, 2019).

Bertsekas, D. P. and J. N. Tsitsiklis, "Neuro-dynamic programming: an overview", in "Proceedings of 1995 34th IEEE Conference on Decision and Control", vol. 1, pp. 560–564 (IEEE, 1995).

Braun, J. E. and K.-H. Lee, "Assessment of demand limiting using building thermal mass in small commercial buildings", ASHRAE Transactions **112**, 1, 547–558 (2006).

Bressan, A., "Viscosity solutions of Hamilton-Jacobi equations and optimal control problems", (2011).

Cai, J., J. Braun, D. Kim and J. Hu, "General approaches for determining the savings potential of optimal control for cooling in commercial buildings having both energy and demand charges", Science and Technology for the Built Environment **22**, 6, 733–750 (2016).

Chakraborty, A., P. Seiler and G. J. Balas, "Nonlinear region of attraction analysis for flight control verification and validation", Control Engineering Practice **19**, 4, 335–345 (2011a).

Chakraborty, A., P. Seiler and G. J. Balas, "Susceptibility of F/A-18 flight controllers to the falling-leaf mode: Nonlinear analysis", Journal of guidance, control, and dynamics **34**, 1, 73–85 (2011b).

Chen, Y. and A. D. Ames, "Duality between density function and value function with applications in constrained optimal control and markov decision process", arXiv preprint arXiv:1902.09583 (2019).

Colbert, B. K. and M. M. Peet, "Using trajectory measurements to estimate the region of attraction of nonlinear systems", in "2018 IEEE Conference on Decision and Control (CDC)", pp. 2341–2347 (IEEE, 2018).

Conti, J. J., "Annual energy outlook 2014 with projections to 2040", US Energy Information Administration (EIA), Independent statistics & Analysis (2014).

Cowlagi, R. V. and P. Tsiotras, "Hierarchical motion planning with dynamical feasibility guarantees for mobile robotic vehicles", IEEE Transactions on Robotics **28**, 2, 379–395 (2011).

Crandall, M. G., "Viscosity solutions: a primer", in "Viscosity solutions and applications", pp. 1–43 (Springer, 1997).

Cunis, T., J.-P. Condomines and L. Burlion, "Sum-of-squares flight control synthesis for deep-stall recovery", Journal of Guidance, Control, and Dynamics pp. 1–14 (2020).

Cuomo, K. M., A. V. Oppenheim and S. H. Strogatz, "Synchronization of lorenz-based chaotic circuits with applications to communications", IEEE Transactions on circuits and systems II: Analog and digital signal processing **40**, 10, 626–633 (1993).

Dabbene, F., D. Henrion and C. Lagoa, "Simple approximations of semialgebraic sets and their application to control", Automatica, Elsevier **78**, 110–118 (2017).

Dadebo, S., K. McAuley and P. McLellan, "On the computation of optimal singular and bang–bang controls", Optimal Control Applications and Methods **19**, 4, 287–297 (1998).

Domingo, A. and M. Sniedovich, “Experiments with dynamic programming algorithms for nonseparable problems”, European Journal of Operational Research **67**, 172–187 (1993).

Dreyfus, S. E., “An appraisal of some shortest-path algorithms”, Operations research **17**, 3, 395–412 (1969).

Dufour, F. and T. Prieto-Rumeau, “Approximation of markov decision processes with general state space”, J. Math. Anal. Appl. **388**, 1254–1267 (2012).

Dunn, B., H. Kamath and J.-M. Tarascon, “Electrical energy storage for the grid: A battery of choices”, Science **334**, 6058, 928–935 (2011).

Esterhuizen, W., T. Aschenbruck and S. Streif, “On maximal robust positively invariant sets in constrained nonlinear systems”, arXiv preprint arXiv:1904.01985 (2019).

Evans, L. C., *Partial differential equations*, vol. 19 (American Mathematical Soc., 2010).

Farhangi, H., “The path of the smart grid”, IEEE Power and Energy Magazine **8**, 1, 18–28 (2010).

Fischer, D., “Essential supremum with the continuous function?”, Mathematics Stack Exchange, uRL:https://math.stackexchange.com/q/618126 (version: 2015-06-09) (2015).

Frankowska, H. and M. Mazzola, “Discontinuous solutions of Hamilton–Jacobi–Bellman equation under state constraints”, Calculus of Variations and Partial Differential Equations **46**, 3-4, 725–747 (2013).

Frankowska, H. and R. Vinter, “Existence of neighboring feasible trajectories: applications to dynamic programming for state-constrained optimal control problems”, Journal of Optimization Theory and Applications **104**, 1, 20–40 (2000).

Gallistl, D., T. Sprekeler and E. Süli, “Mixed finite element approximation of periodic Hamilton–Jacobi–Bellman problems with application to numerical homogenization”, arXiv preprint arXiv:2010.01647 (2020).

Gallo, G. and S. Pallottino, “Shortest path algorithms”, Annals of operations research **13**, 1, 1–79 (1988).

Gao, W., L. Yan, M. Saeedi and H. S. Nik, “Ultimate bound estimation set and chaos synchronization for a financial risk system”, Mathematics and Computers in Simulation **154**, 19–33 (2018).

Gill, P. E., W. Murray and M. A. Saunders, “Snopt: An sqp algorithm for large-scale constrained optimization”, SIAM review **47**, 1, 99–131 (2005).

Glover, K. and J. C. Doyle, “State-space formulae for all stabilizing controllers that satisfy an h_∞-norm bound and relations to relations to risk sensitivity”, Systems & control letters **11**, 3, 167–172 (1988).

Goluskin, D., “Bounding extrema over global attractors using polynomial optimisation”, Nonlinearity **33**, 9, 4878 (2020).

Hahn, W., *The Direct Method of Liapunov*, pp. 93–165 (Springer Berlin Heidelberg, Berlin, Heidelberg, 1967).

Hirsch, M., S. Smale and R. Devaney, *Differential Equations, Dynamical Systems and An Introduction To Choas* (2004).

Huang, J., Q. Chang, J. Zou and J. Arinez, “A real-time maintenance policy for multistage manufacturing systems considering imperfect maintenance effects”, IEEE Access **6**, 62174–62183 (2018).

Jacobson, D., “Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games”, IEEE Transactions on Automatic control **18**, 2, 124–131 (1973).

Jacobson, D., S. Gershwin and M. Lele, “Computation of optimal singular controls”, IEEE Transactions on Automatic Control **15**, 1, 67–73 (1970).

Jennawasin, T., M. Kawanishi and T. Narikiyo, “Performance bounds for optimal control of polynomial systems: A convex optimization approach”, SICE Journal of Control, Measurement, and System Integration **4**, 6, 423–429 (2011).

Jiang, Y. and Z.-P. Jiang, “Global adaptive dynamic programming for continuous-time nonlinear systems”, IEEE Transactions on Automatic Control **60**, 11, 2917–2929 (2015).

Jones, M., H. Mohammadi and M. M. Peet, “Estimating the region of attraction using polynomial optimization: A converse Lyapunov result”, in “Proceedings of the IEEE Conference on Decision and Control", pp. 1796–1802 (2017).

Jones, M. and M. M. Peet, "Solving dynamic programming with supremum terms in the objective and application to optimal battery scheduling for electricity consumers subject to demand charges", in "Conference on Decision and Control (CDC)", pp. 1323–1329 (IEEE, 2017).

Jones, M. and M. M. Peet, "A dynamic programming approach to evaluating multivariate gaussian probabilities", in "The 23rd International Symposium on Mathematical Theory of Networks and Systems (MTNS)", (2018).

Jones, M. and M. M. Peet, "GPU accelerated 3D dynamic programming path planning and obstacle avoidance", https://codeocean.com/capsule/3639299/ (2019a).

Jones, M. and M. M. Peet, "Relaxing the hamilton jacobi bellman equation to construct inner and outer bounds on reachable sets", in "2019 IEEE 58th Annual Conference on Decision and Control (CDC)", (2019b).

Jones, M. and M. M. Peet, "Using SOS for optimal semialgebraic representation of sets: Finding minimal representations of limit cycles, chaotic attractors and unions", in "Proceedings of the IEEE American Control Conference (ACC)", pp. 2084–2091 (IEEE, 2019c).

Jones, M. and M. M. Peet, "Path planning animation", ResearchGate DOI:10.13140/RG.2.2.20466.32968 (2020).

Jones, M. and M. M. Peet, "Converse Lyapunov functions and converging inner approximations to maximal regions of attraction of nonlinear systems", in "2021 IEEE 60th Annual Conference on Decision and Control (CDC)", (2021a).

Jones, M. and M. M. Peet, "A converse sum-of-squares Lyapunov function for outer approximation of minimal attractor sets of nonlinear systems", arXiv preprint arXiv:2110.03093 (2021b).

Jones, M. and M. M. Peet, "Extensions of the dynamic programming framework: Battery scheduling, demand charges, and renewable integration", IEEE Transactions on Automatic Control **66**, 4, 1602–1617 (2021c).

Jones, M. and M. M. Peet, "A generalization of Bellman's Equation with application to path planning, obstacle avoidance and invariant set estimation", Automatica **127**, 109510 (2021d).

Jones, M. and M. M. Peet, “Polynomial approximation of value functions and nonlinear controller design with performance bounds”, arXiv preprint arXiv:2010.06828 (2021e).

Kalise, D. and K. Kunisch, “Polynomial approximation of high-dimensional Hamilton–Jacobi–Bellman equations and applications to feedback control of semilinear parabolic PDEs”, SIAM Journal on Scientific Computing **40**, 2, A629–A652 (2018).

Kamoutsi, A., T. Sutter, P. M. Esfahani and J. Lygeros, “On infinite linear programming and the moment approach to deterministic infinite horizon discounted optimal control problems”, IEEE control systems letters **1**, 1, 134–139 (2017).

Kamyar, R. and M. M. Peet, “Optimal thermostat programming and optimal electricity rates for customers with demand charges”, in “American Control Conference (ACC), 2015”, pp. 4529–4535 (IEEE, 2015).

Kamyar, R. and M. M. Peet, “Multi-objective dynamic programming for constrained optimization of non-separable objective functions with application in energy storage”, in “Decision and Control (CDC), 2016 IEEE 55th Conference on”, pp. 5348–5353 (IEEE, 2016).

Kang, W. and L. C. Wilcox, “Mitigating the curse of dimensionality: sparse grid characteristics method for optimal feedback control and HJB equations”, Computational Optimization and Applications **68**, 2, 289–315 (2017).

Kantner, M. and T. Koprucki, “Beyond just “flattening the curve”: Optimal control of epidemics with purely non-pharmaceutical interventions”, Journal of Mathematics in Industry **10**, 1, 1–23 (2020).

Kellett, C. M., “Classical converse theorems in Lyapunov’s second method”, arXiv preprint arXiv:1502.04809 (2015).

Khalil, H., *Nonlinear Systems* (1996).

Khmelnitsky, E., “On an optimal control problem of train operation”, IEEE transactions on automatic control **45**, 7, 1257–1266 (2000).

Kirmani, F., A. Pal, A. Mudgal, A. Shrestha and A. Siddiqui, “Advantages and disadvantages of hydroelectric power plant”, International Journal of Innovative Science and Research Technology (2021).

Korda, M., D. Henrion and C. N. Jones, "Controller design and value function approximation for nonlinear dynamical systems", Automatica **67**, 54–66 (2016).

Ku, Y. and C. Chen, "Stability study of a third-order servomechanism with multiplicative feedback control", Transactions of the American Institute of Electrical Engineers, Part II: Applications and Industry **77**, 3, 131–136 (1958).

Kunisch, K., S. Volkwein and L. Xie, "HJB-POD-based feedback design for the optimal control of evolution problems", SIAM Journal on Applied Dynamical Systems **3**, 4, 701–722 (2004).

Kurzwel, J., "On the inversion of Ljapunov's second theorem on stability of motion", AMS Translations Series 2 **24**, 19–77 (1963).

Lakshmi, M. V., G. Fantuzzi, J. D. Fernández-Caballero, Y. Hwang and S. I. Chernyshenko, "Finding extremal periodic orbits with polynomial optimization, with application to a nine-mode model of shear flow", SIAM Journal on Applied Dynamical Systems **19**, 2, 763–787 (2020).

Lasserre, J. B., "Tractable approximations of sets defined with quantifiers", Mathematical Programming **151**, 2, 507–527 (2015).

Lasserre, J. B., "Volume of sublevel sets of homogeneous polynomials", SIAM Journal on Applied Algebra and Geometry **3**, 2, 372–389 (2019).

Lasz, P. H., "Geometric analysis", (2014).

Lee, E., "Fundamental limits of cyberphysical systems modeling", ACM Trans. Cyber-Phys (2016).

Leong, Y. P., M. B. Horowitz and J. W. Burdick, "Optimal controller synthesis for nonlinear dynamical systems", arXiv preprint arXiv: 1410.0405 (2014).

Leth, T., R. Wisniewski and C. Sloth, "On the existence of polynomial Lyapunov functions for rationally stable vector fields", in "Proceedings of the IEEE Conference on Decision and Control", pp. 4884–4889 (IEEE, 2017).

Li, D., "Multiple objectives and non-separability in stochastic dynamic programming", International Journal of Systems Science **21**, 5, 933–950 (1990).

Li, D. and Y. Y. Haimes, "The envelope approach for multiobjeetive optimization problems", IEEE Transactions on Systems, Man, and Cybernetics **17**, 6, 1026–1038 (1987).

Li, D. and Y. Y. Haimes, "Multilevel methodology for a class of non-separable optimization problems", International Journal of Systems Science **21**, 11, 2351–2360 (1990a).

Li, D. and Y. Y. Haimes, "New approach for nonseparable dynamic programming problems", Journal of Optimization Theory and Applications **64**, 2, 311–330 (1990b).

Li, D. and Y. Y. Haimes, "Extension of dynamic programming to nonseparable dynamic optimization problems", Computers & Mathematics with Applications **21**, 11-12, 51–56 (1991).

Li, D., J. Lu, X. Wu and G. Chen, "Estimating the bounds for the Lorenz family of chaotic systems", Chaos, Solitons and Fractals **23**, 529–534 (2005).

Liberzon, D., *Calculus of variations and optimal control theory: a concise introduction* (Princeton University Press, 2011).

Lin, Y., E. D. Sontag and Y. Wang, "A smooth converse Lyapunov theorem for robust stability", SIAM Journal on Control and Optimization **34**, 1, 124–160 (1996).

Liu, Q., M. Dong, W. Lv and C. Ye, "Manufacturing system maintenance based on dynamic programming model with prognostics information", Journal of Intelligent Manufacturing **30**, 3, 1155–1173 (2019).

Lofberg, J., "Yalmip: A toolbox for modeling and optimization in matlab", in "2004 IEEE international conference on robotics and automation (IEEE Cat. No. 04CH37508)", pp. 284–289 (IEEE, 2004).

Magnani, A., S. Lall and S. Boyd, "Tractable fitting with convex polynomials via sum of squares", CDC (2005).

Maidens, J., A. Barrau, S. Bonnabel and M. Arcak, "Symmetry reduction for dynamic programming", Automatica **97**, 367–375 (2018).

Maidens, J., A. Packard and M. Arcak, "Parallel dynamic programming for optimal experiment design in nonlinear systems", in "Conference on Decision and Control (CDC)", pp. 2894–2899 (IEEE, 2016).

Maly, D. and K. Kwan, “Optimal battery energy storage system (bess) charge scheduling with dynamic programming”, IEE Proceedings-Science, Measurement and Technology **142**, 6, 453–458 (1995).

Malỳ, J. and W. P. Ziemer, *Fine regularity of solutions of elliptic partial differential equations*, no. 51 (American Mathematical Soc., 1997).

Massera, J. L., “On Liapounoff’s conditions of stability”, Annals of Mathematics pp. 705–721 (1949).

McEneaney, W. M., “A curse-of-dimensionality-free numerical method for solution of certain HJB PDEs”, SIAM journal on Control and Optimization **46**, 4, 1239–1276 (2007).

Mitchell, I. M., A. M. Bayen and C. J. Tomlin, “A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games”, IEEE Transactions on automatic control **50**, 7, 947–957 (2005).

Mohammadi, H., M. Razaviyayn and M. R. Jovanović, “On the stability of gradient flow dynamics for a rank-one matrix approximation problem”, in “Proceedings of the IEEE American Control Conference (ACC)”, pp. 4533–4538 (IEEE, 2018).

Mohd, A., E. Ortjohann, A. Schmelter, N. Hamsic and D. Morton, “Challenges in integrating distributed energy storage systems into future smart grid”, in “IEEE International Symposium on Industrial Electronics”, pp. 1627–1632 (2008).

Moyalan, J., H. Choi, Y. Chen and U. Vaidya, “Sum of squares based convex approach for optimal control synthesis”, in “2021 29th Mediterranean Conference on Control and Automation (MED)”, pp. 1270–1275 (2021).

Pakniyat, A. and R. Vasudevan, “A convex duality approach to optimal control of killed markov processes”, CDC (2019).

Peet, M. M., “Exponentially stable nonlinear systems have polynomial Lyapunov functions on bounded regions”, IEEE Transactions on Automatic Control **54**, 5, 979–987 (2009).

Peet, M. M. and A. Papachristodoulou, “A converse sum-of-squares lyapunov result: An existence proof based on the picard iteration”, in “49th IEEE Conference on Decision and Control (CDC)”, pp. 5949–5954 (IEEE, 2010).

Penmetsa, V., *Climate Change Effects on Electricity Generation from Hydropower, Wind, Solar and Thermal Power Plants*, Ph.D. thesis, ARIZONA STATE UNIVERSITY (2020).

Powell, W. B., *Approximate Dynamic Programming: Solving the curses of dimensionality*, vol. 703 (John Wiley & Sons, 2007).

Prajna, S., A. Papachristodoulou and P. A. Parrilo, "Introducing SOSTOOLS: A general purpose sum of squares programming solver", in "Proceedings of the IEEE Conference on Decision and Control", vol. 1, pp. 741–746 (2002a).

Prajna, S., A. Papachristodoulou and P. A. Parrilo, "Introducing sostools: A general purpose sum of squares programming solver", in "Proceedings of the 41st IEEE Conference on Decision and Control, 2002.", vol. 1, pp. 741–746 (IEEE, 2002b).

Putinar, M., "Positive polynomials on compact semialgebriac sets.", Math J (1993).

Ribeiro, A. M., A. R. Fioravanti, A. Moutinho and E. C. de Paiva, "Control design based on sum of squares programming for non-affine in input systems", in "2020 IEEE 6th International Conference on Control Science and Systems Engineering (ICCSSE)", pp. 130–135 (IEEE, 2020).

Rippel, E., A. Bar-Gill and N. Shimkin, "Fast graph-search algorithms for general-aviation flight trajectory generation", Journal of Guidance, Control, and Dynamics **28**, 4, 801–811 (2005).

Ruszczyński, A., "Risk-averse dynamic programming for markov decision processes", Mathematical programming **125**, 2, 235–261 (2010).

Savkin, A. V. and M. Hoy, "Reactive and the shortest path navigation of a wheeled mobile robot in cluttered environments", Robotica **31**, 2, 323 (2013).

Schlosser, C. and M. Korda, "Converging outer approximations to global attractors using semidefinite programming", arXiv preprint arXiv:2005.03346 (2020).

Shapiro, A., "On a time consistency concept in risk averse multistage stochastic programming", Operations Research Letters **37**, 3, 143–147 (2009).

Shapiro, A. and K. Ugurlu, "Decomposability and time consistency of risk averse multistage programs", Operations Research Letters **44**, 5, 663–665 (2016).

Shear, T., "Today in energy: February archive", US Energy Information Administration (EIA), Independent statistics & Analysis (2014).

Sherwood, L., "U.S. solar market trends 2012", Tech. rep., Interstate Renewable Energy Council (2013).

Solar Energy Power Association, "SEPA comments on utility investments in distributed solar companies", press release (2013).

Soner, H. M., "Optimal control with state-space constraint i", SIAM Journal on Control and Optimization **24**, 3, 552–561 (1986).

Spivak, M., "Calculus on manifolds", (1965).

SRP, "Standard electric price plans", (November 2015).

Sturm, J. F., "Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones", Optimization methods and software **11**, 1-4, 625–653 (1999).

Summers, E., A. Chakraborty, W. Tan, U. Topcu, P. Seiler, G. Balas and A. Packard, "Quantitative local L2-gain and reachability analysis for nonlinear systems", International Journal of Robust and Nonlinear Control **23**, 10, 1115–1135 (2013).

Tan, W. and A. Packard, "Stability region analysis using polynomial and composite polynomial Lyapunov functions and Sum-of-Squares programming", IEEE Transactions on Automatic Control **53**, 2, 565–571 (2008).

Teel, A. R. and L. Praly, "A smooth Lyapunov function from a class-kl estimate involving two positive semidefinite functions", ESAIM: Control, Optimisation and Calculus of Variations **5**, 313–367 (2000).

Tutuncu, R., K. Toh and M. Todd, "Solving semidefinite-quadratic-linear programs using SDPT3", Springer-Verlag (2002).

Valmorbida, G. and J. Anderson, "Region of attraction estimation using invariant sets and rational Lyapunov functions", Automatica **75**, 37–45 (2017).

Vandenberghe, L. and S. Boyd, "Semidefinite programming", SIAM review **38**, 1, 49–95 (1996).

Vannelli, A. and M. Vidyasagar, “Maximal Lyapunov functions and domains of attraction for autonomous nonlinear systems”, Automatica **21**, 1, 69–80 (1985).

Wang, T., S. Lall and M. West, “Polynomial level-set method for attractpr estimation”, Journal of The Franklin Institute **349**, 2783–2798 (2012a).

Wang, T.-C., S. Lall and M. West, “Polynomial level-set method for attractor estimation”, Journal of the Franklin Institute **349**, 9, 2783–2798 (2012b).

Wang, Z., R. M. Jungers and C.-J. Ong, “Computation of the maximal invariant set of discrete-time systems subject to quasi-smooth non-convex constraints”, arXiv preprint arXiv:1912.09727 (2019).

Wilson, F. W., “Smoothing derivatives of functions and applications”, Transactions of the American Mathematical Society **139**, 413–428 (1969).

Xie, J., L. Jin and L. R. Garcia Carrillo, “Optimal path planning for unmanned aerial systems to cover multiple regions”, in “AIAA Scitech 2019 Forum”, p. 1794 (2019).

Xue, B., M. Fränzle and N. Zhan, “Inner-approximating reachable sets for polynomial systems with time-varying uncertainties”, IEEE Transactions on Automatic Control (2019).

Xue, B. and N. Zhan, “Robust invariant sets computation for switched discrete-time polynomial systems”, arXiv preprint arXiv:1811.11454 (2018).

Yin, H., A. Packard, M. Arcak and P. Seiler, “Reachability analysis using dissipation inequalities for nonlinear dynamical systems”, arXiv preprint arXiv:1808.02585 (2018).

Yoshizawa, T., “Stability theory by liapunov’s second method”, (1966).

Yu, P. and X. Liao, “Globally attractive and positive invariant set of the Lorenz system”, Internation Journal of Bifurcation and Chaos **16**, 757–764 (2005).

Zeinalzadeh, A. and V. Gupta, “Minimizing risk of load shedding and renewable energy curtailment in a microgrid with energy storage”, arXiv preprint arXiv:1611.08000 (2016).

Zeng, X. and J. Wang, “Globally energy-optimal speed planning for road vehicles on a given route”, Transportation Research Part C: Emerging Technologies **93**, 148–160

(2018).

Zhang, Y., M. E. Raoufat, K. Tomsovic and S. M. Djouadi, “Set theory-based safety supervisory control for wind turbines to ensure adequate frequency response”, IEEE Transactions on Power Systems **34**, 1, 680–692 (2019).

Zhao, P., S. Mohan and R. Vasudevan, “Control synthesis for nonlinear optimal control via convex relaxations”, in “2017 American Control Conference (ACC)”, pp. 2654–2661 (IEEE, 2017).

Zhao, Y., W. Zhang, H. Su and J. Yang, “Observer-based synchronization of chaotic systems satisfying incremental quadratic constraints and its application in secure communication”, IEEE Transactions on Systems, Man, and Cybernetics: Systems **50**, 12, 5221–5232 (2018).

Zheng, X., Z. She, J. Lu and M. Li, “Computing multiple Lyapunov-like functions for inner estimates of domains of attraction of switched hybrid systems”, International Journal of Robust and Nonlinear Control (2018).

Zheng, Y., A. Sootla and A. Papachristodoulou, “Block factor-width-two matrices and their applications to semidefinite and sum-of-squares optimization”, arXiv preprint arXiv:1909.11076 (2019).

Zhu, Y., D. Zhao, X. Yang and Q. Zhang, “Policy iteration for H_∞ optimal control of polynomial nonlinear systems via sum of squares programming”, IEEE transactions on cybernetics **48**, 2, 500–509 (2017).

Zubov, V. I., *Methods of AM Lyapunov and their application* (P. Noordhoff, 1964).

APPENDIX A

SUBLEVEL SET APPROXIMATION

For sets $A, B \subset \mathbb{R}^n$ we denote the volume metric as $D_V(A,B)$, where

$$D_V(A,B) := \mu((A/B) \cup (B/A)). \tag{A.1}$$

In this appendix we show that the volume metric (D_V in Eq. (A.1)) is indeed a metric. Given a sequence of functions, $\{J_d\}_{d \in \mathbb{N}}$, such that $J_d \to V$ as $d \to \infty$ with respect to some norm, we also present conditions under which the sequence of sublevel sets, $\{x \in \Omega : J_d(x) \le \gamma\}$ (or $\{x \in \Omega : J_d(x) < \gamma\}$) where $\gamma \in \mathbb{R}$, converges to $\{x \in \Omega : V(x) \le \gamma\}$ (or $\{x \in \Omega : V(x) < \gamma\}$) with respect to the volume metric. The sublevel approximation results presented in this appendix are required in the proof of Prop. 5.5, Theorem 7.2 and Cor. 8.3.

Definition A.1. *$D : X \times X \to \mathbb{R}$ is a metric if the following is satisfied for all $x, y \in X$,*

- *$D(x,y) \ge 0$,*
- *$D(x,y) = 0$ if and only if $x = y$,*
- *$D(x,y) = D(y,x)$,*
- *$D(x,z) \le D(x,y) + D(y,z)$.*

Lemma A.1 (Jones and Peet (2019c))**.** *Consider the quotient space,*

$$X := \mathcal{B} \pmod{\{X \subset \mathbb{R}^n : X \neq \emptyset, \mu(X) = 0\}},$$

recalling $\mathcal{B} := \{B \subset \mathbb{R}^n : \mu(B) < \infty\}$ is the set of all bounded sets. Then $D_V : X \times X \to \mathbb{R}$, defined in Eq. (A.1)*, is a metric.*

Lemma A.2 (Jones and Peet (2019c))**.** *If $A, B \in \mathcal{B}$ and $B \subseteq A$ then*

$$D_V(A,B) = \mu(A/B) = \mu(A) - \mu(B).$$

Inspired by an argument used in Lasserre (2015) we now show if two functions are close in the L^1 norm then it follows their sublevel sets are close with respect to the volume metric.

Proposition A.1. *Consider a set $\Lambda \in \mathcal{B}$, a function $V \in L^1(\Lambda, \mathbb{R})$, and a family of functions $\{J_d \in L^1(\Lambda, \mathbb{R}) : d \in \mathbb{N}\}$ that satisfies the following properties:*

1. *For any $d \in \mathbb{N}$ we have $J_d(x) \le V(x)$ for all $x \in \Lambda$.*
2. *$\lim_{d \to \infty} ||V - J_d||_{L^1(\Lambda, \mathbb{R})} = 0$.*

Then for all $\gamma \in \mathbb{R}$

$$\lim_{d \to \infty} D_V\bigg(\{x \in \Lambda : V(x) \le \gamma\}, \{x \in \Lambda : J_d(x) \le \gamma\}\bigg) = 0. \tag{A.2}$$

Proof. To prove Eq. (A.2) we show for all $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $d \geq N$

$$D_V\bigg(\{x \in \Lambda : V(x) \leq \gamma\}, \{x \in \Lambda : J_d(x) \leq \gamma\}\bigg) < \varepsilon. \tag{A.3}$$

In order to do this we first denote the following family of sets for each $n \in \mathbb{N}$

$$A_n := \bigg\{x \in \Lambda : V(x) \leq \gamma + \frac{1}{n}\bigg\}.$$

Since $J_d(x) \leq V(x)$ for all $x \in \Lambda$ and $d \in \mathbb{N}$ we have

$$\{x \in \Lambda : V(x) \leq \gamma\} \subseteq \{x \in \Lambda : J_d(x) \leq \gamma\} \text{ for all } d \in N. \tag{A.4}$$

Moreover, since $\{x \in \Lambda : V(x) \leq \gamma\} \subseteq \Lambda$, $\{x \in \Lambda : J_d(x) \leq \gamma\} \subseteq \Lambda$ and $\Lambda \in \mathcal{B}$ it follows $\{x \in \Lambda : V(x) \leq \gamma\} \in \mathcal{B}$ and $\{x \in \Lambda : J_d(x) \leq \gamma\} \in \mathcal{B}$.

Now for $d \in \mathbb{N}$

$$\begin{aligned} &D_V\bigg(\{x \in \Lambda : V(x) \leq \gamma\}, \{x \in \Lambda : J_d(x) \leq \gamma\}\bigg) \\ &= \mu(\{x \in \Lambda : J_d(x) \leq \gamma\}) - \mu(\{x \in \Lambda : V(x) \leq \gamma\}) \\ &= \mu(\{x \in \Lambda : J_d(x) \leq \gamma\}) - \mu(A_n \cap \{x \in \Lambda : J_d(x) \leq \gamma\}) \\ &\qquad + \mu(A_n \cap \{x \in \Lambda : J_d(x) \leq \gamma\}) - \mu(\{x \in \Lambda : V(x) \leq \gamma\}) \\ &\leq \mu(\{x \in \Lambda : J_d(x) \leq \gamma\}) - \mu(A_n \cap \{x \in \Lambda : J_d(x) \leq \gamma\}) \\ &\qquad + \mu(A_n) - \mu(\{x \in \Lambda : V(x) \leq \gamma\}) \\ &= \mu(\{x \in \Lambda : J_d(x) \leq \gamma\}/A_n) + \mu(A_n/\{x \in \Lambda : V(x) \leq \gamma\}). \end{aligned} \tag{A.5}$$

The first equality of Eq. (A.5) follows by Lemma A.2 (since the sublevel sets of V and J_d are bounded and satisfy Eq. (A.4)). The first inequality follows as $A_n \cap \{x \in \Lambda : J_d(x) \leq \gamma\} \subseteq A_n$ which implies $\mu(A_n \cap \{x \in \Lambda : J_d(x) \leq \gamma\}) \leq \mu(A_n)$. The third equality follows using Lemma A.2 and since $A_n \cap \{x \in \Lambda : J_d(x) \leq \gamma\} \subseteq \{x \in \Lambda : J_d(x) \leq \gamma\}$ and $\{x \in \Lambda : V(x) \leq \gamma\} \subseteq A_n$.

To show that Eq. (A.3) holds for any $\varepsilon > 0$ we will split the remainder of the proof into two parts. In Part 1 we show that there exists $N_1 \in \mathbb{N}$ such that $\mu(A_n/\{x \in \Lambda : V(x) \leq \gamma\}) < \frac{\varepsilon}{2}$ for all $n \geq N_1$. In Part 2 we show that for any $n \in \mathbb{N}$ there exists $N_2 \in \mathbb{N}$ such that $\mu(\{x \in \Lambda : J_d(x) \leq \gamma\}/A_n) < \frac{\varepsilon}{2}$ for all $d \geq N_2$.

Part 1 of proof: In this part of the proof we show that there exists $N_1 \in \mathbb{N}$ such that $\mu(A_n/\{x \in \Lambda : V(x) \leq \gamma\}) < \frac{\varepsilon}{2}$ for all $n > N_1$.

Since $\cap_{n=1}^{\infty} A_n = \{x \in \Lambda : V(x) \leq \gamma\}$ and $A_{n+1} \subseteq A_n$ for all $n \in \mathbb{N}$ we have that $\mu(\{x \in \Lambda : V(x) \leq \gamma\}) = \mu(\cap_{n=1}^{\infty} A_n) = \lim_{n \to \infty} \mu(A_n)$ (using the "continuity from above" property of measures). Thus there exists $N_1 \in \mathbb{N}$ such that

$$|\mu(A_n) - \mu(\{x \in \Lambda : V(x) \leq \gamma\})| < \frac{\varepsilon}{2} \text{ for all } n > N_1.$$

Therefore it follows

$$\begin{aligned} &\mu(A_n/\{x \in \Lambda : V(x) \leq \gamma\}) \\ &\qquad = \mu(A_n) - \mu(\{x \in \Lambda : V(x) \leq \gamma\}) < \frac{\varepsilon}{2} \text{ for all } n > N_1. \end{aligned}$$

Part 2 of proof: For fixed $n > N_1$ we now show there exists $N_2 \in \mathbb{N}$ such that $\mu(\{x \in \Lambda : J_d(x) \leq \gamma\}/A_n) < \frac{\varepsilon}{2}$ for all $d \geq N_2$.

Now

$$\{x \in \Lambda : J_d(x) \leq \gamma\}/A_n \subseteq \{x \in \Lambda : n|J_d(x) - V(x)| \geq 1\} \tag{A.6}$$
$$\text{for all } d \in \mathbb{N}.$$

The set containment in Eq. (A.6) follows since if $y \in \{x \in \Lambda : J_d(x) \leq \gamma\}/A_n$ then $y \in \Lambda$, $J_d(y) \leq \gamma$ and $y \notin A_n$. Since $y \notin A_n$ we have $V(y) > \gamma + \frac{1}{n}$. Thus

$$n|J_d(y) - V(y)| \geq n(V(y) - J_d(y)) \geq n\left(\gamma + \frac{1}{n} - \gamma\right) = 1,$$

which implies $y \in \{x \in \Lambda : n|J_d(x) - V(x)| \geq 1\}$.

Since $\lim_{d\to\infty} \int_\Lambda |V(x) - J_d(x)|dx = 0$ there exists $N_2 \in \mathbb{N}$ such that

$$\int_\Lambda |V(x) - J_d(x)|dx < \frac{\varepsilon}{2n} \text{ for all } d \geq N_2. \tag{A.7}$$

Therefore,

$$\begin{aligned}\mu(\{x \in \Lambda :& J_\delta(x) \leq \gamma\}/A_n) \leq \mu(\{x \in \Lambda : n|J_\delta(x) - V(x)| \geq 1\})\\ &\leq \int_\Lambda n|J_\delta(x) - V(x)|dx < \frac{\varepsilon}{2} \text{ for } d \geq N_2.\end{aligned} \tag{A.8}$$

The first inequality in Eq. (A.8) follows by Eq. (A.6). The second inequality follows by Chebyshev's inequality (Lemma C.5). The third inequality follows by Eq. (A.7). □

Prop. A.1 shows if a sequence of functions $\{J_d\}_{d\in\mathbb{N}}$ converges from bellow to some function V with respect to the L^1 norm then the sequence sublevel sets $\{x \in \Lambda : J_d(x) \leq \gamma\}$ converge to $\{x \in \Lambda : V(x) \leq \gamma\}$ with respect to the volume metric. However, this does not imply the sequence of "strict" sublevel sets $\{x \in \Lambda : J_d(x) < \gamma\}$ converge to $\{x \in \Lambda : V(x) < \gamma\}$ (even if $\{J_d\}_{d\in\mathbb{N}}$ converges from bellow to V with respect to the L^∞ norm). To see this we next consider a counterexample where $\{J_d\}_{d\in\mathbb{N}}$ is a family of functions that can uniformly approximate some given $V \in Lip((0,1),\mathbb{R})$ but $\{x \in \Lambda : J_d(x) < \gamma\}$ does not converge to $\{x \in \Lambda : V(x) < \gamma\}$.

Counterexample A.1. *We show there exists $\gamma \in \mathbb{R}$, $\Lambda \subset \mathbb{R}$, $V \in Lip(\Lambda,\mathbb{R})$ and $\{J_d\}_{d\in\mathbb{N}} \subset Lip(\Lambda,\mathbb{R})$ such that $J_d(x) \leq V(x)$ for all $x \in \Lambda$ and* $\lim_{d\to\infty} \int_\Lambda |V(x) - J_d(x)|dx = 0$ *but*

$$\lim_{d\to\infty} D_V\bigg(\{x \in \Lambda : V(x) < \gamma\}, \{x \in \Lambda : J_d(x) < \gamma\}\bigg) \neq 0$$

Let

$$\Lambda = (0,1), \quad V(x) = \begin{cases} 0 \textit{ if } x \in (0, 0.25] \\ 2(x - 0.25) \textit{ if } x \in (0.25, 0.75), \\ 1 \textit{ if } x \in [0.75, 1) \end{cases}$$

$$J_d(x) = \begin{cases} 0 \textit{ if } x \in (0, 0.25] \\ 2(x - 0.25) \textit{ if } x \in (0.25, 0.75), \\ 1 - \frac{1}{d} \textit{ if } x \in [0.75, 1) \end{cases} \quad \gamma = 1.$$

Now for all $d \in \mathbb{N}$ it is clear that we have $J_d(x) \leq V(x)$ and $V(x) - J_d(x) < \frac{1}{d}$ for all $x \in \Lambda$. This implies

$$\lim_{d \to \infty} \int_\Lambda V(x) - J_d(x) dx \leq \lim_{d \to \infty} \sup_{x \in \Lambda} (V(x) - J_d(x)) \leq \lim_{d \to \infty} \frac{1}{d} = 0.$$

However $\{x \in \Lambda : V(x) < \gamma\} = (0, 0.75)$ and for all $d \in \mathbb{N}$ $\{x \in \Lambda : J_d(x) < \gamma\} = (0, 1)$. Therefore

$$\begin{aligned} &D_V(\{x \in \Lambda : V(x) < \gamma\}, \{x \in \Lambda : J_d(x) < \gamma\}) \\ &\qquad = D_V((0, 0.75), (0, 1)) = 0.25 \textit{ for all } d \in \mathbb{N}. \end{aligned}$$

Hence,

$$\lim_{d \to \infty} D_V (\{x \in \Lambda : V(x) < \gamma\}, \{x \in \Lambda : J_d(x) < \gamma\}) = 0.25 \neq 0.$$

Corollary A.1. *Consider a set $\Lambda \in \mathcal{B}$, a function $V \in L^1(\Lambda, \mathbb{R})$, and a family of functions $\{J_d \in L^1(\Lambda, \mathbb{R}) : d \in \mathbb{N}\}$ that satisfies the following properties:*

1. *For any $d \in \mathbb{N}$ we have $J_d(x) \geq V(x)$ for all $x \in \Lambda$.*
2. $\lim_{d \to \infty} ||V - J_d||_{L^1(\Lambda, \mathbb{R})} = 0$.

Then for all $\gamma \in \mathbb{R}$

$$\lim_{d \to \infty} D_V \bigg(\{x \in \Lambda : V(x) < \gamma\}, \{x \in \Lambda : J_d(x) < \gamma\} \bigg) = 0. \tag{A.9}$$

Proof. Let us denote $\tilde{V}(x) = -V(x)$ and $\tilde{J}_d(x) = -J_d(x)$. It follows that $\tilde{J}_d(x) \leq \tilde{V}(x)$ for all $x \in \Lambda$ and $\lim_{d \to \infty} ||\tilde{V} - \tilde{J}_d||_{L^1(\Lambda, \mathbb{R})} = 0$. Therefore, by Prop. A.1 we have that

$$\lim_{d \to \infty} D_V \bigg(\{x \in \Lambda : \tilde{V}(x) \leq \gamma\}, \{x \in \Lambda : \tilde{J}_d(x) \leq \gamma\} \bigg) = 0. \tag{A.10}$$

Now, $\Lambda = \{x \in \Lambda : V(x) < \gamma\} \cup \{x \in \Lambda : V(x) \geq \gamma\} = \{x \in \Lambda : V(x) < \gamma\} \cup \{x \in \Lambda : \tilde{V}(x) \leq \gamma\}$. Therefore

$$\{x \in \Lambda : V(x) < \gamma\} = \Lambda / \{x \in \Lambda : \tilde{V}(x) \leq \gamma\},$$

and by a similar argument

$$\{x \in \Lambda : J_d(x) < \gamma\} = \Lambda / \{x \in \Lambda : \tilde{J}_d(x) \leq \gamma\}.$$

Thus, by Lem. A.2 and since $\{x \in \Lambda : \tilde{J}_d(x) \leq \gamma\} \subseteq \Lambda$, we have that

$$\begin{aligned} &D_V \bigg(\{x \in \Lambda : V(x) < \gamma\}, \{x \in \Lambda : J_d(x) < \gamma\} \bigg) \\ &= D_V \bigg(\Lambda / \{x \in \Lambda : \tilde{V}(x) < \gamma\}, \Lambda / \{x \in \Lambda : \tilde{J}_d(x) < \gamma\} \bigg) \\ &= D_V \bigg(\{x \in \Lambda : \tilde{V}(x) \leq \gamma\}, \{x \in \Lambda : \tilde{J}_d(x) \leq \gamma\} \bigg). \end{aligned} \tag{A.11}$$

Now by Eqs. (A.10) and (A.11) it follows that Eq. (A.9) holds. □

APPENDIX B

MOLLIFICATION THEORY

This Appendix gives a brief overview on the topic of mollification theory, for a more in depth overview we refer to Evans (2010).

Mollifiers The standard mollifier, $\eta \in C^{\infty}(\mathbb{R}^n, \mathbb{R})$ is defined as

$$\eta(x) := \begin{cases} C \exp\left(\frac{1}{||x||_2^2 - 1}\right) & \text{when } ||x||_2 < 1, \\ 0 \quad \text{when } ||x||_2 \geq 1, \end{cases} \tag{B.1}$$

where $C > 0$ is chosen such that $\int_{\mathbb{R}^n} \eta(x) dx = 1$.

For $\sigma > 0$ we denote the scaled standard mollifier by $\eta_\sigma \in C^{\infty}(\mathbb{R}^n, \mathbb{R})$ such that

$$\eta_\sigma(x) := \frac{1}{\sigma^n} \eta\left(\frac{x}{\sigma}\right).$$

Note, clearly $\eta_\sigma(x) = 0$ for all $x \notin B_\sigma(0)$.

Mollification of a Function (Smooth Approximation) Recall from Chapter 2 that for open sets $\Omega \subset \mathbb{R}^n$ and $\sigma > 0$ we denote $< \Omega >_\sigma := \{x \in \Omega : B_\sigma(x) \subset \Omega\}$. Now, for each $\sigma > 0$ and function $V \in L^1(\Omega, \mathbb{R})$ we denote the *σ-mollification* of V by $[V]_\sigma :< \Omega >_\sigma \to \mathbb{R}$, where

$$[V]_\sigma(x) := \int_{\mathbb{R}^n} \eta_\sigma(x - z) V(z) dz = \int_{B_\sigma(0)} \eta_\sigma(z) V(x - z) dz. \tag{B.2}$$

To calculate the derivative of a mollification we next introduce the concept of weak derivatives.

Definition B.1. *For $\Omega \subset \mathbb{R}^n$ and $F \in L^1(\Omega, \mathbb{R})$ we say any $H \in L^1(\Omega, \mathbb{R})$ is the weak $i \in \{1, .., n\}$-partial derivative of F if*

$$\int_\Omega F(x) \frac{\partial}{\partial x_i} \alpha(x) dx = -\int_\Omega H(x) \alpha(x) dx, \ \textit{for} \, \alpha \in C^{\infty}(\mathbb{R}^n, \mathbb{R}).$$

Weak derivatives are “essentially unique”. That is if H_1 and H_2 are both weak derivatives of a function F then the set of points where $H_1(x) \neq H_2(x)$ has measure zero. If a function is differentiable then its weak derivative is equal to its derivative in the “classical” sense. We will use the same notation for the derivative in the “classical” sense and in the weak sense.

In the next proposition we state some useful properties about Sobolev spaces and mollifications taken from Evans (2010).

Proposition B.1 (Evans (2010))**.** *For $1 \leq p < \infty$ and $k \in \mathbb{N}$ we consider $V \in W^{k,p}(E, \mathbb{R})$, where $E \subset \mathbb{R}^n$ is an open bounded set, and its σ-mollification $[V]_\sigma$. Recalling from Chapter 2 that for an open set $\Omega \subset \mathbb{R}^n$ and $\sigma > 0$ we denote $< \Omega >_\sigma := \{x \in \Omega : B_\sigma(x) \subset \Omega\}$, the following holds:*

1. *For all $\sigma > 0$ we have $[V]_\sigma \in C^{\infty}(< E >_\sigma, \mathbb{R})$.*

2. *For all $\sigma > 0$ we have $\nabla_x [V]_\sigma(x) = [\nabla_x V]_\sigma(x)$ for $x \in \, < E >_\sigma$, where $\nabla_x V$ is a weak derivatives.*

3. *If $V \in C(E, \mathbb{R})$ then for any compact set $K \subset E$ we have $\lim_{\sigma \to 0} \sup_{(x,t) \in K} |V(x,t) - [V]_\sigma(x,t)| = 0$.*

4. *(Meyers-Serrin Local Approximation) For any compact set $K \subset E$ we have $\lim_{\sigma \to 0} \| [V]_\sigma - V \|_{W^{k,p}(K,\mathbb{R})} = 0$.*

Note, in the case $E \subset \mathbb{R}^{n+1}$ and $V \in W^{k,p}(E, \mathbb{R})$ (the same as in Chapter 5) Statement 2 in Prop. B.1 can be stated as: "For all $\sigma > 0$ we have $\nabla_t [V]_\sigma(x,t) = [\nabla_t V]_\sigma(x,t)$ and $\nabla_x [V]_\sigma(x,t) = [\nabla_x V]_\sigma(x,t)$ for $(x,t) \in \, < E >_\sigma$, where $\nabla_t V$ and $\nabla_x V$ are weak derivatives", using our ∇_x and ∇_t notation for functions with arguments in $(x,t) \in \mathbb{R}^{n+1}$.

APPENDIX C

MISCELLANEOUS RESULTS

In this appendix we present several miscellaneous results required in various places throughout the manuscript and not previously found in any of the other appendices.

Lemma C.1 (Exponential inequalities). *The following inequalities hold*

$$\exp(-x) \leq 1 \text{ for all } x \geq 0 \tag{C.1}$$

$$x\exp(-x) \leq 1 \text{ for all } x \in \mathbb{R}. \tag{C.2}$$

$$\exp(x) \geq 1 + x \text{ for all } x \in \mathbb{R}. \tag{C.3}$$

Lemma C.2 (Gronwall's Inequality, Hirsch *et al.* (2004)). *Consider scalars $a, b \in \mathbb{R}$ and functions $u, \beta \in C^1(I, \mathbb{R})$. Suppose*

$$\frac{d}{dt}u(t) \leq \beta(t)u(t) \text{ for all } t \in (a, b).$$

Then it follows that

$$u(t) \leq u(a)\exp\left(\int_a^t \beta(s)ds\right) \text{ for all } t \in [a, b].$$

Theorem C.1 (The Bolzano Weierstrass Theorem). *Consider a sequence $\{x_n\}_{n\in\mathbb{N}} \subset \mathbb{R}^n$. Then the $\{x_n\}_{n\in\mathbb{N}}$ is a bounded sequence, that is there exists $M > 0$ such that $x_n < M$ for all $n \in \mathbb{N}$, if and only if there exists a convergent subsequence $\{y_n\}_{n\in\mathbb{N}} \subset \{x_n\}_{n\in\mathbb{N}}$.*

Lemma C.3 (Sublevel sets of continuous functions are closed). *Suppose $f \in C(\mathbb{R}^n, \mathbb{R})$ and Ω is compact set, then the set $\{x \in \Omega : f(x) \leq c\}$, where $c \in \mathbb{R}$, is closed.*

Theorem C.2 (Polynomial Approximation, Peet (2009)). *Let $E \subset \mathbb{R}^n$ be an open set and $f \in C^1(E, \mathbb{R})$. For any compact set $K \subseteq E$ and $\varepsilon > 0$ there exists $g \in \mathcal{P}(\mathbb{R}^n, \mathbb{R})$ such that*

$$\sup_{x\in K} |D^\alpha f(x) - D^\alpha g(x)| < \varepsilon \text{ for all } |\alpha| \leq 1.$$

We next state a result that can be thought of as a generalization of the Weierstrass approximation theorem. It proves there exists a polynomial that can approximate a sufficiently smooth function arbitrarily well with respect to the $W^{1,\infty}$ norm weighted by a function of form $w(x) = 1/||x||_2^{2\beta}$. This result was first presented in the case of $\beta = 1$ in Peet (2009) and then later extended the to general case of $\beta \in \mathbb{N}$ in Leth *et al.* (2017).

Theorem C.3 (Weighted Polynomial Approximation, Leth *et al.* (2017)). *Let $E \subset \mathbb{R}^n$ be an open set, $\beta \in \mathbb{N}$ and $V \in C^{2\beta+2}(\mathbb{R}^n, \mathbb{R})$. For any compact set $K \subseteq E$ and $\varepsilon > 0$ there exists $g \in \mathcal{P}(\mathbb{R}^n, \mathbb{R})$ such that*

$$|V(x) - g(x)| < \varepsilon ||x||_2^{2\beta} \text{ for all } x \in K,$$
$$||\nabla V(x) - \nabla g(x)||_2 < \varepsilon ||x||_2^{2\beta} \text{ for all } x \in K.$$

Theorem C.4 (Rademacher's Theorem, Malỳ and Ziemer (1997) and Evans (2010))**.** *If $\Omega \subset \mathbb{R}^n$ is an open subset and $V \in Lip(\Omega, \mathbb{R})$, then V is differentiable almost everywhere in Ω with point-wise derivative corresponding to the weak derivative almost everywhere; that is the set of points in Ω where V is not differentiable has Lebesgue measure zero. Moreover,*

$$\operatorname*{ess\,sup}_{x\in\Omega} \left| \frac{\partial}{\partial x_i} V(x) \right| \leq L_V \text{ for all } 1 \leq i \leq n,$$

where $L_V > 0$ is the Lipschitz constant of V and $\frac{\partial}{\partial x_i}V(x)$ is the weak derivative of V.

Lemma C.4 (Infimum of family of Lipshitz functions is Lipschitz, Lasz (2014))**.** *Suppose $\{h_\alpha\}_{\alpha\in I} \subset LocLip(\mathbb{R}^n, \mathbb{R})$ is a family of locally Lipschitz continuous functions. Then $h : \mathbb{R}^n \to \mathbb{R}$ defined as $h(x) := \inf_{\alpha \in I} h_\alpha(x)$ is such that $h \in LocLip(\mathbb{R}^n, \mathbb{R})$ provided there exists $x \in \mathbb{R}^n$ such that $h(x) < \infty$.*

Theorem C.5 (Putinar's Positivstellesatz, Putinar (1993))**.** *Consider the semialgebriac set $X = \{x \in \mathbb{R}^n : g_i(x) \geq 0 \text{ for } i = 1, ..., k\}$. Further suppose $\{x \in \mathbb{R}^n : g_i(x) \geq 0\}$ is compact for some $i \in \{1, .., k\}$. If the polynomial $f : \mathbb{R}^n \to \mathbb{R}$ satisfies $f(x) > 0$ for all $x \in X$, then there exists SOS polynomials $\{s_i\}_{i\in\{1,..,m\}} \subset \sum_{SOS}$ such that,*

$$f - \sum_{i=1}^{m} s_i g_i \in \sum_{SOS}.$$

Definition C.1. *Let $\Omega \subset \mathbb{R}^n$. We say $\{U_i\}_{i=1}^\infty$ is an open cover for Ω if $U_i \subset \mathbb{R}^n$ is an open set for each $i \in \mathbb{N}$ and $\Omega \subseteq \{U_i\}_{i=1}^\infty$.*

Theorem C.6 (Existence of Partitions of Unity, Spivak (1965))**.** *Let $E \subseteq \mathbb{R}^n$ and let $\{E_i\}_{i=1}^\infty$ be an open cover of E. Then there exists a collection of $C^\infty(E, \mathbb{R})$ functions, denoted by $\{\psi\}_{i=1}^\infty$, with the following properties:*

1. *For all $x \in E$ and $i \in \mathbb{N}$ we have $0 \leq \psi_i(x) \leq 1$.*
2. *For all $x \in E$ there exists an open set $S \subseteq E$ containing x such that all but finitely many ψ_i are 0 on S.*
3. *For each $x \in E$ we have $\sum_{i=1}^\infty \psi_i(x) = 1$.*
4. *For each $i \in \mathbb{N}$ we have $\{x \in E : \psi_i(x) \neq 0\} \subseteq E_i$.*

Lemma C.5 (Chebyshev's Inequality)**.** *Let (X, Σ, μ) be a measurable space and $f \in L^1(X, \mathbb{R})$. For any $\varepsilon > 0$ and $0 < p < \infty$,*

$$\mu(\{x \in X : |f(x)| > \varepsilon\}) \leq \frac{1}{\varepsilon^p} \int_X |f(x)|^p dx.$$

Lemma C.6 (Equivalence of essential supremum and supremum, Fischer (2015))**.** *Let $E \subset \mathbb{R}^n$ be an open set and $f \in C(E, \mathbb{R})$. Then $\operatorname{ess\,sup}_{x\in E} |f(x)| = \sup_{x\in E} |f(x)|$.*