2018

Stochastic Control Foundations Of Autonomous Behavior

Santiago Paternain
University of Pennsylvania, spater@seas.upenn.edu

Stochastic Control Foundations Of Autonomous Behavior

Abstract
The goal of this thesis is to develop a mathematical framework for autonomous behavior. We begin by describing a minimum notion of autonomy, understood as the ability that an agent operating in a complex space has to satisfy in the long run a set of constraints imposed by the environment of which the agent does not have information a priori. In particular, we care about endowing agents with greedy algorithms to solve problems of the form previously described. Although autonomous behavior will require logic reasoning, the goal is to understand what is the most complex autonomous behavior that can be achieved through greedy algorithms. Being able to extend the class of problems that can be solved with these simple algorithms can allow to free the logic of the system and to focus it towards high-level reasoning and planning.

The second and third chapters of this thesis focus on the problem of designing gradient controllers that allow an agent to navigate towards the minimum of a convex potential in punctured spaces. Such problem is related to the problem of satisfying constraints since we can interpret each constraint as a separate potential that needs to be minimized. We solve this problem first in the case where the information about the potential and the obstacles is deterministic and complete and later, in Chapter \ref{chap_stochnf}, we consider the case where this information is only available from a stochastic model. In both cases, we derive sufficient conditions in which a Rimon-Koditschek artificial potential can be tuned into a navigation function and hence being able to solve the problem. These conditions relate the geometry of the potential of interest and the geometry of the obstacles.

Chapter \ref{chap_feasibility} considers the problem of satisfying a set of constraints when their temporal evolution is arbitrary. We show that an online version of a saddle point controller generates trajectories whose fit and regret are bounded by sublinear functions. These metrics are associated with online operation and they are analogous to feasibility and optimality in classic deterministic optimization. The fact that these quantities are bounded by sublinear functions suggests that the trajectories approach the optimal solution. Saddle points have the advantage of providing an intuition on the relative hardness of satisfying each constraint. The limit values of the multipliers are a measure of such relative difficulty, the larger the multiplier the larger is the cost in which one incurs if we try to tighten the corresponding constraint. In Chapter \ref{chap_counterfactuals} we exploit this property and modify the saddle point controller to deal with situations in which the problems of interest are not feasible. The modification of the algorithm allows us to identify which are the constraints that are harder to satisfy. This information can later be used by a high logic reasoning to modify the problem of interest to make it feasible.

Before concluding remarks and future work we devote our attention to the problem of non-myopic agents. In Chapter \ref{chap_rl} we consider the setting of reinforcement learning where the objective is to maximize the expected cumulative rewards that the agent gathers, i.e., the Q-function. We model the policy of the agent as a function in a Reproducing Kernel Hilbert Space since this class of functions has the advantage of being quite rich and allows us to compute policy gradients in a simple way. We present an unbiased estimator of the policy gradient that can be constructed in finite time and we establish convergence of the stochastic gradient policy ascent to a function that is a critical point of the Q-function.

Degree Type
Dissertation

STOCHASTIC CONTROL FOUNDATIONS OF AUTONOMOUS BEHAVIOR

Santiago Paternain

A DISSERTATION

in

Electrical and Systems Engineering

Presented to the Faculties of the University of Pennsylvania
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
2018

Supervisor of Dissertation

Alejandro Ribeiro, Rosenbluth Associate Professor of Electrical and Systems Engineering

Graduate Group Chairperson

Alejandro Ribeiro, Rosenbluth Associate Professor of Electrical and Systems Engineering

Dissertation Committee
Daniel E. Koditschek, Alfred Fitler Moore Professor, Electrical and Systems Engineering
Manfred Morari, Distinguished Faculty Fellow, Electrical and Systems Engineering
Alec Koppel, Research Scientist, U.S. Army Research Laboratory

STOCHASTIC CONTROL FOUNDATIONS OF AUTONOMOUS BEHAVIOR

Acknowledgments

The years that I spent as a Ph.D. student were some of the most valuables in my life for both professional and personal reasons. I am extremely grateful to be able to appreciate and acknowledge the influence of my advisor, collaborators and friends on writing this thesis.

First and foremost, I would like to express my sincere gratitude to my advisor Prof. Alejandro Ribeiro for opening me the doors of his laboratory. This thesis would not have been possible without his comments, suggestions and support. I want to thank him for the significant impact in shaping my career but also my personality. It has been my great pleasure to have the privilege of being his student. Alejandro has positively influenced my life at a stage that I will always regards fondly and for that reason I am grateful as well.

I would like to further thank Dr. Daniel E. Koditschek, Dr. Manfred Morari and Dr. Alec Koppel, for agreeing to serve in my doctoral committee and for giving me the opportunity to collaborate with them in different projects. Although not all of them resulted in technical content for this thesis, their positive impact on my research experience which led to writing this thesis is invaluable.

Writing this thesis will not have been possible without the joint effort of my collaborators. I would like to thank Dr. Juan Andrés Bazerque from the Universidad de la República in Uruguay, and Austin Small. Over the course of my PhD, I've been fortunate to make some meaningful friendships that have strengthened my research purpose. Specifically, I benefited greatly from working closely with Dr. Aryan Mokthari and Mahyar Fazlyab.

I would also like to extend a deep sense of gratitude to my friends in the laboratory to which I belonged over the past five years. I would especially like to mention here Luiz Chamon, Fernando Gama, Dr. Miguel Calvo-Fullana and Markos Epitropou. Your support from a personal standpoint and frienship made me feel at home. I'd would also like to mention other friends in graduate school with me, whose presence has made the whole process much more enjoyable, and included but not limited to: Ceyhun Eksin, Santiago Segarra, Weiyu Huang, Mark Eisen, Shi-Ling Phuong, Ekaterina Tolstaya, Harshat Kumar, Luana Ruiz, Maria Pfiefer Alp Aydinoglu and Dr. Antonio García-Marques.

I would like to thank my "Philly" friends from different football – soccer for my friends in the United States– teams and the wonderful people I met at the "Asados". Thanks as

well to my friends back home in Uruguay. Special thanks to my parents, Pilar and Miguel. "Thank you" is insufficient for your consistent support, limitless sacrifices and indescribable love. To my grandmother Olga, cousins, uncles and aunts. Last but not least, to Mercedes. Thank you for always being there for me and for believing in me more than myself. This thesis would not have been written without your unconditional love and support. Dedicating this thesis to you is the least I could do to acknowledge your role in writing this thesis and appreciate your kindness and love.

Santiago Paternain, Philadelphia, May 2018

ABSTRACT

STOCHASTIC CONTROL FOUNDATIONS OF AUTONOMOUS BEHAVIOR

Santiago Paternain
Alejandro Ribeiro

The goal of this thesis is to develop a mathematical framework for autonomous behavior. We begin by describing a minimum notion of autonomy, understood as the ability that an agent operating in a complex space has to satisfy in the long run a set of constraints imposed by the environment of which the agent does not have information a priori. In particular, we care about endowing agents with greedy algorithms to solve problems of the form previously described. Although autonomous behavior will require logic reasoning, the goal is to understand what is the most complex autonomous behavior that can be achieved through greedy algorithms. Being able to extend the class of problems that can be solved with these simple algorithms can allow to free the logic of the system and to focus it towards high-level reasoning and planning.

The second and third chapters of this thesis focus on the problem of designing gradient controllers that allow an agent to navigate towards the minimum of a convex potential in punctured spaces. Such problem is related to the problem of satisfying constraints since we can interpret each constraint as a separate potential that needs to be minimized. We solve this problem first in the case where the information about the potential and the obstacles is deterministic and complete and later, in Chapter 3, we consider the case where this information is only available from a stochastic model. In both cases, we derive sufficient conditions in which a Rimon-Koditschek artificial potential can be tuned into a navigation function and hence being able to solve the problem. These conditions relate the geometry of the potential of interest and the geometry of the obstacles.

Chapter 4 considers the problem of satisfying a set of constraints when their temporal evolution is arbitrary. We show that an online version of a saddle point controller generates trajectories whose fit and regret are bounded by sublinear functions. These metrics are associated with online operation and they are analogous to feasibility and optimality in classic deterministic optimization. The fact that these quantities are bounded by sublinear functions suggests that the trajectories approach the optimal solution. Saddle points have the advantage of providing an intuition on the relative hardness of satisfying each constraint. The limit values of the multipliers are a measure of such relative difficulty, the larger the multiplier the larger is the cost in which one incurs if we try to tighten the corresponding constraint. In Chapter 5 we exploit this property, and modify the saddle point controller to deal with situations in which the problems of interest are not feasible. The modification of

the algorithm allows us to identify which are the constraints that are harder to satisfy. This information can later be used by a high logic reasoning to modify the problem of interest to make it feasible.

Before concluding remarks and future work we devote our attention to the problem of non-myopic agents. In Chapter 6 we consider the setting of reinforcement learning where the objective is to maximize the expected cumulative rewards that the agent gathers, i.e., the Q-function. We model the policy of the agent as a function in a Reproducing Kernel Hilbert Space since this class of functions has the advantage of being quite rich and allows us to compute policy gradients in a simple way. We present an unbiased estimator of the policy gradient that can be constructed in finite time and we establish convergence of the stochastic gradient policy ascent to a function that is a critical point of the Q-function. [1]

[1] Work presented in this thesis has been published and submitted for review to IEEE Transactions on Automatic Control and In Proceedings of the American Control Conference and the Decision and Control Conference. Submissions available at [93, 94, 96–100]. Work in this thesis is supported by ARL DCIST CRA W911NF-17-2-0181.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Systems capable of exhibiting autonomous behavior and that are able to perform complex tasks without human assistance are of importance, especially when deployed in environments that are dangerous for humans, like collapsed buildings or zones with toxic or chemical waste spills. Such systems can also be used in theory to perform complex medical procedures or to simplify human tasks in domestic applications and transportation, and therefore it is not surprising that creating autonomous systems has been an actively pursued goal.

While there are many different perspectives of autonomy, a minimal definition is that an agent exhibits autonomous behavior if it can survive when deployed in a complex environment about which there is no information available a priori. Mathematically, we can think of the environment as presenting the agent with a set of unknown functions and of the agent as selecting an action that results in an equal number of payoffs. The agent has the ability to sense the outcome of his actions and must select actions based on a policy that makes the cumulative reward obtained along its trajectory close to a certain value. As a reference problem that can be formulated in this language, consider a drone that is to be positioned within range of a number of targets whose positions are unknown. The drone has to do so with an energy budget and the space in which this task is to be accomplished may be an open field or a wooded area. The drone survives in an environment if it is endowed with an algorithm that allows it to place himself within range of the targets while satisfying the energy constraint. We say that the drone exhibits autonomous behavior if it can satisfy these constraints irrespectively of the environment in which it is deployed – i.e., irrespectively of the position and number of targets and of whether the environment is open or wooded.

A fundamental question that arises from the previous discussion is about the complexity of the algorithms that are to endow an agent with autonomous behavior. It is more or less an accepted consensus that autonomy algorithms have to rely on high-level decision-making rules. While we do not disagree with this statement, our contentions here are that greedy

stochastic control rules can solve a bigger class of problems than currently considered feasible and, as a consequence of sorts, that the interface between low-level greedy rules and high-level logical reasoning have to take a different form. The goal of this thesis is to expand the foundations of stochastic gradient control to incorporate these problems into the class of problems for which greedy algorithms have provable convergence certificates and to exploit this expansion to propose novel methodologies to interface between low-level control and high-level algorithmic reasoning.

To provide a more detailed explanation we adopt a taxonomy that classifies problems with respect to three main properties, the complexity of the space in which the agent is deployed, the amount of information that is available prior to deployment and the time horizon of the operation. The (configuration) spaces in which the agent is deployed can be assumed to be open, punctured or complex. An open environment is one in which there are no restrictions in the configuration space, a punctured space is one that is characterized by the presence of compact obstacles homeomorphic to a point, and a complex environment is one in which the configuration space has an arbitrary shape. The information available a priori is classified as complete, stochastic, viability, or none. Having complete information means that the environment and the constraints to be satisfied are known. Stochastic information means that a probabilistic model of the world is available. Viability means that no information is given except for the knowledge that there is a strategy that would permit satisfaction of all the constraints that specify the environment. In the extreme case, not even this information is available and the agent has to discover whether the environment is viable or not. If the environment is not viable, an autonomous agent has to be able to fall back into a laxer notion of survivability. Finally, the time horizon of the operation distinguishes between myopic operation, where the agent tries to solve the problem for the current time instance, without taking into account the consequences that such actions could have in the future, and farsighted operation, where the agent might make decisions that in the short term are not optimal, but they imply better rewards in the future.

The current state of the start utilizes greedy control to solve myopic problems in open environments with either complete information or stochastic models of environments. Indeed, a problem of complete information in an open environment can be solved with a saddle point controller [4]. These controllers compute gradients for all of the constraints that the agent has to satisfy and moves along a weighted linear combination of them. The coefficients of this linear combination are adapted according to how far from being satisfied the respective constraint is. In that sense, the weights can be thought of as prices. If a constraint is far from being satisfied it means that its satisfaction is relatively difficult and that a large coefficient, i.e., a large price, is justified in the corresponding element of the linear combination. If we keep the environment open but now assume that only a stochastic

model of the environment is given a priori, the situation is much different but the solution methodology is about the same. If the constraints and their gradients can be estimated locally without bias, a stochastic saddle point controller can be proven to converge to a point at which the constraints are satisfied [65]. Do notice that in both cases we must require the existence of a point at which the constraints are satisfied and that the functions that define the environment are convex. This latter condition is not restrictive if we assume that the agent is equipped with local sensors because in that case, a local solution is all that we can hope for.

While greedy saddle point controllers and their stochastic approximations can solve myopic problems in open environments when information is stochastic, this is not enough for the minimal notion of autonomous behavior previously described. Problems in complex environments are addressed with path planning tools which, in the case of stochastic or viability specifications, are coupled with a preliminary mapping stage. Although there are many different specific path planning methods, they can be broadly considered as a decomposition approach because their overall goal is to separate trajectories in complex environment into mesoscale pieces that are locally open and can be planned using tools that work well in open environments. Problems in which it is necessary to discover a measure of what sort of constraints can be satisfied in the environment are addressed heuristically through trial and error. The second drawback about greedy saddle point controllers is that they can only solve problems where the costs are myopic. This is, they offer a solution for an optimization problem without taking into account the rewards –or payoffs– collected along the trajectory. Operating in a regime where one looks at the future is part of the requirements for autonomy. This means, that it is justified to select an action at a given time that results in a low payoff, but it places the system in a state where better rewards can be collected in the long run. Say for instance that an agent requires to place itself to a given distance from a set of targets but it is running out of battery. A myopic controller might prioritize to follow the targets to maximize the payoff without recharging the battery, and thus failing to complete the mission in the future. On the other hand, stopping for re-charging might produce smaller rewards in the short term, but allows the agent to be in a state – high battery level– that allows him in the future to re-position close to the targets for a longer time.

The objective of this thesis is to develop greedy algorithms that allow to bridge the gap, between myopic operation with full information about the environment in open spaces and farsighted operation with no information about the environment in complex spaces.

1.1 Main Contributions

We next detail the thrusts motivated in the previous paragraphs by outlining the work that is presented in this thesis and the particular questions that we address in each chapter.

1.1.1 Navigation functions in punctured spaces.

In the previous paragraphs the motivation of designing algorithms that allow a mobile robot to avoid obstacles has been presented. Many efforts have been made in this direction in situations in which a desired configuration x_d is provided explicitly to the agent. Formally, obstacles are defined as open sets $\mathcal{O}_i$ in the workspace. The set of valid configurations consist of the set difference of the workspace and the obstacles and its termed the free space. The objective is then to converge to x_d while remaining on the free space at all times. A way of greedily solving this problem is through artificial potentials, see e.g. [38, 49, 131]. These potentials are a superposition of an attractive potential – having its minimum at the desired configuration– and repulsive potentials at the obstacles – taking maximum value at their boundary. In some of these constructions convergence to x_d cannot be ensured because of the presence of local minima due to the superposition of several potentials. However, the construction in [57] ensures convergence to the desired point from almost all initial conditions, in a space with spherical obstacles. The potential build in [57] has some defining properties that ensures convergence to x_d and obstacle avoidance. These are that the potential has a unique minimum that coincides with x_d, that all critical points are non degenerate and that the maximum of the potential coincides with the boundary of the obstacles. A potential satisfying these properties define what the authors call a navigation function. In [57] it is also shown that navigation from all initial positions is not possible, and therefore almost sure navigation is the best that one can achieve. The ideas in [57] have been extended to generic star obstacles in [106], yet to do so, a diffeomorphism mapping the world into a spherical world needs to be constructed and to do so, complete information of the environment is required beforehand. The first advantage of this framework compared to some of the other path planning algorithms – such as visibility graphs [81] or cell decomposition [22, 23, 69, 73] – is that it does not require the use of logic and can, therefore, be programmed at a very low level. This can release the logic of these simple task and be available to develop some high level reasoning. A second advantage of gradient descent like algorithms is that they can be easily generalized to systems with intrinsic dynamics. Other path planning algorithms do not take into account the dynamics of the system and therefore may provide trajectories that are not feasible for the robot.

The drawback with the previous approach is that the attractive potential needs to be spherically symmetric. This situation might arise in some problems of interest, but it is

typically the result of knowing the desired destination x_d beforehand. In other applications, however, it is more reasonable to have the desired configuration given as the solution of an optimization problem. As a reference example, think about a robot that is trying to reach the top of a hill. It is more reasonable to assume that the agent can sense its way up the hill using, for instance, an Inertial Measurement Unit (IMU) instead of requiring the location of the top. Along the same lines, it might be of interest to be able the find the source of a given signal, for example, the source of a gas leak which might be expressed as the position for which the gas concentration is the highest. In this setting, it is not reasonable to assume that the position x_d is known and the agent needs to follow the gradient of the intensity of the signal it receives to localize its source.

In Chapter 2 we generalize the artificial potentials from [57] to construct navigation functions in situations in which the attractive potential is not necessarily rotationally symmetric. In particular, we provide sufficient conditions for the possibility of constructing navigation functions of the form in [57]. These conditions relate the geometry of the potential with the geometry of the free-space and the intuition behind them are that the flatter the obstacles are with respect to the level curves of the attractive potential, the hardest it is to tune construct a navigation function.

1.1.2 Online observation of obstacles and environment.

The approach based on navigation functions requires some restrictive assumptions regarding the gradient and the value of the objective function being known exactly at each location. For instance, suppose that a terrestrial robot is trying to reach the top of a hill. The slope of the hill is estimated using measures from onboard accelerometers. These sensors provide noisy measure and hence the estimation cannot be exact. Likewise, it requires complete knowledge of the obstacles, when it is more reasonable to assume that obstacles that are far away from the current position should not influence the behavior of the agent. In addition, the knowledge of the obstacles is inferred using sensors, e.g. LIDARs, and in that sense, the estimates will be contaminated with measurement noise.

Measurements of the objective function $f_0(x)$ or its gradient $\nabla f_0(x)$ can be used to construct an estimator of the gradient of the objective function. This estimate is a random variable denoted by $\hat{\nabla} f_0(x_t, \theta_t)$ which depends on the configuration of the agent at time t and on a random variable θ_t that accounts for measurement noise. If the estimate is unbiased it means that on average the estimate at a given location is equal to the gradient of the function at that point. Formally, it means that the expectation of the noisy gradient with respect to the noise is the gradient itself, i.e., $\mathbb{E}_{\theta_t}\left[\hat{\nabla} f_0(x_t, \theta_t)\right] = \nabla f_0(x_t)$. If we consider the simpler version of the problem in which obstacles are not present, a stochastic version of the gradient descent algorithm ensures convergence to the minimum of f_0 with

probability one (see e.g [107]).

The first question to answer is how to use inexact information to build estimates of the obstacles. For instance, if obstacles are spherical, estimates of the radius and the distance to an obstacle are the minimum information required to avoid them. A naive approach could be to artificially enlarge the obstacles to take into account estimation errors. The insights obtained in the deterministic setting (cf., Chapter 2) suggest that the larger the obstacles and the closer they are to the desired position, make the navigation harder. This has also been observed in [35, 57]. Hence, the previous solution could be over conservative and yield unnecessary stiff trajectories and even make navigation impossible. A second issue to consider is that the navigation function framework relies upon the complete knowledge of the obstacles – shape, position, and size. It is clear that by considering a robot that senses the obstacles as it moves in the space this assumption must be dropped. This fact introduces a mismatch between local estimates the obstacle – obtained for instance by fitting an osculating circle at the closest point of the obstacle from the agent – and the true world.

In Chapter 3 we show that if that said mismatch is not large as compared to the gradient of the navigation function, safe navigation to a neighborhood of the desired configuration is achieved from all initial positions with probability one.

1.1.3 Viability and strategic behavior

The third thrust is related to being able to perform tasks in environments that are time-varying, meaning that the objective functions or the constraints could change over time. In particular, we are interested in adversarial environments, where the change of the function at time t, is such that the action decided at time $t-1$ is the worst choice that we could have selected. To illustrate this idea we can think of the robot as playing a game against the environment. The game is as follows, at time t the agent is allowed to select an action to play at time $t+1$, based on the information of the function that he is trying to minimize at time t. The objective function now is a set of functions $\{f_{0,t}(x), t \in \mathbb{N}\}$ of which the agent only knows at time T the value of the functions $f_{0,t}(x_t)$ for $t = 0 \dots T$. Because of the adversarial nature of the environment and the lack of information about the evolution of it, we cannot possibly expect that the agent minimizes the function $f_{0,t}$ at any time. Therefore the success of an agent in this kind of environment is established through the idea of regret. Regret is the difference between the total loss in which the agent incurs and the loss in which a clairvoyant agent would have incurred if he was allowed to play always the same action. Formally, regret at time T can be expressed as

$$\mathcal{R}_T = \sum_{t=0}^{T} f_{0,t}(x_t) - \min_{x \in \mathbb{R}^n} \sum_{t=0}^{T} f_{0,t}(x). \tag{1.1}$$

If the above quantity is large, having known the evolution of the system we could have chosen a strategy in which the cost incurred is smaller. In that sense, the above quantity measures how much we regret not having that information available. This framework was introduced first in [128] and it has been shown in [138] that an online version of gradient descent achieves regret bounded by $\mathcal{O}(\sqrt{T})$. Having sublinear regret means that the action that we are selecting is approaching the optimal solution. Further works show that by changing the step size of the update can improve the bounds on regret. For instance, in [42] it is shown that online gradient descent with diminishing step size for strongly convex functions archives regret bounded by $\mathcal{O}(\log(T))$. In Chapter 4 we present a continuous time version of this problem and establish regret bounded by a constant. We can think of the problem of satisfying a set of constraint in an adversarial environment as well using a similar concept to that of regret named Fit. The latter is the total constraint violation in which an agent incurs

$$\mathcal{F}_T = \int_0^T f(t, x(t))\, dx. \tag{1.2}$$

This quantity measures – in the same sense that regret measures optimality– how far we are from satisfying the constraints. If there is an action that satisfies the constraints for all times, having known the evolution of the system we could determine this action and have a negative Fit. By having a total constraint violation that grows sublinearly gives the idea of approaching the action that is feasible for all times. In Chapter 4 it is shown that an online version of the algorithm by Arrow Hurwicz proposed in [4] achieves bounded fit irrespectively of the time horizon T. Furthermore, we show that if an optimality criterion is added regret is still bounded by a constant but the fit now is bounded by a function that grows as $\mathcal{O}(\sqrt{T})$.

1.1.4 Price interfaces

Our interest in variations of Arrow and Hurwicz algorithm in [4] is based in two of its characteristics. First of all its simplicity allows it to be implemented in low level controllers. Therefore, releasing the logical reasoning of tasks that up to date is in charge of performing, and allowing it to devote its power to more sophisticated computations. On the other hand, the algorithm provides a very useful way to identify the constraints that are not satisfiable. Saddle point controllers updates the action by descending along a weighted combination of the gradients of the constraints, so to push all the values towards satisfiablity. The weights of the linear combinations are updated in operation time, they are increased if the corresponding constraint is being violated and they are decreased if the constraint is satisfied. The larger a multiplier is the harder it is to satisfy that particular constraint. This observation is the keystone to the integration with logical reasoning. Notice that

through a saddle point algorithm it is easy to identify if the problem is not feasible, because the multipliers keep increasing. With this information the part of the system in charge of the logical reasoning has the information about which constraints should be modified to succeed on its goal or at least it has the information about which of the constraints does not allows him to perform a given task. For instance, let us consider the surveillance problem in which we are interested in tracking several obstacles. Suppose that there is no way of being close to all of the targets, then a mechanism to identify which one of the constraints is the hardest to satisfy can be used by the logical reasoning part of the system to decide a different policy. For instance it could change the problem of being at a given distance of all the targets for a new problem stated as being at a given distance of the target whose multipliers are bounded and adding an optimality criteria given by being as close as possible to the remaining targets. The problem of deciding the policy that must be accomplished is the task of the logical reasoning part of the system, and as discussed the information arising from the low level control is a fundamental piece of information to effectively chose the strategy to follow.

In Chapter 5 we propose a modification of the saddle point algorithm for both the deterministic setting and the setting where a probabilistic model of the constraints and the objective function is available to the agent. This modification introduces adaptive slack variables for each constraint and updates them by increasing its value if the corresponding multiplier is positive and decreases the value if the slacks grows too much. The algorithm is such that it converges to the primal-dual solution for a slack that is proportional to the dual variable. By analyzing the slacks, and the value of the multipliers, we get a relative measure of which constraints are harder to satisfy.

1.1.5 Non-myopic behavior

In the previous discussions, we always consider agents that aim to minimize a given function or to satisfy a set of constraints for which it suffices to find the configuration that allows the agent to get the minimum reward, without taking into account all the payoffs obtained along the trajectory. The last thrust of interest is in situations where we care about non-myopic decision making. This is cases where the agent cares about a policy that allows him to maximize its expected cumulative reward. A common model for these behaviors is based on Markov Decision Processes (MPD), where the state to which the system transitions at a given time is a random variable, whose probability distribution depends on the current state and the action selected by the decision maker. The actions selected by the agent determine instantaneous rewards that can be aggregated over a trajectory to determine cumulative rewards. Hence, the cumulative reward is a measure of the quality of the decision making policy and the objective is therefore not to find the best action but the best policy, i.e.,

the policy that maximizes the expectation of the cumulative reward, also known as the Q-function of the MDP. A solution to these problems can be found in the reinforcement learning literature. This is a model-free control framework for MDPs, where the transition probabilities from one state to next one are not known but the decision policy is based on the rewards obtained. When the state and action spaces are discrete, the solutions to these problems can be divided among those that learn the Q–function to then chose for any given state the action that maximizes the function [132] and those that attempt to directly learn the optimal policy by running gradient ascent in the space of policies [27, 120].

A major drawback of the previous algorithms for reinforcement learning is that they suffer from the curse of dimensionality, this is, the complexity of the problem scales exponentially with the number of actions and states [37]. This is, in particular, the case of continuous spaces, where any reasonable discretization leads to a very large number of states and possible actions. Efforts to sidestep this issue assume that either the Q-function or the policy admits some parametrization [13, 119], or that it belongs to a Reproducing Kernel Hilbert Space (RKHS) [61, 71, 126]. The latter provides the ability to approximate functions using nonparametric functional representations. Although the structure of the space is determined by the choice of the kernel, the set of functions that can be represented is sufficiently rich to permit a good approximation of a large class of functions.

In Chapter 6 we consider policy learning in RKHS and we show, that it is possible to learn a policy that is a stationary point of the Q–function. To do so, we propose an estimate of the gradient of the Q function that is unbiased and that can be computed in finite time. With said estimate, by running stochastic gradient ascent in the space of functions one can establish convergence with probability one.

Chapter 2

Navigation Functions for Convex Potentials in a Space with Convex Obstacles

Given a convex potential in a space with convex obstacles, an artificial potential is used to navigate to the minimum of the natural potential while avoiding collisions. The artificial potential combines the natural potential with potentials that repel the agent from the border of the obstacles. This is a popular approach to navigation problems because it can be implemented with spatially local information that is acquired during operation time. Artificial potentials can, however, have local minima that prevent navigation to the minimum of the natural potential. In this chapter we derive conditions that guarantee artificial potentials have a single minimum that is arbitrarily close to the minimum of the natural potential. The qualitative implication is that artificial potentials succeed when either the condition number– the ratio of the maximum over the minimum eigenvalue– of the Hessian of the natural potential is not large and the obstacles are not too flat or when the destination is not close to the border of an obstacle. Numerical analyses explore the practical value of these theoretical conclusions.

2.1 Introduction

It is customary in navigation problems to define the task of a robot as a given goal in its configuration space; e.g. [24, 68]. A drawback of this approach is the need for global information to provide the goal configuration. In a hill climbing problem, for instance, this means that the position of the top of the hill must be known, when it is more reasonable to assume that the robot senses its way to the top [45,46]. In general, the ability to localize the

source of a specific signal can be used by mobile robots to perform complex missions such as environmental monitoring [92, 117], surveillance and reconnaissance [110], and search and rescue operations [64]. In all these scenarios the desired configuration is not available beforehand but a high level task is nonetheless well defined through the ability to sense the environment.

These task formulations can be abstracted by defining goals that minimize a convex potential, or equivalently, maximize a concave objective. The potential is unknown a priori but its values and, more importantly, its gradients can be estimated from sensory inputs. The gradient estimates derived from sensory data become inputs to a gradient controller that drives the robot to the potential's minimum if it operates in an open convex environment, e.g [43, 122]. These gradient controllers are appealing not only because they exploit sensory information without needing an explicit target configuration, but also because of their simplicity and the fact that they operate using local information only.

In this chapter we consider cases where the configuration space is not convex because it includes a number of nonintersecting convex obstacles. The goal is to design a modified gradient controller that relies on local observations of the objective function and local observations of the obstacles to drive the robot to the minimum of the potential while avoiding collisions. Both, objective function and obstacle observations are acquired at operation time. As a reference example think of navigation towards the top of a wooded hill. The hill is modeled as a concave potential and the trunks a set of nonintersecting convex punctures. The robot is equipped with an inertial measurement unit (IMU) providing the slope's directional derivative, a GPS to measure the current height and a lidar unit giving range and bearing to nearby physical obstacles [45, 46]. We then obtain local gradient measurement from the IMU, local height measurements from the GPS and local models of observed obstacles from the lidar unit and we want to design a controller that uses this spatially local information to drive the robot to the top of the hill.

A possible solution to this problem is available in the form of artificial potentials, which have been widely used in navigation problems, see e.g. [10,11,25,33–36,49–51,57,62,74,75,77, 78,80,91,106,109,131]. The idea is to mix the attractive potential to the goal configuration with repulsive artificial fields that push the robot away from the obstacles. This combination of potentials is bound to yield a function with multiple critical points. However, we can attempt to design combinations in which all but one of the critical points are saddles with the remaining critical point being close to the minimum of the natural potential. If this is possible, a gradient controller that follows this artificial potential reaches the desired target destination while avoiding collisions with the obstacles for almost all initial conditions (see Section 2.2).

The design of mechanisms to combine potentials that end up having a unique minimum

has been widely studied when the natural potential is rotationally symmetric. Koditschek-Rimon artificial potentials are a common alternative that has long been known to work for spherical quadratic potentials and spherical holes [57] and more recently generalized to focally admissible obstacles [35]. In the case of spherical worlds local constructions of these artificial potentials have been provided in [34]. Further relaxations to these restrictions rely on the use of diffeomorphisms that map more generic environments. Notable examples are Koditschek-Rimon potentials in star shaped worlds [105, 106] and artificial potentials based on harmonic functions for navigation of topological complex three dimensional spaces [77,78]. These efforts have proven successful but can be used only when the space is globally known because that information is needed to design a suitable diffeomorphism. Alternative solutions that are applicable without global knowledge of the environment are the use of polynomial navigation functions [74] for n-dimensional configuration spaces with spherical obstacles and [75] for 2-dimensional spaces with convex obstacles, as well as adaptations used for collision avoidance in multiagent systems [28, 109, 124].

Perhaps the most comprehensive development in terms of expanding the applicability of artificial potentials is done in [33, 35, 36]. This series of contributions reach the conclusion that Koditschek-Rimon potentials can be proven to have a unique minimum in spaces much more generic than those punctured by spherical holes. In particular it is possible to navigate any environment that is sufficiently curved. This is defined as situations in which the goals are sufficiently far apart from the borders of the obstacles as measured relative to their flatness. These ideas provides a substantive increase in the range of applicability of artificial potentials as they are shown to fail only when the obstacles are very flat or when the goal is very close to some obstacle border. These curvature conditions seems to be a fundamental requirement of the problem itself rather than of the solution proposed, since it is present as well in other navigation approaches such as navigation via separating hyperplanes [5–7].

Spherical quadratic potentials appear in some specific applications but are most often the result of knowing the goal configuration. Thus, the methods in [10, 11, 25, 33–36, 49–51, 57, 62,74,75,77,78,80,91,106,109,131] are applicable, for the most part, when the goal is known a priori and not when potential gradients are measured during deployment. To overcome this limitation, this work extends the theoretical convergence guarantees of Koditscheck-Rimon functions to problems in which the attractive potential is an arbitrary strongly convex function and the free space is a convex set with a finite number of nonintersecting smooth and strongly convex obstacles (Section 2.2) under mild conditions (Section 2.3). The qualitative implication of these general conditions is that artificial potentials have a unique minimum when one of the following two conditions are met (Theorem 2): (i) The condition number of the Hessian of the natural potential is not large and the obstacles are

not too flat. (ii) The distance from the obstacles' borders to the minimum of the natural potential is large relative to the size of the obstacles. These conditions are compatible with the definition of sufficiently curved worlds in [33]. To gain further insight we consider the particular case of a space with ellipsoidal obstacles (Section 2.3.1). In this scenario the condition to avoid local minima is to have the minimum of the natural potential sufficiently separated from the border of all obstacles as measured by the product of the condition number of the objective and the eccentricity of the respective ellipsoidal obstacle (Theorem 3). The influence on the eccentricity of the obstacles had already been noticed in [33, 36], however the results of Theorem 3 refine those of the literature by providing an algebraic expression to check focal admissibility of the surface.

Results described above are characteristics of the navigation function. The construction of a modified gradient controller that utilizes local observations of this function to navigate to the desired destination is addressed next (Section 2.5). Convergence of a controller that relies on availability of local gradient observations of the natural potential and a local model of the obstacles is proven under the same hypothesis that guarantee the existence of a unique minimum of the potential function (Theorem 4). The local obstacle model required for this result assumes that only obstacles close to the agent are observed and incorporated into the navigation function but that once an obstacle is observed its exact form becomes known. In practice, this requires a space with sufficient regularity so that obstacles can be modeled as members of a class whose complete shape can be estimated from observations of a piece. In, e.g., the wooded hill navigation problem this can be accomplished by using the lidar measurements to fit a circle or an ellipse around each of the tree trunks. The practical implications of these theoretical conclusions are explored in numerical simulations (Section 2.6).

2.2 Problem formulation

We are interested in navigating a punctured space while reaching a target point defined as the minimum of a convex potential function. Formally, let $\mathcal{X} \in \mathbb{R}^n$ be a non empty compact convex set and let $f_0 : \mathcal{X} \to \mathbb{R}_+$ be a convex function whose minimum is the agent's goal. Further consider a set of obstacles $\mathcal{O}_i \subset \mathcal{X}$ with $i = 1 \dots m$ which are assumed to be open convex sets with nonempty interior and smooth boundary $\partial\mathcal{O}_i$. The free space, representing the set of points accessible to the agent, is then given by the set difference between the space $\mathcal{X}$ and the union of the obstacles $\mathcal{O}_i$,

$$\mathcal{F} := \mathcal{X} \setminus \bigcup_{i=1}^{m} \mathcal{O}_i. \tag{2.1}$$

The free space in (2.1) represents a convex set with convex holes; see, e.g., Figure 2.4. We assume here that the optimal point is in the interior $\text{int}(\mathcal{F})$ of free space.

Further let $t \in [0, \infty)$ denote a time index and let x^* be the minimum of the objective function, i.e. $x^* := \text{argmin}_{x \in \mathbb{R}^n} f_0(x)$. Then, the navigation problem of interest is to generate a trajectory $x(t)$ that remains in the free space at all times and reaches x^* at least asymptotically,

$$x(t) \in \mathcal{F}, \ \forall t \in [0, \infty), \quad \text{and} \quad \lim_{t \to \infty} x(t) = x^*. \tag{2.2}$$

In the canonical problem of navigating a convex objective defined over a convex set with a fully controllable agent, convergence to the optimal point as in (2.2) can be assured by defining a trajectory that varies along the negative gradient of the objective function,

$$\dot{x} = -\nabla f_0(x). \tag{2.3}$$

In a space with convex holes, however, the trajectories arising from the dynamical system defined by (2.3) satisfy the second goal in (2.2) but not the first because they are not guaranteed to avoid the obstacles. We aim here to build an alternative function $\varphi(x)$ such that the trajectory defined by the negative gradient of $\varphi(x)$ satisfies both conditions. It is possible to achieve this goal, if the function $\varphi(x)$ is a navigation function whose formal definition we introduce next [57].

Definition 1 (Navigation Function). *Let $\mathcal{F} \subset \mathbb{R}^n$ be a compact connected analytic manifold with boundary. A map $\varphi : \mathcal{F} \to [0, 1]$, is a navigation function in $\mathcal{F}$ if:*

Differentiable. *It is twice continuously differentiable in $\mathcal{F}$.*

Polar at x^*. *It has a unique minimum at x^* which belongs to the interior of the free space, i.e., $x^* \in \text{int}(\mathcal{F})$.*

Morse. *Its critical points on $\mathcal{F}$ are nondegenerate.*

Admissible. *All boundary components have the same maximal value, namely $\partial\mathcal{F} = \varphi^{-1}(1)$.*

The properties of navigation functions in Definition 1 are such that the solutions of the controller $\dot{x} = -\nabla\varphi(x)$ satisfy (2.2) for almost all initial conditions. To see why this is true observe that the trajectories arising from gradient flows of a function φ, converge to the critical points and that the value of the function along the trajectory is monotonically decreasing,

$$\varphi(x(t_1)) \geq \varphi(x(t_2)), \quad \text{for any} \quad t_1 < t_2. \tag{2.4}$$

Admissibility, combined with the observation in (2.4), ensures that every trajectory whose initial condition is in the free space remains on free space for all future times, thus satisfying

the first condition in (2.2). For the second condition observe that, as per (2.4), the only trajectory that can have as a limit set a maximum, is a trajectory starting at the maximum itself. This is a set of zero measure if the function satisfies the Morse property. Furthermore, if the function is Morse, the set of initial conditions that have a saddle point as a limit is the stable manifold of the saddle which can be shown to have zero measure as well. It follows that the set of initial conditions for which the trajectories of the system converge to the local minima of φ has measure one. If the function is polar, this minimum is x^* and the second condition in (2.2) is thereby satisfied. We formally state this result in the next Theorem.

Theorem 1. *Let φ be a navigation function on $\mathcal{F}$ as per Definition 1. Then, the flow given by the gradient control law*

$$\dot{x} = -\nabla\varphi(x), \tag{2.5}$$

has the following properties:

(i) *$\mathcal{F}$ is a positive invariant set of the flow.*

(ii) *The positive limit set of $\mathcal{F}$ consists of the critical points of φ.*

(iii) *There is a set of measure one, $\tilde{\mathcal{F}} \subset \mathcal{F}$, whose limit set consists of x^*.*

Proof. See [55]. □

Theorem 1 implies that if $\varphi(x)$ is a navigation function as defined in 1, the trajectories defined by (2.5) are such that $x(t) \in \mathcal{F}$ for all $t \in [0, \infty)$ and that the limit of $x(t)$ is the minimum x^* for almost every initial condition. This means that (2.2) is satisfied for almost all initial conditions. We can therefore recast the original problem (2.2) as the problem of finding a navigation function $\varphi(x)$. Observe that Theorem 1 guarantees that a navigation function can be used to drive a fully controllable agent [cf. (2.5)]. However, navigation functions can also be used to drive agents with nontrivial dynamics as we explain in Remark 1.

To construct a navigation function $\varphi(x)$ it is convenient to provide a different characterization of free space. To that end, let $\beta_0 : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable concave function such that

$$\mathcal{X} = \left\{ x \in \mathbb{R}^n \,\middle|\, \beta_0(x) \geq 0 \right\}. \tag{2.6}$$

Since the function β_0 is assumed concave its super level sets are convex, thus a function satisfying (2.6) can always be found because the set $\mathcal{X}$ is also convex. The boundary $\partial\mathcal{X}$, which is given by the set of points for which $\beta_0(x) = 0$, is called the external boundary of free

space. Further consider the m obstacles $\mathcal{O}_i$ and define m twice continuously differentiable convex functions $\beta_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1 \dots m$. The function β_i is associated with obstacle $\mathcal{O}_i$ and satisfies

$$\mathcal{O}_i = \left\{ x \in \mathbb{R}^n \,\middle|\, \beta_i(x) < 0 \right\}. \tag{2.7}$$

Functions β_i exist because the sets $\mathcal{O}_i$ are convex and the sublevel sets of convex functions are convex.

Given the definitions of the β_i functions in (2.6) and (2.7), the free space $\mathcal{F}$ can be written as the set of points at which all of these functions are nonnegative. For a more succinct characterization, define the function $\beta : \mathbb{R}^n \to \mathbb{R}$ as the product of the $m + 1$ functions β_i,

$$\beta(x) := \prod_{i=0}^{m} \beta_i(x). \tag{2.8}$$

If the obstacles do not intersect, the function $\beta(x)$ is nonnegative if and only if all of the functions $\beta_i(x)$ are nonnegative. This means that $x \in \mathcal{F}$ is equivalent to $\beta(x) \geq 0$ and that we can then define the free space as the set of points for which $\beta(x)$ is nonnegative – when obstacles are nonintersecting. We state this assumption and definition formally in the following.

AS1 (Obstacles do not intersect). *Let $x \in \mathbb{R}^n$. If for some $i = 1 \dots m$ we have that $\beta_i(x) \leq 0$, then $\beta_j(x) > 0$ for all $j = 0 \dots m$ with $j \neq i$.*

Definition 2 (Free space). *The free space is the set of points $x \in \mathbb{R}^n$ where the function β in (2.8) is nonnegative,*

$$\mathcal{F} = \{ x \in \mathbb{R}^n : \beta(x) \geq 0 \}. \tag{2.9}$$

Observe that we have assumed that the optimal point x^* is in the interior of free space. We have also assumed that the objective function f_0 is strongly convex and twice continuously differentiable and that the same is true of the obstacle functions β_i. We state these assumptions formally for later reference.

AS2. *The objective function f_0, the obstacle functions β_i and the free space $\mathcal{F}$ are such that:*

Optimal point. *$x^* := \operatorname{argmin}_x f_0(x)$ is such that $f_0(x^*) \geq 0$ and it is in the interior of the free space,*

$$x^* \in \mathit{int}(\mathcal{F}). \tag{2.10}$$

Twice differentiable strongly convex objective *The function f_0 is twice continuously differentiable and strongly convex in $\mathcal{X}$. The eigenvalues of the Hessian $\nabla^2 f_0(x)$ are therefore contained in the interval $[\lambda_{\min}, \lambda_{\max}]$ with $0 < \lambda_{\min}$. In particular, strong convexity*

implies that for all $x, y \in \mathcal{X}$,

$$f_0(y) \geq f_0(x) + \nabla f_0(x)^T (y - x) + \frac{\lambda_{\min}}{2} \|x - y\|^2, \tag{2.11}$$

and, equivalently,

$$\left(\nabla f_0(y) - \nabla f_0(x)\right)^T (y - x) \geq \lambda_{\min} \|x - y\|^2. \tag{2.12}$$

Twice differentiable strongly convex obstacles *The function* β_i *is twice continuously differentiable and strongly convex in* $\mathcal{X}$. *The eigenvalues of the Hessian* $\nabla^2 \beta_i(x)$ *are therefore contained in the interval* $[\mu^i_{\min}, \mu^i_{\max}]$ *with* $0 < \mu^i_{\min}$.

The goal in this chapter is to find a navigation function φ for the free space $\mathcal{F}$ of the form of Definition 2 when assumptions 1 and 2 hold. Finding this navigation function is equivalent to attaining the goal in (2.2) for almost all initial conditions. We find sufficient conditions for this to be possible when the minimum of the objective function takes the value $f(x^*) = 0$. When $f(x^*) \neq 0$ we find sufficient conditions to construct a function that satisfies the properties in Definition 1 except for the polar condition that we relax to the function φ having its minimum within a predefined distance of the minimum x^* of the potential f_0. The construction and conditions are presented in the following section after two pertinent remarks.

Remark 1 (**System with dynamics).** If the system has integrator dynamics, then (2.5) can be imposed and problem (2.2) be solved by a navigation function. If the system has nontrivial dynamics, a minor modification can be used [56]. Indeed, let $M(x)$ be the inertia matrix of the agent, $g(x, \dot{x})$ and $h(x)$ be fictitious and gravitational forces, and $\tau(x, \dot{x})$ the torque control input. The agent's dynamics can then be written as

$$M(x)\ddot{x} + g(x, \dot{x}) + h(x) = \tau(x, \dot{x}). \tag{2.13}$$

The model in (2.13) is of control inputs that generate a torque $\tau(x, \dot{x})$ that acts through the inertia $M(x)$ in the presence of the external forces $g(x, \dot{x})$ and $h(x)$. Let $d(x, \dot{x})$ be a dissipative field, i.e., satisfying $\dot{x}^T d(x, \dot{x}) < 0$. Then, by selecting the torque input

$$\tau(x, \dot{x}) = -\nabla\varphi(x) + d(x, \dot{x}), \tag{2.14}$$

the behavior of the agent converges asymptotically to solutions of the gradient dynamical system (2.5) [56]. In particular, the goal in (2.2) is achieved for a system with nontrivial dynamics. Furthermore the torque input above presents a minimal energy solution to the obstacle-avoidance problem [121].

Remark 2 (Example objective functions). *The attractive potential* $f_0(x) = \|x - x^*\|^2$ *is commonly used to navigate to position* x^**. In this work we are interested in more general potentials that may arise in applications where* x^* *is unknown a priori. As a first example consider a target location problem in which the location of the target is measured with uncertainty. This results in the determination of a probability distribution* $p_{x_0}(x_0)$ *for the location* x_0 *of the target. A possible strategy here is to navigate to the expected target position. This can be accomplished if we define the potential*

$$f_0(x) := \mathbb{E}\left[\|x - x_0\|\right] = \int_{\mathcal{F}} \|x - x_0\|\, p_{x_0}(x_0)\, dx_0 \tag{2.15}$$

which is non spherical but convex and differentiable as long as $p_{x_0}(x_0)$ *is a nonatomic dsitribution. Alternatives uses of the distribution* $p_{x_0}(x_0)$ *are possible. An example would be a robust version of* (2.16) *in which we navigate to a point that balances the expected proximity to the target with its variance. This can be formulated by the use of the potential* $f_0(x) := \mathbb{E}\left[\|x - x_0\|\right] + \lambda \mathrm{var}\left[\|x - x_0\|\right]$ *for some* $\lambda > 0$*.*

We can also consider p *targets with location uncertainties captured by probability distributions* $p_{x_i}(x_i)$ *and importance weights* ω_i*. We can navigate to the expected position of the weighted centroid using the potential*

$$f_0(x) := \sum_{i=1}^{p} \omega_i \int_{\mathcal{F}} \|x - x_i\|\, p_{x_i}(x_i)\, dx_i. \tag{2.16}$$

Robust formulations of (2.16) *are also possible.*

2.3 Navigation Function

Following the development in [57] we introduce an order parameter $k > 0$ and define the function φ_k as

$$\varphi_k(x) := \frac{f_0(x)}{\left(f_0^k(x) + \beta(x)\right)^{1/k}}. \tag{2.17}$$

In this section we state sufficient conditions such that for large enough order parameter k, the artificial potential (2.17) is a navigation function in the sense of Definition 1. These conditions relate the bounds on the eigenvalues of the Hessian of the objective function $\lambda_{\min}$ and $\lambda_{\max}$ as well as the bounds on the eigenvalues of the Hessian of the obstacle functions $\mu^i_{\min}$ and $\mu^i_{\max}$ with the size of the obstacles and their distance to the minimum of the objective function x^*. The first result concerns the general case where obstacles are defined through general convex functions.

Theorem 2. *Let* $\mathcal{F}$ *be the free space defined in* (2.9) *satisfying Assumption 1 and let*

$\varphi_k : \mathcal{F} \to [0,1]$ be the function defined in (2.17)*. Let $\lambda_{\max}$, $\lambda_{\min}$ and $\mu^i_{\min}$ be the bounds in Assumption 2. Further let the following condition hold for all $i = 1 \dots m$ and for all x_s in the boundary of $\mathcal{O}_i$*

$$\frac{\lambda_{\max}}{\lambda_{\min}} \frac{\nabla \beta_i(x_s)^T (x_s - x^*)}{\|x_s - x^*\|^2} < \mu^i_{\min}. \tag{2.18}$$

Then, for any $\varepsilon > 0$ there exists a constant $K(\varepsilon)$ such that if $k > K(\varepsilon)$, the function φ_k in (2.17) *is a navigation function with minimum at $\bar{x}$, where $\|\bar{x} - x^*\| < \varepsilon$. Furthermore if $f_0(x^*) = 0$ or $\nabla \beta(x^*) = 0$, then $\bar{x} = x^*$.*

Proof. See Section 2.4. □

Theorem 2 establishes sufficient conditions on the obstacles and objective function for which φ_k defined in (2.17) is guaranteed to be a navigation function for sufficiently large order k. This implies that an agent that follows the flow (2.5) will succeed in navigating towards x^* when $f_0(x^*) = 0$. In cases where this is not the case the agent converges to a neighborhood of x^*. This neighborhood can be made arbitrarily small by increasing k. Of these conditions (2.18) is the hardest to check and thus the most interesting. Here we make the distinction between verifying the condition in terms of design – understood as using the result to define which environments can be navigated – and its verification in operation time. We discuss the former next and we present a heuristic to do the latter in Remark 5. Observe that even if (2.18) needs to be satisfied at all the points that lie in the boundary of an obstacle, it is not difficult to check numerically in low dimensions. This is because the functions are smooth and thus it is possible to discretize the boundary set with a thin partition to obtain accurate approximations of both sides of (2.18). In addition, as we explain next, in practice there is no need check the condition on every point of the boundary. Observe first that, generically, (2.18) is easier to satisfy when the ratio $\lambda_{\max}/\lambda_{\min}$ is small and when the minimum eigenvalue $\mu^i_{\min}$ is large. The first condition means that we want the objective to be as close to spherical as possible and the second condition that we do not want the obstacle to be too flat. Further note that the left hand side of (2.18) is negative if $\nabla \beta_i(x_s)$ and $x_s - x^*$ point in opposite directions. This means that the condition can be violated only by points in the border that are "behind" the obstacle as seen from the minimum point. For these points the worst possible situation is when the gradient at the border point x_s is aligned with the line that goes from that point to the minimum x^*. In that case we want the gradient $\nabla \beta_i(x_s)$ and the ratio $(x_s - x^*)/\|x_s - x^*\|^2$ to be small. The gradient $\nabla \beta_i(x_s)$ being small with respect to $\mu^i_{\min}$ means that we do not want the obstacle to have sharp curvature and the ratio $(x_s - x^*)/\|x_s - x^*\|^2$ being small means that we do not want the destination x^* to be too close to the border. In summary, the simplest navigation problems have objectives and obstacles close to spherical and minima that are not close to the border of the obstacles.

The insights described above notwithstanding, a limitation of Theorem 2 is that it does not provide a trivial way to determine if it is possible to build a navigation function with the form in (2.17) for a given space and objective. In the following section after remarks we consider ellipsoidal obstacles and derive a condition that is easy to check.

Remark 3 (Sufficiently curved worlds [33,35,36]). In cases where the objective function is rotationally symmetric for instance $f_0 = \|x - x^*\|^2$ we have that $\lambda_{\max} = \lambda_{min}$. Let θ_i be the angle between $\nabla\beta_i(x_s)$ and $\nabla f_0(x_s)$, thus (2.18) yields

$$\frac{\|\nabla\beta_i(x_s)\|\cos(\theta_i)}{\|x_s - x^*\|} < \mu^i_{\min}. \tag{2.19}$$

For a world to be sufficiently curved there must exist a direction $\hat{t}_i$ such that

$$\frac{\|\nabla\beta_i(x_s)\|\cos(\theta_i)\hat{t}_i^T D^2 f_0(x_s)\hat{t}_i}{\|\nabla f_0(x_s)\|} < \hat{t}_i^T\nabla^2\beta_i(x_s)\hat{t}_i. \tag{2.20}$$

Since the potential is rotationally symmetric the left hand side of the above equation is equal to the left hand side of (2.19). Observe that, the right hand side of condition (2.19) is the worst case scenario of the right hand side of condition (2.20). These curvature conditions seems to be a fundamental requirement of the problem itself rather than of the solution proposed, since it is present as well in other navigation approaches such as navigation via separating hyperplanes [5–7].

Remark 4. *The condition presented in Theorem 2 is sufficient but not necessary. In that sense, and as shown by the numerical example presented after Thorem 3, it is possible that the artificial potential is a navigation function even when the condition* (2.18) *is violated. Furthermore, in the case of spherical potentials it has been show that the artificial potential yields a navigation function for partially non convex obstacles and for obstacles that yield degenerate criticals points [35, 36]. In terms of the objective function it is possible to ensure navigation by assuming local strict convexity at the goal. However under this assumption condition* (2.18) *takes a form that is not as neat and thus we chose to provide a weaker result in favor of simplicity.*

2.3.1 Ellipsoidal obstacles

Here we consider the particular case where the obstacles are ellipsoids. Let $A_i \in \mathcal{M}^{n\times n}$ with $i = 1\dots m$ be $n \times n$ symmetric positive definite matrices and x_i and r_i be the center and the length of the largest semi-axis of each obstacle $\mathcal{O}_i$. Then, for each $i = 1\dots m$ we define $\beta_i(x)$ as

$$\beta_i(x) := (x - x_i)^T A_i (x - x_i) - \mu^i_{\min} r_i^2, \tag{2.21}$$

The obstacle $\mathcal{O}_i$ is defined as those points in $\mathbb{R}^n$ where $\beta_i(x)$ is not positive. In particular its boundary, $\beta_i(x) = 0$, defines an ellipsoid whose largest semi-axis has length r_i

$$\frac{1}{\mu^i_{\min}} (x - x_i)^T A_i (x - x_i) = r_i^2. \tag{2.22}$$

For the particular geometry of the obstacles considered in this section, Theorem 2 takes the following simplified form.

Theorem 3. *Let $\mathcal{F}$ be the free space defined in* (2.9) *satisfying Assumption 1, and $\varphi_k : \mathcal{F} \to [0, 1]$ be the function defined in* (2.17)*. Let $\lambda_{\max}$, $\lambda_{\min}$, $\mu^i_{\max}$ and $\mu^i_{\min}$ be the bounds from Assumption 2. Assume that β_i takes the form of* (2.21) *and the following inequality holds for all $i = 1..m$*

$$\frac{\lambda_{\max}}{\lambda_{\min}} \frac{\mu^i_{\max}}{\mu^i_{\min}} < 1 + \frac{d_i}{r_i}, \tag{2.23}$$

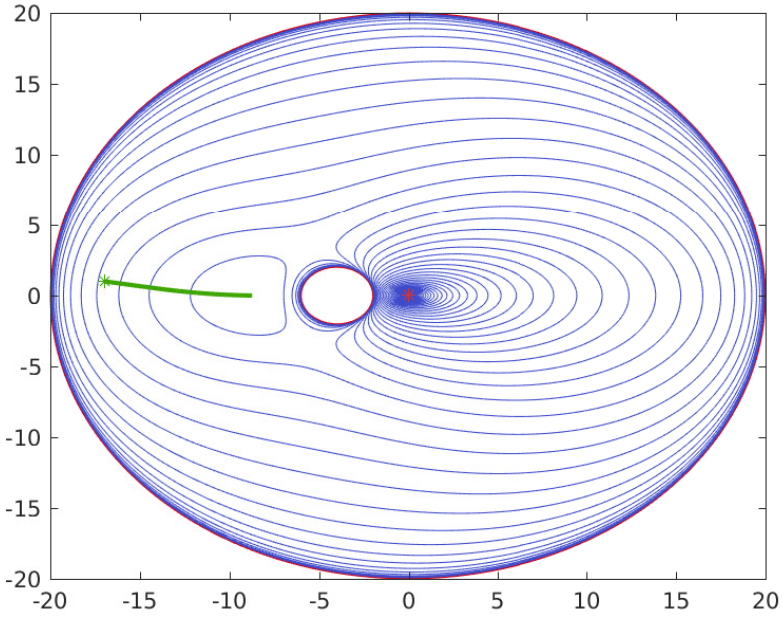
where $d_i := \|x_i - x^*\|$. Then, for any $\varepsilon > 0$ there exists a constant $K(\varepsilon)$, such that if $k > K(\varepsilon)$, the function φ_k in (2.17) *is a navigation function with minimum at $\bar{x}$, where $\|\bar{x} - x^*\| < \varepsilon$. Furthermore if $f_0(x^*) = 0$ or $\nabla\beta(x^*) = 0$, then $\bar{x} = \bar{x^*}$.*

Proof. See Appendix A.1.4. □

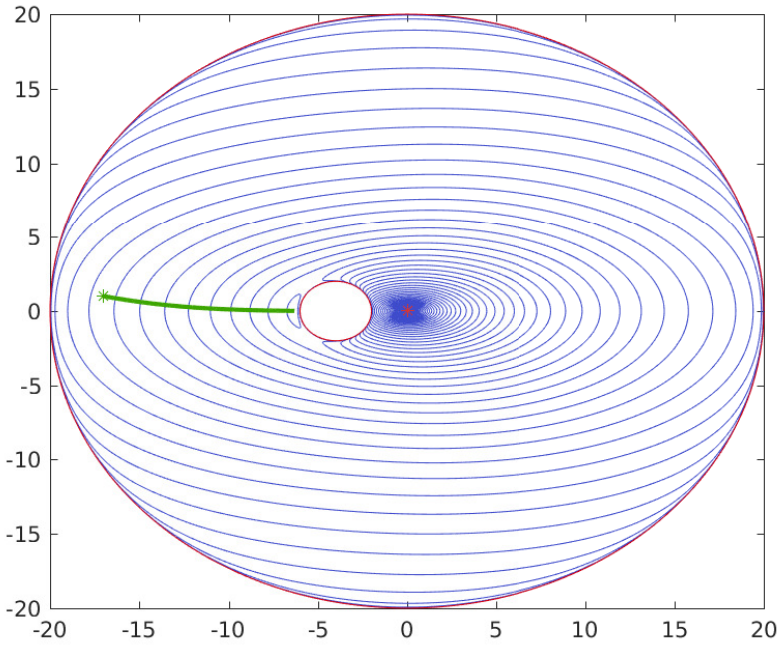
Condition (2.23) gives a simple test to establish that in a given space with ellipsoidal obstacles it is possible to build a Koditscheck-Rimon navigation function. If the inequality is satisfied then it is always possible to select sufficiently large k to make (2.17) a navigation function.

Observe that the more eccentric the obstacles and the level sets of the objective function are, the larger the left hand side of (2.23) becomes and the more difficult it is to guarantee successful navigation. In particular, for a flat obstacle – understood as an ellipse having its minimum eigenvalue equal to zero– the considered condition is impossible to satisfy. For a given eccentricity of the obstacles and the level sets of the objective, the proximity of x^* to the obstacles plays a role. Increasing the distance d_i between the center of the obstacles and the objective, or, equivalently, by decreasing the size of the obstacles r_i, we increase the ratio in the right hand side of (2.23), thereby making it easier to navigate the environment with the potential φ_k. Both of these observations are consistent with Theorem 2. We emphasize that, as is also the case with Theorem 2, the inability to guarantee that it will work, does not mean a navigation function of the proposed form does not exist in the given environment (cf., Remark 4). Conditions (2.18) and (2.23) are shown to be sufficient but not necessary. If the conditions are violated it may nonetheless be possible to build a world in which the proposed artificial potential is a navigation function.

To illustrate ideas, consider an example world in $\mathbb{R}^2$ with only one circular obstacle of



(a) $k = 2$



(b) $k = 10$

Figure 2.1: The artificial potential fails to be a navigation function for $k = 2$ and $k = 10$ when (2.23) is violated and the direction defined by the center of the obstacle and the goal is collinear to the direction of the eigenvector corresponding to the smallest eigenvalue of the Hessian of the objective function.

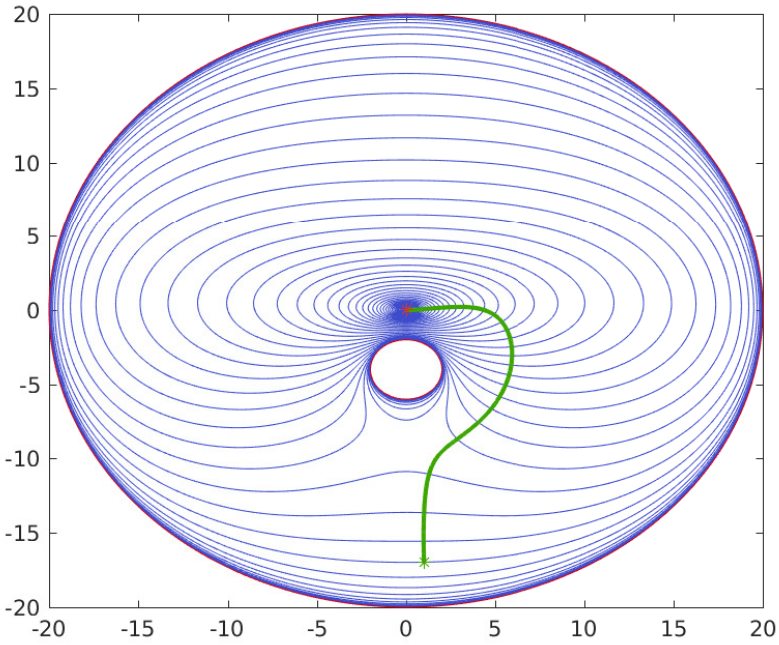


Figure 2.2: For $k = 2$ the artificial potential is a navigation function even though (2.23) is violated but the direction defined by the center of the obstacle and the objective is perpendicular to the direction of the eigenvector corresponding to the smallest eigenvalue of the Hessian of the objective function. Recall that when those directions are collinear (Figures 2.1(a) and 2.1(b)), the potential φ_k fails to be a navigation function.

radius 2 and objective function

$$f_0(x) = x^T \begin{pmatrix} 1 & 0 \\ 0 & \lambda_{\max} \end{pmatrix} x. \tag{2.24}$$

In this example, the minimum of the objective function is attained at the origin and the left hand side of (2.23) takes the value λ_{max}. In the first two simulations we consider the case in which the direction $x_i - x^*$ is aligned with the direction of the eigenvector associated with the smaller eigenvalue of the objective function. This is achieved by placing the center of the obstacle in the horizontal axis at $(-4, 0)$. The right hand side of (2.23) takes therefore the value 3. In the simulations depicted in figures 2.1(a)–2.2, $\lambda_{\max} = 3$, therefore violating condition (2.23). As it can be observed in figures 2.1(a) and 2.1(b) a local minimum other than x^* is present to the left of the obstacle, to which the trajectory converges. Thus, the potential defined in (2.17) fails to be a navigation function. Note that increasing the tuning parameter does not turn the potential into a navigation function since it does not get rid of the local minimum. On the contrary it makes the situation worst, since it pushes the local minimum closer to the obstacle. In Figure 2.2 we observe an example in which the trajectory converges to x^* and condition (2.23) is violated at the same time. Here, the center of the obstacle is placed at $(0, -4)$, and therefore the direction $x_i - x^*$ is no longer aligned with the eigenvector of the Hessian of the objective function associated to the minimum eigenvalue. Hence showing that condition (2.23) is loose when those directions are not collinear.

Notice that the problem of navigating a spherical world to reach a desired destination x^* [57] can be understood as particular case where the objective function takes the form $\|x - x^*\|^2$ and the obstacles are spheres. In this case φ_k is a navigation function for large enough k for every valid world (satisfying Assumption 1), irrespectively of the size and placement of the obstacles. This result can be derived as a corollary of Theorem 3 by showing that condition (2.23) is always satisfied in the setting of [57].

Corollary 1. *Let $\mathcal{F} \subset E^n$ be the set defined in* (2.9) *and let $\varphi_k : \mathcal{F} \to [0, 1]$ be the function defined in* (2.17)*. Let $\mathcal{F}$ verify Assumption 1 and let $f_0(x) = \|x - x^*\|^2$. Let the obstacles be hyper spheres of centers x_i and radii r_i for all $i = 1..m$. Then there exists a constant K such that if k in* (2.17) *is larger than K, then φ_k is a navigation function.*

Proof. Since spherical obstacles are a particular case of ellipsoids the hypothesis of Theorem 3 are satisfied. To show that φ_k is a navigation function we need to show that condition (2.23) is satisfied. For this obstacle geometry we have $\mu^i_{\min} = \mu^i_{\max}$ for all $i = 1 \dots m$. On the other hand, the Hessian of the function $f_0(x) = \|x - x^*\|^2$ is given by $\nabla^2 f_0(x) = 2I$, where I is the $n \times n$ identity matrix. Thus, all its eigenvalues are equal. This implies that the left hand side of (2.23) takes the value one. On the other hand, since d_i and r_i are positive quantities the right hand side of (2.23) is strictly larger than one. Hence the

condition is always satisfied and therefore $\varphi_k(x)$ is a navigation function for some large enough k. □

2.4 Proof of Theorem 2

In this section we show that φ_k, defined in (2.17) is a navigation function under the hypotheses of Theorem 2 by showing that it satisfies Definition 1.

2.4.1 Twice Differentiability and Admissibility

The following lemma shows that the artificial potential (2.17) is twice continuously differentiable and admissible.

Lemma 1 (Differentiability and admissibility). *Let $\mathcal{F}$ be the set defined in* (2.9) *and let $\varphi_k : \mathcal{F} \to [0,1]$ be the function defined in* (2.17)*. Then, φ_k is admissible and twice continuously differentiable on $\mathcal{F}$.*

Proof. Let us show first that φ_k is twice continuously differentiable. To that end we first show that the denominator of (2.17) is strictly positive. For any $x \in \text{int}(\mathcal{F})$ it holds that $\beta(x) > 0$ (cf., (2.9)). Hence $f_0^k(x) + \beta(x) > 0$ because f_0 is nonnegative (cf., Assumption 2). The same holds for $x \in \partial\mathcal{F}$ because the minimum of f_0 is not in $\partial\mathcal{F}$ (cf., Assumption 2). Therefore $\left(f_0^k(x) + \beta(x)\right)^{-1/k}$ is twice continuously differentiable in the free space since f_0 and β are twice continuously differentiable (cf., Assumption 2). Hence φ_k is twice continuously differentiable since it is the product of twice continuously differentiable functions. To show admissibility observe that on one hand for every $x \in \text{int}(\mathcal{F})$ we have that $\beta(x) > 0$, thus $\varphi_k(x) < 1$. On the other hand, if $x \in \partial\mathcal{F}$ we have that $\beta(x) = 0$, hence $\varphi_k(x) = 1$. Thus, the pre image of 1 by φ_k is the boundary of the free space. This completes the proof. □

2.4.2 The Koditschek-Rimon potential φ_k is polar on $\mathcal{F}$

In this section we show that the function φ_k defined in (2.17) is polar on the free space $\mathcal{F}$ defined in (2.9). Furthermore we show that if $f_0(x^*) = 0$ or if $\nabla\beta(x^*) = 0$, then its minimum coincides with the minimum of f_0. If this is not the case, then the minimum of $\varphi_k(x)$ can be placed arbitrarily close to x^* by increasing the order parameter k. In what follows it is convenient to define the product of all the obstacle functions except β_i

$$\bar{\beta}_i(x) := \prod_{j=0, j\neq i}^{m} \beta_j(x). \tag{2.25}$$

Then, for any $i = 0 \dots m$, the gradient of the obstacle function can be written as

$$\nabla\beta(x) = \beta_i(x)\nabla\bar{\beta}_i(x) + \bar{\beta}_i(x)\nabla\beta_i(x). \tag{2.26}$$

The next lemma establishes that $\varphi_k(x)$ does not have critical points at the boundary of the free space.

Lemma 2. *Let $\mathcal{F}$ be the set defined in (2.9) satisfying Assumption 1 and let $\varphi_k : \mathcal{F} \to [0, 1]$ be the function defined in (2.17). Then if Assumption 2 holds there are not critical points of φ_k in the boundary of the free space.*

Proof. For any $x \in \mathcal{F}$ the gradient of φ_k is given by

$$\nabla\varphi_k(x) = \left(f_0^k(x) + \beta(x)\right)^{-1-\frac{1}{k}} \left(\beta(x)\nabla f_0(x) - \frac{f_0(x)\nabla\beta(x)}{k}\right). \tag{2.27}$$

In particular, if $x \in \partial\mathcal{F}$ we have that $\beta(x) = 0$ (cf., (2.9)) and the above expression reduces to

$$\nabla\varphi_k(x) = -\frac{f_0^{-k}(x)}{k}\nabla\beta(x). \tag{2.28}$$

Since f_0 is nonnegative and its minimum is not in the boundary of the free space (cf., Assumption 2), it must be the case that $f_0(x) > 0$. It is left to show that $\nabla\beta(x) \neq 0$ for all $x \in \partial\mathcal{F}$. In virtue of Assumption 1 the obstacles do not intersect. Hence if $x \in \partial\mathcal{F}$, it must be the case that for exactly one of the indices $i = 0 \dots m$ we have that $\beta_i(x) = 0$ (cf., (2.8)). Denote by i^* this particular index. Then (2.26) reduces to

$$\nabla\beta(x) = \bar{\beta}_{i^*}(x)\nabla\beta_{i^*}(x). \tag{2.29}$$

Furthermore we have that for all $j \neq i^*$, $\beta_j(x) > 0$ (cf., (2.21)) hence $\bar{\beta}(x)_{i^*} > 0$. Since the obstacles are non empty open sets and in its boundary $\beta_{i^*}(x) = 0$ and in its interior $\beta_{i^*} < 0$, because β_{i^*} is convex it must be the case that $\nabla\beta_{i^*}(x) \neq 0$ for any $x \in \partial\mathcal{O}_{i^*}$. An analogous argument holds for the case of β_0. This shows that $\nabla\beta(x) \neq 0$ and therefore, there are no critical points in the boundary of the free space. □

In the previous lemma we showed that there are not critical points at the boundary of $\varphi_k(x)$, however we show next that these are either placed arbitrarily close to the boundary of the free space or to x^*. We formalize this result next.

Lemma 3. *Let $\mathcal{F}$ be the free space defined in (2.9) satisfying Assumption 1 and let $\varphi_k : \mathcal{F} \to [0, 1]$ be the function defined in (2.17). Then $\varphi_k(x)$ has critical points $x_c \in int(\mathcal{F})$ for all $k > 0$ and there exists $\varepsilon_0 > 0$ such that for and any $\varepsilon \in (0, \varepsilon_0]$ there exits $K_0(\varepsilon) > 0$ such that if $k > K_0(\varepsilon)$ either $\|\nabla f_0(x_c)\| < \varepsilon$ or $\|\beta_i(x_c)\| < \varepsilon$ for exactly one $i = 1 \dots m$.*

Proof. See appendix A.1.1. □

The previous lemma shows that the critical points of the navigation function can be pushed arbitrarily close to the boundary of one of the obstacles or arbitrarily close to the minimum of the objective function by selecting k sufficiently large. In the next Lemma we show that for large enough k the critical points close to the boundary of the obstacles cannot be local minima. The following lemma as well as Lemma 6 can be derived from [33, 35, 36]. We report the proofs since they are shorter for the particular class of obstacles here considered.

Lemma 4. *Let $\mathcal{F}$ be the free space defined in* (2.9) *satisfying Assumption 1 and let $\varphi_k : \mathcal{F} \to [0, 1]$ be the function defined in* (2.17)*. Let $\lambda_{\max}$, $\lambda_{\min}$ and $\mu^i_{\min}$ be the bounds in Assumption 2. Further let* (2.18) *hold for all $i = 1 \dots m$ and for any $x \in \partial\mathcal{O}_i$. Then, there exists $\varepsilon_1 > 0$ such that for any $\varepsilon \in (0, \varepsilon_1]$, there exists $K_1(\varepsilon)$ such that if $k > K_1(\varepsilon)$, no critical point x_c such that $\beta_i(x_c) < \varepsilon$ is a local minimum.*

Proof. See Appendix A.1.2. □

In the previous Lemma we established that the critical points near the boundary of the free space are not local minima. Therefore the critical points close to x^* have to be. In the next Lemma we formalize this result and we show that for large enough k there is only one nondegenerate minimum.

Lemma 5. *Let $\mathcal{F}$ be the free space defined in* (2.9) *satisfying Assumption 1 and let $\varphi_k : \mathcal{F} \to [0, 1]$ be the function defined in* (2.17)*. Let $\lambda_{\max}$, $\lambda_{\min}$ and $\mu^i_{\min}$ be the bounds in Assumption 2. Further let* (2.18) *hold for all $i = 1 \dots m$ and for all x_s in the boundary of $\mathcal{O}_i$. Then, for any $\varepsilon \in (0, \varepsilon_1]$ there exists $K_2(\varepsilon) > 0$ such that if $k > K_2(\varepsilon)$, φ_k is polar with minimum $\bar{x}$ such that $\|\bar{x} - x^*\| < \varepsilon$. Moreover if $f_0(x^*) = 0$ or $\nabla\beta(x^*) = 0$, then $\bar{x} = x^*$.*

Proof. See Appendix A.1.3. □

The previous lemma establishes that φ_k is polar, with its minimum arbitrarily close to x^* hence we are left to show that the $\varphi_k(x)$ is Morse which we do next.

2.4.3 Nondegeneracy of the critical points

In the previous section, we showed that the navigation function is polar and that the minimum is nondegenerate. Hence, to complete the proof we need to show that the critical points close to the boundary are not degenerate. We formalize this in the following lemma.

Lemma 6. *Let $\mathcal{F}$ be the free space defined in* (2.9) *satisfying Assumption 1 and let $\varphi_k : \mathcal{F} \to [0, 1]$ be the function defined in* (2.17)*. Let $\lambda_{\max}$, $\lambda_{\min}$ and $\mu^i_{\min}$ be the bounds in*

Assumption 2. Further let (2.18) *hold for all* $i = 1 \dots m$ *and for all points in the boundary of* $\mathcal{O}_i$*. Then, for any* $\varepsilon \in (0, \varepsilon_0)$ *there exists* $K_3(\varepsilon)$ *such that if* $k > K_3(\varepsilon)$ *the critical points* x_s *of* φ_k *satisfying* $\beta_i(x_s) < \varepsilon$ *for* $i = 1 \dots m$ *are nondegenerate.*

Proof. We showed in 4 that the Hessian of φ_k evaluated at the critical points satisfying $\beta_i(x_s) < \varepsilon < \varepsilon_0$ has $n-1$ negative eigenvalues when $k > K_1(\varepsilon)$. In particular the subspace of negative eigenvalues is the plane normal to $\nabla\beta(x_s)$. Hence, to show that φ_k is Morse it remains to be shown that the quadratic form associated to $\nabla^2\varphi_k$ at the critical points close to the boundary of the free space is positive when evaluated in the direction of $v = \nabla\beta(x_s)/\|\nabla\beta(x_s)\|$. As previously argued $v^T\nabla\varphi_k(x_s)v > 0$ if and only if

$$v^T \left(\beta(x_s)\nabla^2 f_0(x_s) + (1 - \frac{1}{k})\nabla\beta(x_s)\nabla f_0^T(x_s) \quad - \frac{f_0(x_s)}{k}\nabla^2\beta(x_s) \right) v > 0. \tag{2.30}$$

Note that $\beta(x_s)v^T\nabla^2 f_0(x_s)v$ is positive since f_0 is convex (cf., Assumption 2) and $\beta(x) \geq 0$ for all $x \in \mathcal{F}$ (cf., (2.9)).

For any $k > 1$ the second term in the above equation is positive since $\nabla f_0(x_s)$ and $\nabla\beta(x_s)$ point in the same direction. Moreover since at the boundary of the obstacle $\nabla\beta(x) \neq 0$ (see Lemma 2), for any $\delta > 0$, there exists $K_3\prime(\delta)$ such that if $k > K_3(\delta)$, then $\|\nabla\beta(x_s)\| > \delta$. By virtue of Lemma 3 $\|\nabla f_0(x_s)\| > \varepsilon_0$ hence the second term in the above equation is bounded away from zeros by a constant independent of k. Finally since f_0 and β are twice continuously differentiable $f_0(x)\nabla^2\beta(x)$ is bounded by a constant independent of k for all $x \in \mathcal{F}$. Hence there exists $K_3(\varepsilon) > 0$ such that if $k > K_3(\varepsilon)$ (2.30) holds and therefore the critical points are nondegenerate. □

To complete the proof of Theorem 2 it suffices to choose $K = \max\{K_2(\varepsilon), K_3(\varepsilon)\}$.

2.5 Practical considerations

The gradient controller in (2.5) utilizing the navigation function $\varphi = \varphi_k$ in (2.17) succeeds in reaching a point arbitrarily close to the minimum x^* under the conditions of Theorem 2 or Theorem 3. However, the controller is not strictly local because constructing φ_k requires knowledge of all the obstacles. This limitation can be remedied by noting that the encoding of the obstacles is through the function $\beta(x)$ which is defined by the product of the functions $\beta_i(x)$ [cf., (2.8)]. We can then modify $\beta(x)$ to include only the obstacles that have already been visited. Let $c > 0$ be the a constant defining the range of the sensor that estimates the obstacles and define the c-neighborhood of obstacle $\mathcal{O}_i$ as the set of points with $\beta_i(x) \leq c$. For given time t, we define the set of obstacles of which the agent is aware as the set of

obstacles of which the agent has visited their c-neighborhood at some time $s \in [0, t]$,

$$\mathcal{A}_c(t) := \Big\{ i : \beta_i(x(s)) \leq c, \text{ for some } s \in [0, t] \Big\}. \tag{2.31}$$

The above set can be used to construct a modified version of $\beta(x)$ that includes only the obstacles visited by the agent,

$$\beta_{\mathcal{A}_c(t)}(x) := \beta_0(x) \prod_{i \in \mathcal{A}_c(t)} \beta_i(x). \tag{2.32}$$

Observe that the above function depends on the time through the set $\mathcal{A}_c(t)$ however this dependence is not explicit as the set is only modified when the agent reaches the neighborhood of a new obstacle. In that sense $\mathcal{A}_c(t)$ behaves as a switch depending only of the position of the agent. Proceeding by analogy to (2.17), we use the function $\beta_{\mathcal{A}_c(t)}(x)$ in (2.32) to define the switched potential $\varphi_{k,\mathcal{A}_c(t)}(x) : \mathcal{F}_{\mathcal{A}_c(t)} \to \mathbb{R}$ taking values

$$\varphi_{k,\mathcal{A}_c(t)}(x) := \frac{f_0(x)}{\big(f_0^k(x) + \beta_{\mathcal{A}_c(t)}(x)\big)^{1/k}}. \tag{2.33}$$

The free space $\mathcal{F}_{\mathcal{A}_c(t)}$ is defined as in (2.1), with the difference that we remove only those obstacles for which $i \in \mathcal{A}_c(t)$. Observe that $\mathcal{F}_{\mathcal{A}_c(t)} \subseteq \mathcal{F}_{\mathcal{A}_c(s)}$ if $t > s$. We use this potential to navigate the free space $\mathcal{F}$ according to the switched controller

$$\dot{x} = -\nabla \varphi_{k,\mathcal{A}_c(t)}(x). \tag{2.34}$$

Given that $\varphi_{k,\mathcal{A}_c(t)}(x)$ is a switched potential, it has points of discontinuity. The switched gradient controller in (2.34) is interpreted as following the left limit at the discontinuities. The solution of system (2.34) converges to the minimum of $f_0(x)$ while avoiding the obstacles for a set of initial conditions whose measure is one, as we formally state next.

Theorem 4. *Let $\mathcal{F}$ be the free space defined in* (2.9) *verifying Assumption 1 and let $\mathcal{A}_c(t)$ for any $c > 0$ be the set defined in* (A.46). *Consider the switched navigation function $\varphi_{k,\mathcal{A}_c(t)} : \mathcal{F}_{\mathcal{A}_c(t)} \to [0, 1]$ to be the function defined in* (2.33). *Further let condition* (2.18) *hold for all $i = 1 \dots m$ and for all x_s in the boundary of $\mathcal{O}_i$. Then, for any $\varepsilon > 0$ there exists a constant $K(\varepsilon) > 0$, such that if $k > K(\varepsilon)$, for a set of initial conditions of measure one, the solution of the dynamical system* (2.34) *verifies that $x(t) \in \mathcal{F}$ for all $t \in [0, \infty)$ and its limit is $\bar{x}$, where $\|\bar{x} - x^*\| < \varepsilon$. Furthermore if $f_0(x^*) = 0$ or $\nabla\beta(x^*) = 0$, then $\bar{x} = x^*$.*

Proof. See Appendix A.1.5. □

Theorem 4 shows that it is possible to navigate the free space $\mathcal{F}$ and converge asymptotically to the minimum of the objective function $f_0(x)$ by implementing the switched

dynamical system (2.34). This dynamical system only uses information about the obstacles that the agent has already visited. Therefore, the controller in (2.34) is a spatially local algorithm because the free space is not known a priori but observed as the agent navigates. Do notice that the observation of the obstacles is not entirely local because their complete shape is assumed to become known when the agent visits their respective c-neighborhoods. Incremental discovery of obstacles is also considered in [34] for the case of spherical worlds and the proof is similar to that of Theorem 4. We also point out that a minor modification of (2.34) can be used for systems with dynamics as we formalize in the next proposition.

Corollary 2. *Consider the system given by* (2.13)*. Let* $\varphi_{k,\mathcal{A}_c(t)}(x)$ *be the function given by* (2.33) *and let* $d(x,\dot{x})$ *be a dissipative field, then by selecting the torque input*

$$\tau(x,\dot{x}) = -\nabla\varphi_{k,\mathcal{A}_c(t)}(x) + d(x,\dot{x}), \tag{2.35}$$

the behavior of the agent converges asymptotically to solutions of the gradient dynamical system (2.34)*.*

Proof. From the proof of Theorem 4 it follows that there exists a finite time $T > 0$ such that $\mathcal{A}_c(t)$ is constant for any $t \geq T$ [cf.(A.42)]. Then for any $t \geq T$ the dynamical system given by (2.13) with the torque input (2.35) is equivalent to the system discussed in Remark 1 and the proof of [56] follows. □

The above corollary shows that the goal in (2.2) is achieved for a system with nontrivial dynamics when the obstacles are observed in real time.

Remark 5 (**Selection of navigation function order** k)**.** *Theorems 2 - 4 give conditions for the existence of a constant* K *such that for all* $k \geq K$ *the function* φ_k *in* (2.17) *enables successful navigation to the minimum of the potential function* f_0*. The value of* k *is, however, limited by implementation considerations. E.g., as* k *grows the weight of* $\nabla\beta$ *relative to* ∇f_0 *diminishes [cf.,* (A.44)*], pushing trajectories closer to the obstacles. This is unsafe because noise in sensor inputs and actuation might result in collisions. A pre-design solution is to experiment on the type of environment in which the agent is to be deployed and select a* k *that works in most configurations (Section 2.6). With this implementation restriction Theorems 2 - 4 can not guarantee absence of local minima but rather assure that it is possible to select a* k *that will make them rare for a given family of spatial geometries – indeed, they vanish as* k *grows. Alternatively, and given that using a* k *that is as small as possible is beneficial, algorithms to adapt* k *can be used. For a certain maximum allowable value of* k*, Theorems 2 - 4 do not guarantee absence of local minima but they indicate that local minima are rare. In either case, the agent may get stuck in a local minimum of the artificial potential* φ_k *– this may happen because* k *is not large enough or because the*

geometry of the problem is unworkable for any k. Practical deployments must be combined with a decision making module to dislodge the agent from a local minimum when one is encountered. One possible approach to identifying local minima is to verify that the navigation gradient is $\nabla\varphi_k(x) \approx 0$ but the potential gradient is $\nabla f_0(x) \not\approx 0$.

2.6 Numerical experiments

We evaluate the performance of the navigation function (2.33) in different scenarios. To do so, we consider a discrete approximation of the gradient flow (2.34)

$$x_{t+1} = x_t - \varepsilon_t \nabla\varphi_{k,\mathcal{A}_c(t)}(x_t). \tag{2.36}$$

Where x_0 is selected at random and ε_t is a diminishing step size. In Section 2.6.1 we consider a free space where the obstacles considered are ellipsoids –the obstacle functions $\beta_i(x)$ for $i = 1 \dots m$ take the form (2.21). In particular we study the effect of diminishing the distance between the obstacles while keeping the length of its mayor semi-axis constant. In this section we build the free space such that condition (2.23) is satisfied. As already shown through a numerical experiment in Section 2.3 navigation is still possible if (2.23) is violated (cf., Figure 2.2). This observation motivates the study in Section 2.6.3 where we consider worlds were (2.23) is violated. In 2.6.2 we consider egg shaped obstacles as an example of convex obstacles other than ellipsoids. The numerical section concludes in Section 2.6.4 and 2.6.5 where we consider respectively a system with double integrator dynamics and a wheeled robot.

2.6.1 Elliptical obstacles in $\mathbb{R}^2$ and $\mathbb{R}^3$

In this section we consider m elliptical obstacles in $\mathbb{R}^n$, where $\beta_i(x)$ is of the form (2.22), with $n = 2$ and $n = 3$. We set the number of obstacle to be $m = 2^n$ and we define the external boundary to be a spherical shell of center x_0 and radius r_0. The center of each ellipsoid is placed the position $d\,(\pm 1, \pm 1, \dots, \pm 1)$ and then we perturb this position by adding a vector drawn uniformly from $[-\Delta, \Delta]^n$, where $0 < \Delta < d$. The maximum semi-axis of the ellipse –r_i – is drawn uniformly from $[r_0/10, r_0/5]$. We build orthogonal matrices A_i for $i = 1 \dots m$ where their eigenvalues are drawn from the uniform distribution over $[1, 2]$. We verify that the obstacles selected through the previous process do not intersect and if they do, we re draw all previous parameters. For the objective function we consider a quadratic cost given by

$$f_0(x) = (x - x^*)^T Q\,(x - x^*)\,, \tag{2.37}$$

d	10	9	9	6	6	5	5	3	3
k	2	2	5	5	7	7	10	10	15
Max. final dist.	0.0445	17.25	0.0445	21.61	0.0474	22.29	0.0473	14.28	0.0465
Min initial dist.	10.06	10.01	10.01	10.01	10.02	10.03	10.05	10.12	10.80
Colissions	0	0	0	0	0	0	0	0	0

a Results for the experimental setting described in Section 2.6.1. Observe that the smaller the value of d – the closer the obstacles are between them – the environment becomes harder to navigate, i.e. k must be increased to converge to the minimum of f_0.

d	10	10	9	9	6	6	5	5	3
k	2	15	5	15	7	15	10	15	15
μ_r	1.07	1.01	1.03	1.01	1.19	1.03	1.06	1.05	1.06
$\sigma_r^2(\times 10^{-3})$	6.53	0.07	2.10	0.77	10.1	1.59	6.14	2.57	6.60

b Mean and variance of the ratio between the path length and the initial distance to the minimum. For each scenario 100 simulations were considered. Observe that the smaller the value of d the larger the ratio becomes.

where $x^* = \operatorname{argmin} f_0(x)$ and $Q \in \mathcal{M}^{n\times n}$ is a positive symmetric $n \times n$ matrix. x^* is drawn uniformly over $[-r_0/2, r_0/2]^n$ and we verify that it is in the free space. Then, for each obstacle we compute the maximum condition number, i.e, the ratio of the absolute value of the maximum and minimum eigenvalues, of Q such that (2.18) is satisfied. Let N_{cond} be the largest condition number that satisfies all the constraints. Then, the eigenvalues of Q are selected randomly from $[1, N_{cond} - 1]$, hence ensuring that (2.18) is satisfied. Finally the initial position is also selected randomly over $[-r_0, r_0]^n$ and it is checked that it lies on the free space.

For this experiments we set $r_0 = 20$ and $\Delta = 1$. We run 100 simulations varying the parameter d – controlling the proximity of the obstacles– and k. With this information we build Table 2.1a, where we report the number of collisions, the maximal distance of the last iterate to the minimum of f_0 and the minimal initial distance to the minimum of f_0. As we can conclude from Table 2.1a, the artificial potential (2.33) provides collision free paths. Notice that the smaller the distance between the obstacles the harder is to navigate the environment and k needs to be further increased to achieve the goal. For instance we observe that setting $k = 5$ is sufficient to navigate the world when $d = 9$, yet it is not enough to navigate an environment where $d = 6$. The trajectories arising from artificial potentials typically produce paths whose length is larger than the distance between the initial position and the minimum. We perform a statistical study reporting in Table 2.1b the mean and the variance of the ratio between these two quantities. We only consider those values of d and k that always achieve convergence (cf., Table 2.1a). Observe that when the distance d is reduced while keeping k constant the ratio increases. On the

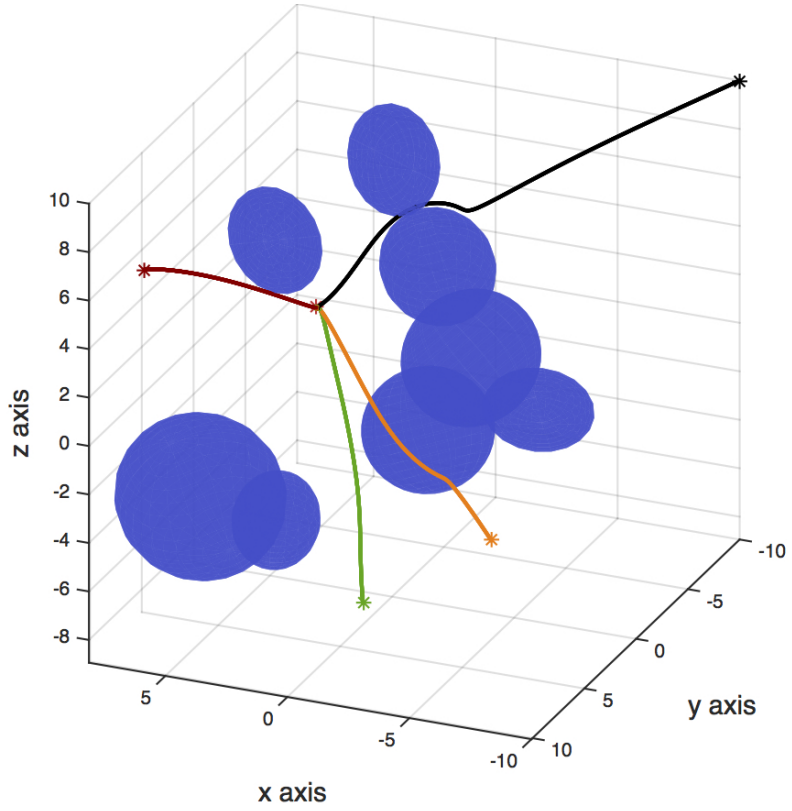


Figure 2.3: Trajectories for different initial conditions in an elliptical world in $\mathbb{R}^3$. As per Theorem 3 and 4 the trajectory converges to the minimum of the objective function while avoiding the obstacles. In this example we have $d = 10$ and $k = 25$.

contrary if d is maintained constant and k is increased the ratio becomes smaller, meaning that the trajectory approaches the optimal one. In Figure 2.3 we simulate one instance of an elliptical world in $\mathbb{R}^3$, with $d = 10$ and $k = 25$. For four initial conditions we observe that the trajectories reach the minimum of f_0.

2.6.2 Egg shaped obstacles

In this section we consider the class of egg shaped obstacles. We draw the center of the each obstacle, x_i, from a uniform distribution over $[-d/2, d/2] \times [-d/2, d/2]$. The distance between the "tip" and the "bottom" of the egg, r_i, is drawn uniformly over $[r_0/10; r_0/5]$ and with probability 0.5, β_i is

$$\beta_i(x) = \|x - x_i\|^4 - 2r_i \left(x^{(1)} - x_i^{(1)}\right)^3, \tag{2.38}$$

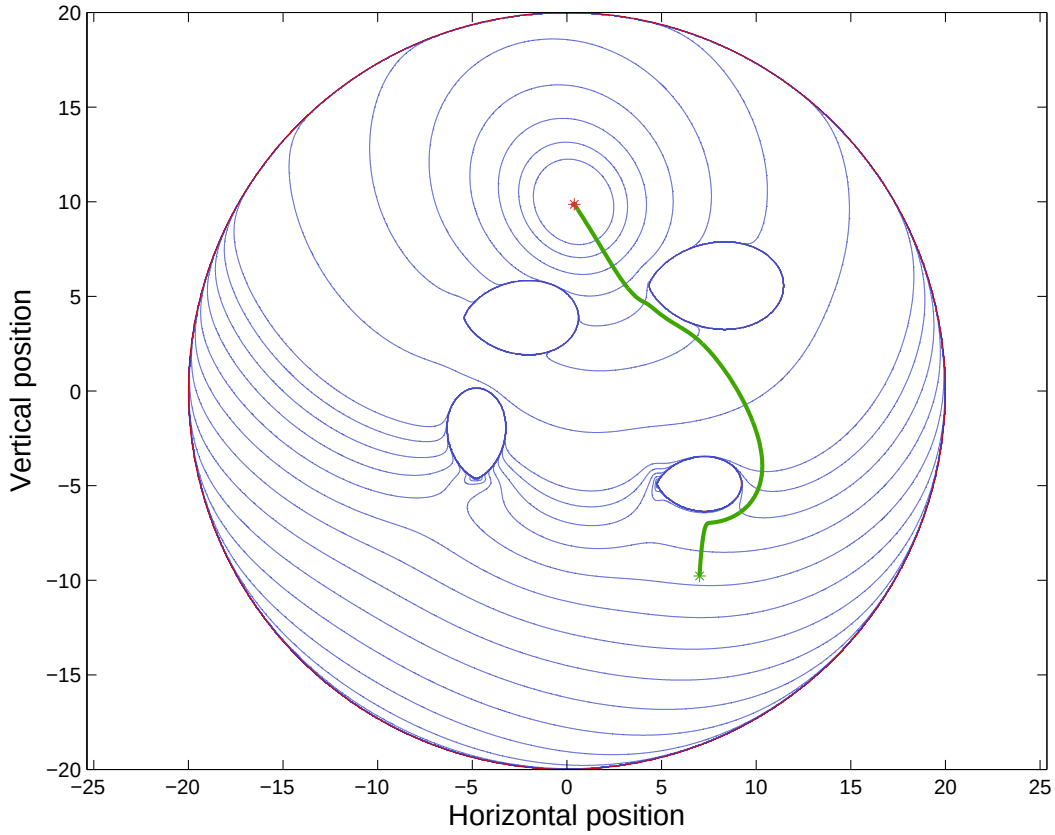


Figure 2.4: Navigation function in an Egg shaped world. As predicted by Theorem 4 the trajectory arising from (2.36) converges to the minimum of the objective function f_0 while avoiding the obstacles.

resulting in a horizontal egg. The superscript (1) refers to first component of a vector. With probability 0.5 the egg is vertical

$$\beta_i(x) = \|x - x_i\|^4 - 2r_i \left(x^{(2)} - x_i^{(2)}\right)^3. \tag{2.39}$$

Notice that the functions β_i as defined above are not convex on $\mathbb{R}^2$, however their Hessians are positive definite outside the obstacles. To be formal we should define a convex extension of the function inside the obstacles in order to say that the function describing the obstacle is convex. This extension is not needed in practice because our interest is limited to the exterior of the obstacle. In Figure 2.4 we observe the level sets of the navigation function and a trajectory arising from (2.36) when we set $k = 25$, $r_0 = 20$ and $d = 10$. In this example the hypotheses of Theorem 2 are satisfied, hence the function φ_k is a navigation function and trajectories arising from the gradient flow (2.34) converge to the optimum of f_0 without running into the free space boundary (cf., Theorem 4).

d	10	9	6	5	3
k	2	5	7	10	15
Success	99%	95%	81%	82%	82%

Table 2.2: Percentage of successful simulations when the condition guaranteeing that φ_k is a navigation function is violated. We observe that as the distance between obstacles becomes smaller the failure percentage increases.

2.6.3 Violation of condition (2.23)

In this section we generate objective functions such that condition (2.23) is violated. To do so, we generate the obstacles as in Section 2.6.1 and the objective function is such that all the eigenvalues of the Hessian are set to be one, except for the maximum which is set to be $\max_{i=1\dots m} N_{cond}+1$, hence assuring that condition (2.23) is violated for all the obstacles. In this simulation Theorem 3 does not ensure that φ_k is a navigation function so it is expected that the trajectory fails to converge. We run 100 simulations for different values of d and k and we report the percentage of successful simulations in Table 2.2. For each value of d the selection of k was done based on Table 2.1a, where k is such that all the simulations attain the minimum of the objective function. Observe that when the distance between the obstacles is decreased the probability of converging to a local minimum different than x^* increases.

2.6.4 Double integrator dynamics

In this section we consider a system with the following simplified version of the dynamics (2.13)

$$\ddot{x} = \tau, \tag{2.40}$$

and the following control law

$$\tau = -\nabla\varphi_k(x) - K\dot{x}. \tag{2.41}$$

In Figure 2.5 we observe the behavior of the system (2.40) when the control law (2.41) is used (green trajectories) against the behavior of the gradient flow system (2.34) (orange trajectory). Thee light green line correspond to a system where the damping constant $K = 4 \times 10^3$ and the dark green correspond to a damping constant of 5×10^3. As we can observe the larger the damping constant the closer the trajectory is to the one of the kinematic system.

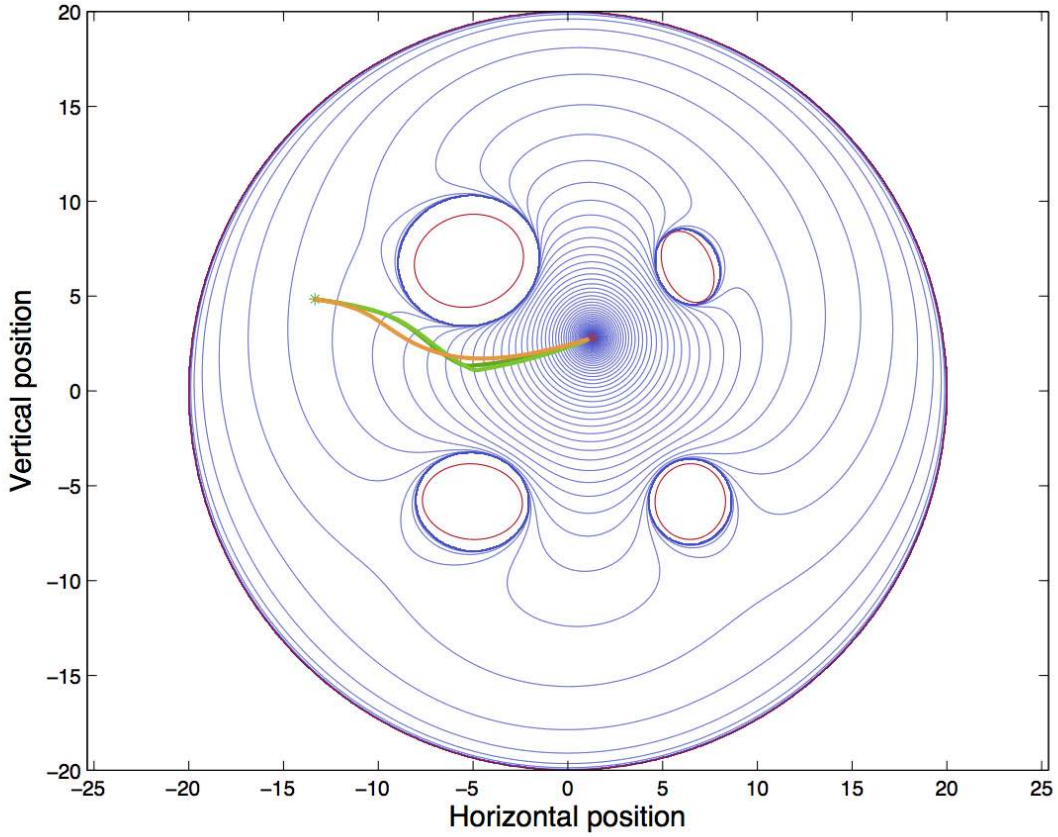


Figure 2.5: In orange we observe the trajectory arising from the system without dynamics (cf., (2.34)). In green we observe trajectories arising from the system (2.40) when we the control law (2.41) is applied. The trajectory in dark green has a larger damping constant than the trajectory in light green and therefore it is closer to the trajectory of the system without dynamics.

2.6.5 Differential drive robot

In this section we consider a disk shaped differential drive robot $(x, \theta) \in \mathbb{R}^2 \times (-\pi, \pi]$, centered at $x \in \mathbb{R}^2$ with body radius $r > 0$ and orientation $\theta \in (-\pi, \pi]$. Its kinematics are given by

$$\dot{x} = v \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}, \quad \dot{\theta} = \omega, \tag{2.42}$$

where v and ω are the linear and angular velocity. The control inputs τ_v and τ_ω actuate respectively over their derivatives

$$\dot{v} = \tau_v, \quad \dot{\omega} = \tau_\omega. \tag{2.43}$$

Observe that the robot described by (2.42) and (2.43) is an under actuated example of the general robot (2.13). Because of the under actuation it is not possible to follow the exact approach described in Remark 1. [125] presents a control law that guarantees theoretical convergence to the minimum of the navigation function for the kinematic model of the differential drive robot. Define the desired angle

$$\theta_d = \arg\left(\frac{\partial\varphi_k(x,y)}{\partial x} + i\frac{\partial\varphi_k(x,y)}{\partial y}\right), \tag{2.44}$$

where $\arg(a + ib)$ is the argument of the complex number $a + ib$. Then the commanded speed is

$$\begin{aligned} v_c = -sgn\left(\frac{\partial\varphi_k(x,y)}{\partial x}\cos\theta + \frac{\partial\varphi_k(x,y)}{\partial y}\sin\theta\right) \\ \left\{k_v\left[\left(\frac{\partial\varphi_k(x,y)}{\partial x}\right)^2 + \left(\frac{\partial\varphi_k(x,y)}{\partial x}\right)^2\right]\right\}. \end{aligned} \tag{2.45}$$

In the above equation $sgn(x)$ is the sign function defined as $sgn(x) = 1$ if ≥ 0 and $sgn(x) = -1$ otherwise. The commanded angular speed is then given by

$$\omega_c = k_\omega\left(\theta_d - \theta\right). \tag{2.46}$$

We propose to extend the previous control law for the dynamic system by setting the linear acceleration to be

$$\tau_v = -v_c - k_{v,d}v, \tag{2.47}$$

and the angular acceleration to be

$$\tau_\omega = -\omega_c - k_{\omega,d}\omega. \tag{2.48}$$

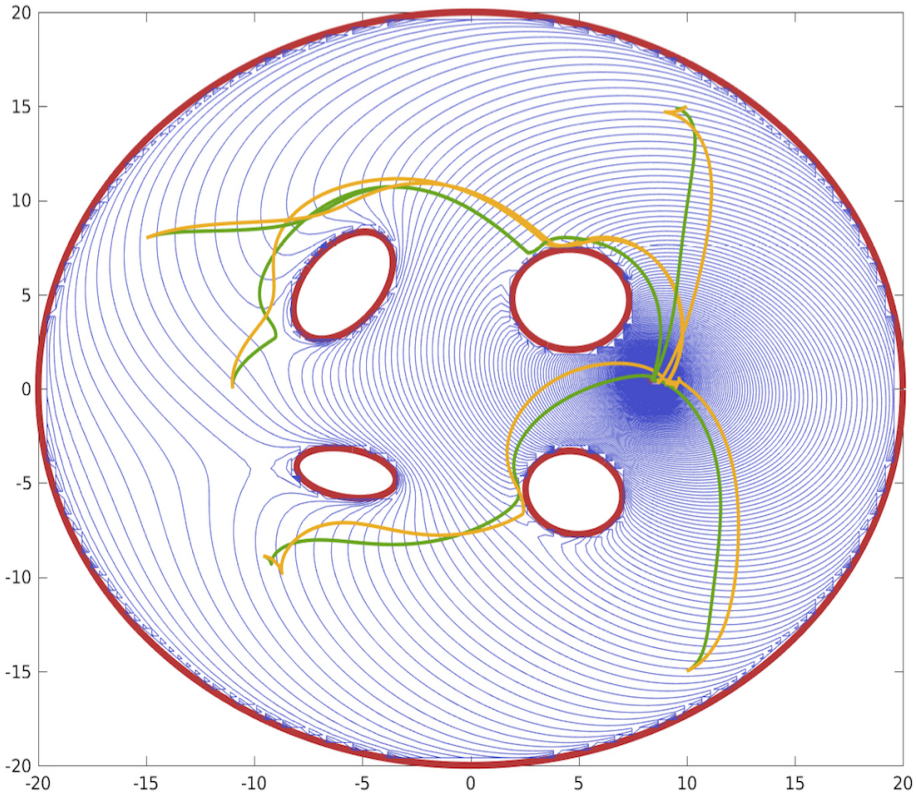


Figure 2.6: In green we depict the trajectories of the kinematic differential drive robot (2.42) , when the control law is given by (2.45) and (2.46). In orange we depict the trajectories of the dynamic differential drive robot (2.42) and (2.43) , when the control law is given by (2.47) and (2.48). In both cases we select $k_v = k_\omega = 1$ and for the dynamic system $k_{v,d} = 4$ and $k_{\omega,d} = 10$. As it can be observed the agent reaches the desired configuration while avoiding the obstacles.

We emphasize that the proposed control does not provide stability guarantees and we are presenting it as an illustration on how to extend the navigation function to systems with dynamics. In Figure 2.6 we depict in green the trajectories of the kinematic differential drive robot (2.42), when the control law is given by (2.45) and (2.46). In orange we depict the trajectories of the dynamic differential drive robot (2.42) and (2.43), when the control law is given by (2.47) and (2.48). In these examples we observe that for $k_v = k_\omega = 1$ and $k_{v,d} = 4$ and $k_{\omega,d} = 10$ the wheeled robot succeeds in reaching the minimum of the objective function while avoiding the obstacles.

2.7 Conclusions

We considered a set with convex holes in which an agent must navigate to the minimum of a convex potential. This function is unknown and only local information about it was used, in particular its gradient and its value at the current location. We defined an artificial

potential function and we showed that under some conditions of the free space geometry and the objective function, this function was a navigation function. Then a controller that moves along the direction of the negative gradient of this function ensures convergence to the minimum of the objective function while avoiding the obstacles. In order to avoid knowing the environment beforehand, a switched controller based on the previous navigation function is defined. This controller only takes into account information about the obstacles that the agent has visited. Numerical experiments support the theoretical results.

Chapter 3

Stochastic Artificial Potentials for Online Safe Navigation

In this Chapter we consider the same type of problems than in Chapter 2. The main difference, is that instead of constructing a navigation function using complete information about the obstacles, we build a stochastic estimate of its gradient, with local information only. The main theoretical contribution is to show that if the estimate available to the agent is unbiased, convergence to the desired location while avoiding the obstacles present in the environment is guaranteed with probability one under the same geometrical conditions than in the deterministic case. Qualitatively these conditions are that the ratio between the maximum and minimum eigenvalue of the Hessian of the objective function is not too large and that the obstacles are not too flat or too close to the desired destination. Moreover, we show that for biased estimates convergence to a point arbitrarily close to the goal is achieved with probability one under assumptions on the bias. These assumptions are motivated by the study of the estimate of the gradient of a Rimon-Koditschek navigation function for sensor models that fit circles around the obstacles. Numerical examples explore the practical value of these theoretical results.

3.1 Introduction

The main drawback of the navigation functions proposed in [57] and explored in part in the previous chapter is that it assume the measurement of the obstacles to be of arbitrary precision. In real robotic systems, however, information about potentials and obstacles is gathered by sensors with noise figures that are not necessarily negligible. This results in observations that are noisy and that, as we explain in Section 3.2.1, are likely to be biased in the case of obstacle estimation. The main contribution of this chapter is to generalize the results in Chapter 2 to stochastic scenarios, understood as settings in which the information

available to the agent comes from a probability distribution instead of being deterministic (Section 3.2). In particular, we show that if the agent is able to construct an unbiased estimate of the gradient of the navigation function, non-collision and convergence to the minimum of the objective function can be ensured with probability one (Theorem 5 Section 3.3).

In most cases, however, constructing an unbiased estimate is not possible, because there exists a mismatch between the real world and the model the agent has of it. This mismatch may be due to not being able to sense all obstacles, or because a simplified model of the world is assumed. However, as long as the bias is not too large compared to the gradient of the navigation function the same theoretical guarantees than in the unbiased case can be provided (Theorem 5). The practical implications of these theoretical conclusions are explored in numerical simulations (Section 3.6) in which we consider the problem of reaching the minimum of non-rotational symmetric potentials in a space where the obstacles are ellipses (Section 3.6.1) and where the obstacles are egg-shaped as an example of a generic convex obstacle (Section 3.6.2). We also consider an artificial potential based on a logarithmic barrier to show that the results of this work are not limited the Rimon-Koditschek artificial potential.

3.2 Problem formulation

In this chapter, we are interested in navigating towards the minimum of a convex potential in a space with convex holes in cases where the information available to the agent about the potential and the free space is local and inexact. As discussed in Chapter 2, a solution to the problem can be obtained through dynamics that follow the negative gradient of a navigation function 1. In particular, in the last chapter we established sufficient conditions for a navigation function of the Koditschek-Rimon form (cf., (2.17) to exists under Assumptions 1 and 2 (cf., Theorem 2). While the navigation function approach provides a provable way of navigating towards the minimum of a convex potential in a cluttered workspace, its drawbacks are twofold: (i) It requires complete characterization of the obstacles to construct the function $\varphi_k(x)$ defined in (2.17). (ii) The measurements of the objective function and the obstacles need to be exact. The main contribution presented in this chapter is to relax these assumptions by considering only local and stochastic information. We describe this framework in the following section.

3.2.1 Navigable Estimates

To model the stochastic nature of the problem we introduce the following probability space $(\Omega, \mathcal{G}, P)$ and we define the following filtration defined as a sequence of increasing sigma

algebras $\{\emptyset, \Omega\} = \mathcal{G}_0 \subset \mathcal{G}_1 \subset \ldots \subset \mathcal{G}_t \subset \ldots \subset \mathcal{G}$. For each $t \geq 0$, define a random vector θ_t to be $\mathcal{G}_t$ measurable. This vector represents the noise of the system at time $t \in \mathbb{N}$. The effect of the noise is that of making the gradient of the navigation $\nabla\varphi_k(x)$ and its estimate $\hat{g}(x_t, \theta_t)$ not collinear. In expectation however, the estimate is related to the gradient of the navigation function as follows

$$\mathbb{E}\left[\hat{g}(x_t, \theta_t)\middle|\mathcal{G}_t\right] = \alpha(x)\left(\nabla\varphi_k(x) + b_k(x)\right), \tag{3.1}$$

where $\alpha : \mathcal{F} \to \mathbb{R}$ is a strictly positive function and $b_k : \mathcal{F} \to \mathbb{R}^n$ is piece-wise differentiable. The bias $b_k(x)$ accounts for a mismatch between the free space and the belief that the agent has of it. Ideally, if the model is perfect, the estimate $\hat{g}(x, \theta)$ is unbiased, i.e., $b_k(x) \equiv 0$. However, if one tries to estimate obstacles using simple models, this is a restrictive assumption. This is observed for instance, in Appendix A.2.1 where we explore the case when the agent assumes the obstacles to be spherical. The origin of the bias is therefore in systematic errors in the estimation and it is not necessarily related to the stochastic nature of the problem; even for noiseless measurements, i.e. $\hat{g}(x_t, \theta_t) = \hat{g}(x_t)$ the bias would be present if the model assumed for the obstacles is not correct.

In what follows we introduce assumptions about the estimate $\hat{g}(x, \theta)$ and the bias $b_k(x)$ that allow navigation towards the minimum of the objective function while avoiding the obstacles (Theorems 5 and 6 which generalize Theorem 2 to the stochastic setting). Estimates satisfying these assumptions are termed navigable estimates and the assumptions presented are motivated in more detail in Appendix A.2.1 where we discuss the case of an agent whose belief is that obstacles are spherical. In the deterministic setting, navigation functions ensure non-collision with the obstacles because in their vicinity, the negative of the gradient of the navigation function is directed outwards the obstacles. Due to the stochastic nature of the estimate $\hat{g}(x, \theta)$ – even if the estimate were to be unbiased – there is no guarantee that this will be the case in general. However, for levels of noise that are not too high this will hold. In order to show that the navigation is collision free the latter assumption along with the boundedness of the estimator are required. We formalize these next.

AS3. *The estimate of the gradient of the navigation function $\hat{g}(x_t, \theta_t)$ satisfies the following properties.*

Bounded *There exists a strictly positive constant B such that for all $x \in \mathcal{F}$ and for all θ we have that*

$$\|\hat{g}(x, \theta)\| \leq B. \tag{3.2}$$

Points outwards the obstacles *Let $d_i(x)$ be the distance between the agent and the obstacle $\mathcal{O}_i$, i.e.,*

$$d_i(x) := \min_{z \in \mathcal{O}_i \cup \partial\mathcal{O}_i} \|x - z\|. \tag{3.3}$$

For each obstacle $\mathcal{O}_i$ there exists a constant $\gamma_i > 0$ such that if $d_i(x) < \gamma_i$ we have for all θ,

$$-\hat{g}(x,\theta)^\top \nabla \beta_i(x) > 0. \tag{3.4}$$

Omnidirectional *There exists a constant $\zeta > 0$ such that for all $y \in \mathbb{R}^n$ with $\|y\| = 1$ we have that*

$$|\mathbb{E}\left[\hat{g}(x,\theta_t)^\top y | \mathcal{G}_t\right] \geq \zeta. \tag{3.5}$$

Biased *Let $\alpha(x) : \mathcal{F} \to \mathbb{R}_{++}$ be a piece-wise differentiable function bounded away from zero and let $b_k(x) : \mathcal{F} \to \mathbb{R}^n$ be piece-wise differentiable on the free space and let $\varphi_k(x)$ be the function defined in* (2.17)*. Then the expected value of the estimate $\hat{g}(x_t,\theta_t)$ with respect to the sigma algebra $\mathcal{G}_t$ satisfies*

$$\mathbb{E}\left[\hat{g}(x_t,\theta_t)\middle|\mathcal{G}_t\right] = \alpha(x_t)\left(\nabla\varphi_k(x_t) + b_k(x_t)\right). \tag{3.6}$$

The omnidirectional assumption is required to ensure that the noise is not driving the system to a specific location. In particular, we use the latter to show that the system does not converge to a saddle point of the navigation function. By Definition 1, navigation functions are Morse in $\mathcal{F}$, and therefore the vector field of its negative gradient is structurally stable in $\mathcal{F}$ [108]. This ensures that if we perturb the vector field by a small quantity, then the resulting flow is topologically equivalent to the original. And therefore the qualitative behavior of the system persists. Since the free space is a manifold with boundary, we need the perturbed vector field not to be tangential at the boundary [108]. This is ensured, for instance, if the bias is zero at the boundary, because $\nabla\varphi_k(x)$ is perpendicular to it. The following assumptions on the bias are sufficient to preserve the qualitative behavior of the system and they are verified, for the most part, by the estimate based on the belief that obstacles are spherical.

AS 4. *The bias defined in* (3.1) *is piece-wise differentiable on the free space and has the following properties.*

Unbiased at the boundary *The bias $b_k(x)$ is such that for any $x \in \partial\mathcal{F}$ we have that $b_k(x) = 0$ for all k.*

Dependence with k *The scaled bias*

$$\tilde{b}_k(x) = b_k(x)\left(f_0(x)^k + \beta(x)\right)^{1+1/k}\Big) \tag{3.7}$$

is such that for any point x in the interior of the free space $\mathcal{F}$ we have that

$$\|\tilde{b}_k(x)\| = O(1/k), \tag{3.8}$$

where $O(1/k)$ is a function satisfying $\lim_{k\to\infty} O(1/k)k = M$ with M a positive constant.

Discontinuities away from the boundary *There exists a constant $D > 0$ such that the function $b_k(x)$ is differentiable for all $x \in \mathcal{F}$ satisfying $\beta_i(x) < D$ for every $i = 1 \dots m$. $b_k(x)$ is also differentiable at x^*.*

Regularity Assumption *Let $\mathcal{U}_k^i$ be the set defined as*

$$\mathcal{U}_k^i = \left\{ x \in \mathcal{F} \middle| \nabla\varphi_k(x)^\top \left(\nabla\varphi_k(x) + b_k(x)\right) \leq 0 \right\} \cap \left\{ x \in \mathcal{F} \middle| \beta_i(x) \leq D \right\}. \tag{3.9}$$

The flows of $\dot{x} = -\nabla\varphi_k(x)$ and $\dot{x} = -\nabla\varphi_k(x) + b_k(x)$ are topologically equivalent in $\mathcal{U}_k^i$.

The regularity assumptions can be interpreted as the bias not being large enough to modify the qualitative behavior of the system. Indeed, topological equivalence can be showed if the norm of the bias in the C^1 sense[1] is sufficiently small with respect to the gradient of the navigation function. For instance, if the model of the world were to be perfect, the bias would be zero and the regularity assumption would hold trivially. As we start simplifying the model of the estimates, the bias will increase. In the case of an agent that fits spheres around the obstacles (cf., Section A.2.1) we can show that the norm of the bias is smaller than that of the gradient of the navigation function. Hence, it is not unreasonable that the regularity assumption holds. An estimate satisfying assumptions 3 and 4 is termed a navigable estimate. We formally define the concept for future reference.

Definition 3 (**Navigable Estimates**). *Let $\theta \in \mathbb{R}^p$ be a random vector and let $\hat{g}(x, \theta)$ be an estimate of the gradient of navigation function of the form* (2.17) *for the free space $\mathcal{F}$ defined in* (2.9)*. We say that $\hat{g}(x, \theta)$ is a Navigable Estimate if it satisfies assumptions 3 and 4.*

Drawing inspiration from the deterministic scenario we propose a stochastic gradient descent scheme to solve (2.2) using only local and stochastic information, based on navigable estimates $\hat{g}(x, \theta)$, in which the agent updates its configuration recursively as

$$x_{t+1} = x_t - \eta_t \hat{g}(x_t, \theta_t), \tag{3.10}$$

where η_t is a step size assumed to be not summable and square summable. A particular step size that satisfies the previous conditions is $\eta_t = \eta_0/(1 + \zeta t)$, where η_0 is the initial step size and ζ controls the rate at which the step size is decreased. We formalize the assumption on he step size for future reference.

[1] Given a vector field $f(x)$ we denote its n-derivative by $D^{(n)}f(x)$. We define the C^n norm of a vector field $f(x)$ in a manifold M as $\|f(x)\|_{C^n} = \sup_{x\in M} \left\{ \|f(x)\|, \|Df(x)\|, \dots, \|D^{(n)}f(x)\| \right\}$.

AS5. *The step size η_t for the update* (3.10) *is a positive and strictly decreasing sequence that satisfies*

$$\sum_{t=0}^{\infty} \eta_t = \infty, \quad \sum_{t=0}^{\infty} \eta_t^2 < \infty, \quad \eta_0 < \min_i \frac{\gamma_i}{B}, \tag{3.11}$$

where γ_i and B are the constants defined in Assumption 3.

The main contribution of this chapter is to show that an agent operating in a workspace with convex holes, that is given a navigable estimate of the form (3.1) is able to reach the minimum of an unknown convex function without running into the free space boundary with probability one (Section 3.4). Before doing so, we present a preliminary result for unbiased estimates (Section 3.3).

3.3 Unbiased Estimator

In this section we consider the particular case of an agent that has access to an unbiased estimator of the gradient of the navigation function rather than the general model presented in (3.1). This means that the bias is identically zero $b_k(x) \equiv 0$. Notice that, such choise of the bias satisfies Assumption 4 trivially. We show that, in this case, an agent that follows the gradient update (3.10) converges to the minimum of the navigation function $\varphi_k(x)$ defined in (2.17) while avoiding the obstacles with probability one. Therefore solving problem (2.2). We start by establishing obstacle avoidance in the following lemma.

Lemma 7. *Let $\mathcal{F}$ be the free space defined in* (2) *verifying Assumption 1. Furthermore, let $\hat{g}(x_t, \theta_t)$ be an estimate of the gradient of the navigation function* (2.17) *satisfying Assumption 3. Then, the update* (3.10)*, with step size satisfying Assumption 5, ensures that $\{x_t, t \geq 0\} \in \mathcal{F}$.*

Proof. Denote by $d_i(x)$ the euclidean distance betwen $x \in \mathcal{F}$ and the set $\mathcal{O}_i$ and note that the triangle inequality implies

$$d_i(x_{t+1}) \geq d_i(x_t) - \eta_t \|\hat{g}(x_t, \theta_t)\|. \tag{3.12}$$

Because the estimate of the gradient of the navigation function satisfies that $\|\hat{g}(x_t, \theta_t)\| \leq B$ (cf., Assumption 3) and η_t is a decreasing sequence with $\eta_0 \leq \min_i \gamma_i / B$ (cf., Assumption 5), we have that $\eta_t \|\hat{g}(x_t, \theta_t)\| < \min_i \{\gamma_i\}$ for all t. Therefore, for cases in which $d_i(x_t) \geq \gamma_i$ (3.12) can be lower bounded by

$$d_i(x_{t+1}) > \gamma_i - \min_i \gamma_i \geq 0. \tag{3.13}$$

Which shows that if the distance between the iterate x_t and the obstacle $\mathcal{O}_i$ is larger

than γ_i, then $x_{t+1} \in \mathcal{F}$. In the opposite case, by virtue of Assumption 3, we have that $-\hat{g}(x_t, \theta_t)^\top \nabla \beta_i(x_t) > 0$ and therefore non-collision with obstacle $\mathcal{O}_i$ is ensured trivially. $\square$

The previous lemma shows that the sequence generated by (3.10) avoids the obstacles. We turn to showing convergence to the minimum of $\varphi_k(x)$ starting by showing convergence to the set of its critical points with probability one. To do so, we use a supermartingale convergence result. A supermartingale is a sequence of random variables that decreases in expectation. In a way, it is a stochastic generalization of a decreasing sequence. If such sequence is bounded below, then the convergence of the supermartingale can be established [29, Theorem 5.2.9]. This result is hence a generalization of the deterministic result stating that a decreasing sequence that is bounded below converges.

Lemma 8. *Let $\mathcal{F}$ be the free space defined in* (2) *verifying Assumption 1 and let* (2.18) *hold. Denote by $\hat{g}(x_t, \theta_t)$ an unbiased navigable estimate as in definition 3 of the gradient of the artificial potential* (2.17) *and let η_t be a sequence satisfying Assumption 5. Then, for any $x_0 \in \mathcal{F}$ the sequence generated by* (3.10) *is such that*

$$\lim_{t \to \infty} x_t = X_c \quad a.e., \tag{3.14}$$

where X_c is a random variable taking values on the set of the critical points of $\varphi_k(x)$.

Proof. Let us write $\varphi_k(x_{t+1})$ in terms of the previous iterate using its Taylor expansion around x_t and the update (3.10)

$$\varphi_k(x_{t+1}) = \varphi_k(x_t) - \eta_t \nabla \varphi_k(x_t) \hat{g}(x_t, \theta_t) + \frac{\eta_t^2}{2} \hat{g}(x_t)^\top \nabla^2 \varphi_k(z) \hat{g}(x_t), \tag{3.15}$$

where z is a point in the segment $x_t - \mu \eta_t \hat{g}(x_t)$ with $\mu \in [0, 1]$. Said segment lies in the free space by virtue of Lemma 7. The free space being a compact set and $\varphi_k(x)$ being a twice differentiable function (cf., Definition 1), implies that the maximum eigenvalue of $\nabla^2 \varphi_k(x)$ is upper bounded by a positive constant L. Then using the bound on the norm of the estimate of the gradient (cf., Assumption 3) the quadratic term in (3.15) can be bounded by

$$\hat{g}(x_t, \theta_t)^\top \nabla^2 \varphi_k(z) \hat{g}(x_t, \theta_t) \leq L B^2. \tag{3.16}$$

Consider the expectation with respect to the sigma field $\mathcal{G}_t$ on both sides of (3.15). Using the linearity of the expectation, the fact that $\varphi_k(x_t)$ is $\mathcal{G}_t$ measurable and the bound derived in (3.16) we have that

$$\mathbb{E}\left[\varphi_k(x_{t+1}) \middle| \mathcal{G}_t\right] \leq \varphi_k(x_t) - \eta_t \mathbb{E}\left[\nabla \varphi_k(x_t)^\top \hat{g}(x_t, \theta_t) \middle| \mathcal{G}_t\right] + \eta_t^2 \frac{L B^2}{2}. \tag{3.17}$$

We next show that the following sequence is a nonnegative supermartingale

$$S_t = \varphi_k(x_t) + \sum_{s=t}^{\infty} \eta_s^2 \frac{LB^2}{2}. \tag{3.18}$$

Since $\varphi_k(x)$ is a navigation function it is nonnegative and bounded (cf., Definition 1), therefore S_t is a nonnegative sequence and bounded because η_t is square summable (cf., Assumption 5). S_t is also adapted to $\mathcal{G}_t$ since x_t is. Thus, in order to show that S_t is a nonnegative supermartingale it suffices to prove that $\mathbb{E}\left[S_{t+1}\middle|\mathcal{G}_t\right] \leq S_t$, which we do next. Using the linearity of the expectation and the bound for $\mathbb{E}\left[\varphi_k(x_{t+1})\middle|\mathcal{G}_t\right]$ derived in (3.17) we have that

$$\mathbb{E}\left[S_{t+1}\middle|\mathcal{G}_t\right] \leq \varphi_k(x_t) + \sum_{s=t}^{\infty} \eta_s^2 \frac{LB^2}{2} - \eta_t \mathbb{E}\left[\nabla\varphi_k(x_t)^\top \hat{g}(x_t, \theta_t)\middle|\mathcal{G}_t\right]. \tag{3.19}$$

Since the estimator is navigable and it is unbiased, we have that $\mathbb{E}\left[\hat{g}(x_t, \theta_t)\middle|\mathcal{G}_t\right] = \alpha(x_t)\nabla\varphi_k(x_t)$ and therefore

$$\mathbb{E}\left[\nabla\varphi_k(x_t)^\top \hat{g}(x_t, \theta_t)\middle|\mathcal{G}_t\right] = \alpha(x)\|\nabla\varphi_k(x_t)\|^2 \geq 0, \tag{3.20}$$

where the las inequality holds because $\alpha(x)$ is strictly positive (cf., Assumption 3). Thus S_t is nonnegative supermartingale and it holds that (see e.g. Theorem 5.2.9 in [29])

$$\lim_{t\to\infty} S_t = S \quad \text{a.e.}, \tag{3.21}$$

where S is a random variable such that $\mathbb{E}[S] \leq \mathbb{E}[S_0]$ and

$$\sum_{t=0}^{\infty} \eta_t \alpha(x_t)\|\nabla\varphi_k(x_t)\|^2 < \infty \quad \text{a.e.}. \tag{3.22}$$

Since the sequence of step sizes $\{\eta_t, t \geq 0\}$ is not summable and $\alpha(x)$ is bounded away from zero (cf., Assumption 3) the convergence of the above series implies that

$$\liminf_{t\to\infty} \|\nabla\varphi_k(x_t)\|^2 = 0 \quad \text{a.e.}. \tag{3.23}$$

We are left to show that $\limsup_{t\to\infty} \|\nabla\varphi_k(x_t)\| = 0$ almost everywhere. Before doing so, observe that if this is the case there exists a subsequence $\{x_{t_s}, s \in \mathbb{N} \cup \{0\}\}$ that converges to the set of critical points of the navigation function $\varphi_k(x)$. Since the limit of S_t exists we have that

$$\lim_{s\to\infty} S_{t_s} = \lim_{s\to\infty} \varphi_k(x_{t_s}) = S \quad \text{a.e.} \tag{3.24}$$

Then, observe that the critical points of the navigation function are nondegenerate (cf.,

Definition 1), and therefore the limit of the sequence x_t generated by the update (3.10) is either the minimum of $\varphi_k(x)$ or one of the saddles of $\varphi_k(x)$. To complete the proof of the Lemma we show by contradiction that $\limsup_{t\to\infty} \|\nabla\varphi_k(x_t)\| = 0$ almost everywhere.

Assume that $\limsup_{t\to\infty} \|\nabla\varphi_k(x_t(\omega))\| = \delta > 0$ for some $\omega \in \Omega$. That being the case, there exists sequences $\{T_s\}$ and $\{T'_s\}$ such that $T_s < T'_s < T_{s+1}$ and

$$\frac{\delta}{3} < \|\nabla\varphi_k(x_t)\| \quad \text{for} \quad T_s \le t < T'_s \quad \text{and} \quad \|\nabla\varphi_k(x_t)\| \le \frac{\delta}{3} \quad \text{for} \quad T'_s \le t < T_{s+1}. \tag{3.25}$$

Then choose $T \in \{T_s, \ldots, T'_s\}$. Using the fact that $\nabla^2\varphi_k(x)$ is bounded by L for all $x \in \mathcal{F}$, it is possible to bound the norm of the difference of the gradients of $\varphi(x)$ evaluated at $t = T$ and $t = T'_s$ by

$$\left\|\nabla\varphi(x_T) - \nabla\varphi(x_{T'_s})\right\| \le L \left\|x_T - x_{T'_s}\right\| = L \left\|\sum_{t=T}^{T'_s-1} \eta_t\alpha(x_t)\hat{g}(x_t,\theta_t)\right\|. \tag{3.26}$$

Define the error $e(x_t,\theta_t) = \hat{g}(x_t,\theta_t) - \alpha(x_t)\nabla\varphi_k(x_t)$ and use the triangle inequality to further upper bound the difference of gradients by

$$\left\|\nabla\varphi(x_T) - \nabla\varphi(x_{T'_s})\right\| \le L \left\|\sum_{t=T}^{T'_s-1} \eta_t\alpha(x_t)\nabla\varphi_k(x_t)\right\| + L \left\|\sum_{t=T}^{T'_s-1} \eta_t e(x_t,\theta_t)\right\| \tag{3.27}$$

We next show that $\sum_{t=0}^{T} \eta_t\alpha(x_t)e(x_t,\theta_t)$ is a square integrable martingale. First of all, $\{e(x_t,\theta_t)\}$ is addapted to the sequence of sigma-algebras $\{\mathcal{G}_t\}$, since $\hat{g}(x_t,\theta_t) \in \mathcal{G}_t$ for all $t \ge 0$. Next write the expectation of $\sum_{t=0}^{u} \eta_t\alpha(x_t)e(x_t,\theta_t)$ with respect to $\mathcal{G}_u$ as

$$\mathbb{E}\left[\sum_{t=0}^{u} \eta_t e(x_t,\theta_t)\Big|\mathcal{G}_u\right] = \sum_{t=0}^{u-1} \eta_t e(x_t,\theta_t) + \eta_u \mathbb{E}\left[e(x_u,\theta_u)\big|\mathcal{G}_u\right] = \sum_{t=0}^{u-1} \eta_t e(x_t,\theta_t), \tag{3.28}$$

where the first equality follows from the fact that $e(\mathbf{x}_t,\theta_t)$ is measurable with respect to $\mathcal{G}_u$ for all $t < u$ and the second one from the fact that $\hat{g}(x_t,\theta_t)$ is an unbiased estimate of $\nabla\varphi_k(x_t)$. Hence $\sum_{t=0}^{T} \eta_t\alpha(x_t)e(x_t,\theta_t)$ is a margingale. We are left to bound its second moment. To do so,take its the expectation with respect to $\mathcal{G}_u$

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{t=0}^{u} \eta_t e(x_t,\theta_t)\right\|^2 \Big|\mathcal{G}_u\right] = \left\|\sum_{t=0}^{u-1} \eta_t e(x_t,\theta_t)\right\|^2 + \eta_u^2 \mathbb{E}\left[\|e(x_u,\theta_u)\|^2\big|\mathcal{G}_u\right] \\ +2\mathbb{E}\left[\eta_u e_u^\top\big|\mathcal{G}_u\right]\sum_{s=0}^{u-1} \eta_s e_s.\end{aligned} \tag{3.29}$$

Observe that, because $\hat{g}(x_t, \theta)$ is an unbiased estimate of $\nabla\varphi_k(x_t)$ for all t the expectation in the last term is equal to zero. Using the fact, that $\hat{g}(x_t, \theta_t)$ is bounded (cf., Assumption 3) and that $\nabla\varphi_k(x)$ is bounded as well, it follows that there exists $\sigma > 0$ such that $\|e(x_t, \theta_t)\| < \sigma$ for all $t \geq 0$. Thus, the previous inequality can be upper bounded by

$$\mathbb{E}\left[\left\|\sum_{t=0}^{u} \eta_t e(x_t, \theta_t)\right\|^2 \Big| \mathcal{G}_u\right] \leq \sum_{t=0}^{u-1} \eta_t^2 \|e(x_t, \theta_t)\|^2 + \eta_u^2 \sigma^2. \tag{3.30}$$

By recursively conditioning with respect to previous sigma algebras we can upper bound the expectation of $\|\sum_{t=0}^{u} \eta_t e(x_t, \theta_t)\|^2$ by

$$\mathbb{E}\left[\left\|\sum_{t=0}^{u} \eta_t e(x_t, \theta_t)\right\|^2\right] \leq \sum_{t=0}^{u} \eta_t^2 \sigma^2. \tag{3.31}$$

The latter shows that the martingale is square integrable because η_t are square summable (cf., Assumption 5). Hence, it converges almost everywhere [29, Theorem 5.4.9.] and we can chose s large enough so $\left\|\sum_{t=T}^{T'_s} \eta_t e(x_t, \theta_t)\right\| < \frac{\delta}{6L}$. Combining this fact with the fact that for all $T_s < t < T'_s$ we have that $\|\nabla\varphi_k(x_t)\| > \delta/3$, we can upper bound (3.27) by

$$\left\|\nabla\varphi_k(x_T) - \nabla\varphi_k(x_{T'_s})\right\| \leq \frac{3L}{\delta} \sum_{t=T}^{T'_s-1} \eta_t \alpha(x_t) \|\nabla\varphi_k(x_t)\|^2 + \frac{\delta}{6}. \tag{3.32}$$

Likewise, chose s large enough so that $\sum_{t=T}^{T'_s-1} \eta_t \alpha(x_t) \|\nabla\varphi_k(x_t)\|^2 \leq \delta^2/(18L)$, then the previous expression means that

$$\left\|\nabla\varphi_k(x_T) - \nabla\varphi_k(x_{T'_s})\right\| \leq \frac{\delta}{3}. \tag{3.33}$$

Which means that $\|\nabla\varphi_k(x_T)\| < 2\delta/3$ which contradicts the fact that the limit superior is larger than zero. □

The previous lemma states that with probability one, the update (3.10) results in a sequence that converges to either the minimum of the navigation function $\varphi_k(x)$ or to one of its saddle points. In the deterministic framework, the stable manifold of the saddles has zero measure and therefore, convergence to the minimum is guaranteed for almost every initial condition. The next lemma is the stochastic counterpart of this result, where we claim that the probability of converging to a saddle is zero. We state the result in its generic form for any Morse function.

Lemma 9. *Let $V(x) : \mathcal{F} \to \mathbb{R}$ be a Morse function and let $\hat{g}(x, \theta_t)$ satisfy Assumption 3*

such that

$$\mathbb{E}\left[\hat{g}(x,\theta_t)^\top \nabla V(x)\middle|\mathcal{G}_t\right] > 0, \tag{3.34}$$

for every $x \in \mathcal{F}$ *satisfying* $\nabla V(x) \neq 0$. *Then for any* $x_0 \in \mathcal{F}$, *the probability of the sequence* $\{x_t, t \geq 0\}$, *generated by the update* (3.10), *converging to a saddle point of* $V(x)$ *is zero.*

Proof. See [101]. □

Notice that in the specific case where $V(x)$ is $\varphi_k(x)$ and $\hat{g}(x,\theta)$ is an unbiased estimator of the gradient of the navigation function, the left hand side of (3.34) yields $\alpha(x_t)\|\varphi_k(x_t)\|^2$ which is strictly positive unless x_t is a critical point of $\varphi_k(x)$. In this case, the previous lemma states that the probability of the limit of sequence $\{x_t \in \mathbb{R}^n, t \in \mathbb{N} \cup \{0\}\}$, given by the update (3.10), being a saddle point of $\varphi_k(x)$ is zero. Thus, by combining lemmas 8 and 9 we can show convergence to the minimum of the navigation function with probability one. Combining these facts with the result of Theorem 2, convergence to x^* if $f_0(x^*) = 0$ or to a point that is arbitrarily close to it if $f_0(x^*) \neq 0$ can be guaranteed with probability one. We formalize this result in the next Theorem.

Theorem 5. *Let* $\mathcal{F}$ *be the free space defined in* (2.9) *verifying Assumption 1 and let* $f_0 : \mathcal{X} \to \mathbb{R}$ *be a function satisfying Assumption 2 with minimum at* x^*. *Consider the artificial potential* $\varphi_k : \mathcal{F} \to [0,1]$ *defined in* (2.17) *and let* $\hat{g}(x_t,\theta_t)$ *be an unbiased navigable estimate of* $\nabla\varphi_k(x)$ *as per Definition 3. Also let* (2.18) *hold for all* $i = 1 \dots m$. *Let* $\{x_t, t \geq 0\}$ *be the sequence generated by the update* (3.10) *with a step size satisfying Assumption 5. Then, for every* $\varepsilon > 0$, *there exists a constant* $K(\varepsilon)$ *such that if* $k > K(\varepsilon)$, *we have that* $\{x_t, t \geq 0\} \in \mathcal{F}$ *and* $\lim_{t\to\infty} x_t = \bar{x}$ *a.e., where* $\|\bar{x} - x^*\| < \varepsilon$. *Furthermore, if* $f_0(x^*) = 0$ *it holds that* $\bar{x} = x^*$.

Proof. From Theorem 2 it follows that for every $\varepsilon > 0$ there exists some $K(\varepsilon) > 0$ such that for any $k > K(\varepsilon)$ the artificial potential $\varphi_k(x)$ is a navigation function with minimum at $\bar{x}$ satisfying $\|\bar{x} - x^*\| < \varepsilon$ if $f_0(x^*) \neq 0$ and $\bar{x} = x^*$ if $f_0(x^*) = 0$. The fact that $\{x_t, t \geq 0\} \in \mathcal{F}$ follows from Lemma 7 and the convergence to $\bar{x}$ is a consequence of lemmas 8 and 9. □

The previous theorem states that by following an unbiased estimate of the gradient of a Rimon-Koditschek navigation function, with probability one, the robot converges to a neighborhood of the minimum of the objective function. We generalize the previous result in two forms. In the following section, we consider the case of biased estimates, and in Section 3.5, we study the case of arbitrary spaces – and suitable navigation functions.

3.4 Biased Estimator

In this section, we generalize Theorem 5 to the case of biased estimators satisfying assumptions 3 and 4. The main difference is that, due to the bias, the estimate $\hat{g}(x_t, \theta_t)$ is not a descent direction in expectation for the navigation function $\varphi_k(x)$. However, it can be shown that it is a descent direction for a different Morse function whose critical points are close to those of $\varphi_k(x)$ and have the same index. We formalize this result next.

Lemma 10. *Let $\mathcal{F}$ be the free space defined in* (2.9) *verifying Assumption 1 and let $f_0 : \mathcal{X} \to \mathbb{R}$ be a function satisfying Assumption 2 with minimum at x^*. Consider the artificial potential $\varphi_k : \mathcal{F} \to [0, 1]$ defined in* (2.17) *and let $\hat{g}(x, \theta)$ be a navigable estimate of $\nabla\varphi_k(x)$ as per Definition 3. Also let* (2.18) *hold for all $i = 1 \dots m$. Then, for every $\delta > 0$ there is a constant K such that if $k > K$, there exists a twice differentiable Morse function $V_k : \mathcal{F} \to \mathbb{R}_+$ satisfying:*

(i) *All critical points $\tilde{x}_c$ of V_k are such that $\|\tilde{x}_c - x_c\| < \delta$, where x_c is a critical point of of $\varphi_k(x)$*

(ii) *$ind(x_c) = ind(\tilde{x}_c)$*

(iii) *for every x that is not a critical point of $V_k(x)$*

$$\mathbb{E}\left[\hat{g}(x, \theta_t)^\top \nabla V_k(x) \middle| \mathcal{G}_t\right] > 0, \tag{3.35}$$

Proof. See Appendix A.2.2. □

In the previous lemma we established the existence of an energy function for which the estimate of the gradient of the navigation function $\hat{g}(x_t, \theta_t)$ is a descent direction in expectation. Hence, similarly to Lemma 8 we can show that a sequence generated by a biased estimate converges to the critical points of the energy function $V_k(x)$ with probability one. Since the indices of the latter are the same as those of $\varphi_k(x)$, the convergence is, with probability one, to the unique minimum of the energy function which is arbitrarily close to that of $\varphi_k(x)$. This is the subject of the following theorem.

Theorem 6. *Let $\mathcal{F}$ be the free space defined in* (2.9) *verifying Assumption 1 and let $f_0 : \mathcal{X} \to \mathbb{R}$ be a function satisfying Assumption 2 with minimum at x^*. Consider the artificial potential $\varphi_k : \mathcal{F} \to [0, 1]$ defined in* (2.17) *and let $\hat{g}(x, \theta)$ be a navigable estimate of $\nabla\varphi_k(x)$ as per Definition 3. Also let* (2.18) *hold for all $i = 1 \dots m$. Let $\{x_t, t \geq 0\}$ be the sequence generated by the update* (3.10) *with a step size satisfying Assumption 5. Then for every $\varepsilon > 0$, there exists a constant $K(\varepsilon) > 0$ such that if $k > K(\varepsilon)$, we have that $\{x_t, t \geq 0\} \in \mathcal{F}$*

and that exists $\bar{x} \in \mathcal{F}$ *satisfying* $\|\bar{x} - x^*\| < \varepsilon$ *and*

$$\lim_{t\to\infty} x_t = \bar{x} \quad a.e. \tag{3.36}$$

Proof. Observe that non collision is ensured by virtue of Lemma 7. Notice that Lemma 10 ensures existence of an energy function such that its critical points are arbitrarily close to those of $\varphi_k(x)$ and the indices of said critical points are the same for both functions. Since for k large enough $\varphi_k(x)$ is a navigation function (cf., Theorem 2), $V_k(x)$ has only one minimum at $\bar{x}$ that satisfies $\|\bar{x} - x^*\| < \varepsilon$ and the other critical points are nondegenerate saddles (cf., Lemma 10). Hence, lemmas 8 and 9 ensure convergence to $\bar{x}$ with probability one. □

The above theorem states that under the same conditions on the free space and the objective function than in the deterministic case, by following the update (3.10) the agent is able, with probability one, to reach a point arbitrarily close to the minimum of the objective function $f_0(x)$ without running into the free space boundary. The main advantage as compared to the results presented in Chapter 2 is that it allows the agents to perform an update using only local information. The mismatch between the real world and the local estimator, based on a given belief that the agent has, may result in a biased estimator. Yet, the bias does not affect the qualitative behavior of the agent. Furthermore, instead of requiring exact information about both the objective function and the obstacles, stochastic measurements suffice to solve the problem of interest.

A second difference between the results in Theorem 2 – complete and deterministic information– and Theorems 5 and 6 – local and stochastic information– is in the sense in which the navigation is almost surely. In the deterministic case, this means that except for a set of initial configurations of measure zero –the stable manifold of the saddle points of $\varphi_k(x)$ – the solutions of the dynamical system $\dot{x} = -\nabla\varphi_k(x)$ converge to the minimum of the objective function; while in the stochastic case the goal is achieved with probability one for every initial position.

3.5 Alternative Artificial Potentials

Throughout this chapter, we focused on navigation functions that are of the Rimon-Koditschek form, however the results here presented can be generalized to any artificial potentials, hence extending the stochastic navigation framework to more complex spaces. For instance, if one can construct an unbiased estimate of harmonic navigation functions, then navigating topologically complex spaces [77, 78] becomes possible with noisy information. The following corollary generalizes this result.

Corollary 3. *Let $\mathcal{F}$ be a free space and let $\varphi : \mathcal{F} \to [0,1]$ be a navigation function (cf., Definition 1) with minimum at the agent's goal x^*. Let $\hat{g}(x_t, \theta_t)$ be an unbiased estimate of the gradient of the navigation function satisfying Assumption 3. Then the update rule* (3.10) *generates a sequence $\{x_t, t \geq 0\} \in \mathcal{F}$ such that $\lim_{t\to\infty} x_t = x^*$.*

Proof. The non-collision proof is a direct consequence of 7 and the convergence to the minimum of the navigation function follows from lemmas 8 and 9. Observe that these do not depend on the specific form of the free space nor the navigation function selected. □

The previous result generalizes Theorem 5 for any free space geometry. This is, by following the negative direction of an unbiased stochastic gradient of a suitable navigation function, one can succeed in navigating towards the minimum of the objective function without running into the free space boundary. If the estimates are biased similar guarantees could be proved but the form of the assumption about the bias (Assumption 4) should be adapted to the specific navigation function, since the ones considered here, are highly related to Rimon-Koditschek potentials. However, with the same assumptions we can extend the result of Theorem 6 for a different class of artificial potentials, that of logarithmic barriers. Inspired in the optimization literature we define the following barrier function

$$\phi_k(x) = f_0(x) - \frac{1}{k}\log(\beta(x)). \tag{3.37}$$

The previous potential is not a navigation function since it is not bounded and it is not defined in the boundary of the free space. However its supremum is at the boundary of the free space and it is possible to show that all the critical points of the previous equation are nondegenerate and it has a unique minimum. Differentiate (3.37) to get

$$\nabla\phi_k(x) = \nabla f_0(x) - \frac{\nabla\beta(x)}{k\beta(x)}. \tag{3.38}$$

Observe that the previous expression is similar to that of the direction of the gradient of the Rimon-Koditschek artificial potential. In particular, the same fundamental properties of the critical points hold, i.e., nondegeneracy and presence of a unique minimum follow from analogous proofs to those in Chapter 2. Since $\nabla\beta(x)$ is not zero in the boundary of the free space (see proof of Lemma 2) the critical points can be pushed by increasing k either arbitrarily close to the minimum of $f_0(x)$ or arbitrarily close to $\beta(x)$. In particular, the first one can be showed to be a unique local minimum and the rest to be saddle points. Furthermore the eigenvalues of the Hessian of these critical points depend on k with the same order as in the case of Rimon-Koditschek artificial potentials. Hence, Assumptions 3

and 4 are appropriate in this case too for the following estimate of the gradient of $\phi_k(x)$

$$\hat{g}(x_t, \theta) = \hat{\beta}(x_t, \theta_t)\hat{\nabla} f_0(x_t, \theta_t) - \frac{\hat{\nabla}\beta(x_t, \theta_t)}{k}. \tag{3.39}$$

Hence by following the negative direction of the gradient of $\phi_k(x)$ the agent converges to a point arbitrarily close to the minimum of $f_0(x)$. We formally state this result next.

Theorem 7. *Let $\mathcal{F}$ be the free space defined in* (2.9) *verifying Assumption 1 and let $f_0 : \mathcal{X} \to \mathbb{R}$ be a function satisfying Assumption 2 with minimum at x^*. Consider the artificial potential $\phi_k : \mathcal{F} \to \mathbb{R}$ defined in* (3.37) *and let $\hat{g}(x_t, \theta_t)$, the estimate defined in* (3.39) *be navigable as per Definition 3. Also let* (2.18) *hold for all $i = 1 \dots m$. Let $\{x_t, t \geq 0\}$ be the sequence generated by the update* (3.10) *with a step size satisfying Assumption 5. Then for every $\varepsilon > 0$, there exists a constant $K(\varepsilon)$ such that if $k > K(\varepsilon)$, we have that $\{x_t, t \geq 0\} \in \mathcal{F}$ and*

$$\lim_{t\to\infty} x_t = \bar{x} \quad a.e., \tag{3.40}$$

where $\bar{x}$ is satisfies $\|\bar{x} - x^*\| < \varepsilon$.

Proof. Observe that non-collision is ensured by virtue of Lemma 7. The proof that the critical points of $\phi_k(x)$ are nondegenerate and that only one of them is a minimum and it can be pushed arbitrarily close to the minimum of $f_0(x)$, is analogous to that of Lemmas 2–6. Hence by virtue of Lemma 10, there exists an energy function such that its critical points are arbitrarily close to those of $\phi_k(x)$ and the indexes of said critical points are the same for both functions. Thus Lemma 8 holds for the energy function. The proof is completed by virtue of Lemma 9 and because all critical points but one are nondegenerate saddles for large enough k. □

The previous Theorem extends the result of the biased estimate of the Rimon-Koditschek navigation function to a new class of artificial potentials under the same conditions over the geometry of the free space and the bias. In the next section we study the implications of Theorems 6 and 7 numerically.

3.6 Numerical Examples

We evaluate the performance of the local stochastic approximation of the gradient of the Rimon-Koditschek potential in two different scenarios in which condition (2.18) is satisfied. Each obstacle is estimated as the osculating circle at the closest point of the obstacle from the agent's position as in Appendix A.2.1. In sections 3.6.1 and 3.6.2 we consider ellipsoidal and egg-shaped obstacles respectively. The performance of the local approximation of the

logarithmic barrier (3.37) is evaluated in Section 3.6.3. In all three cases, the external boundary of the free space is a spherical shell of center c_0 and radius r_0.

3.6.1 Elliptical obstacles

In this section we consider m elliptical obstacles in $\mathbb{R}^2$. For $i = 1 \dots m$, let $A_i \in \mathcal{M}^{2\times 2}$ be symmetric and positive definite matrices, and let $\mu^i_{\min} > 0$ be the minimum eigenvalue of matrix A_i and define the following obstacle functions

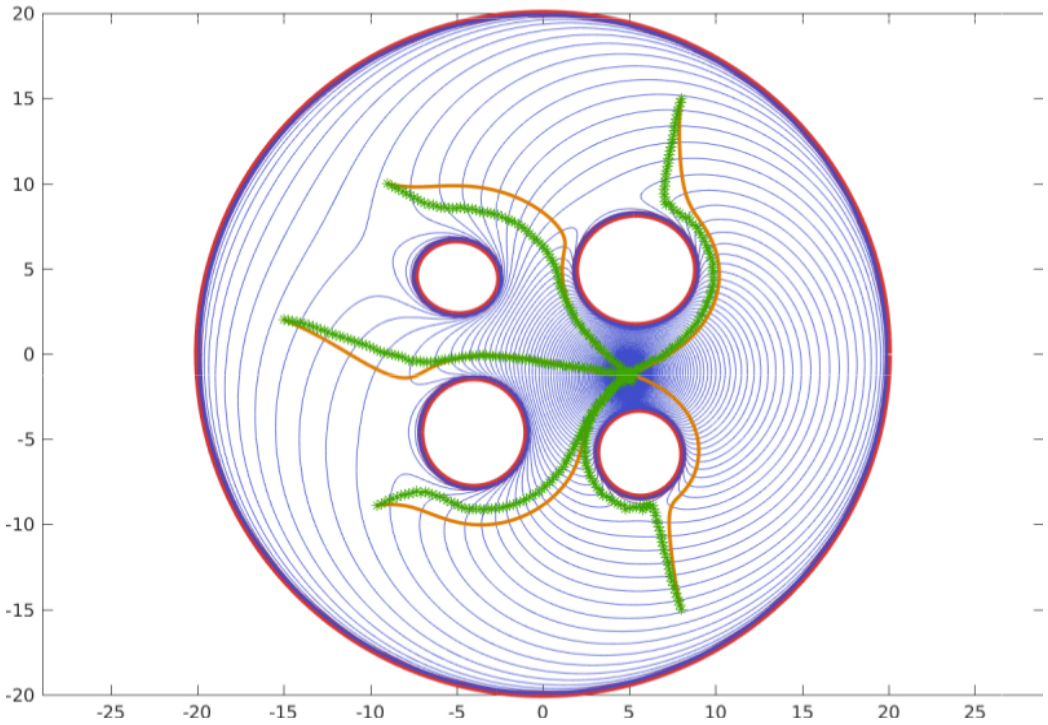
$$\beta_i(x) = (x - c_i)^\top A_i (x - c_i) - \mu^i_{\min} r_i^2. \tag{3.41}$$

where $c_i \in \mathcal{X}$ is the center of the i-th ellipse and $r_i > 0$ is the length of its largest axis. Each obstacle is then defined as

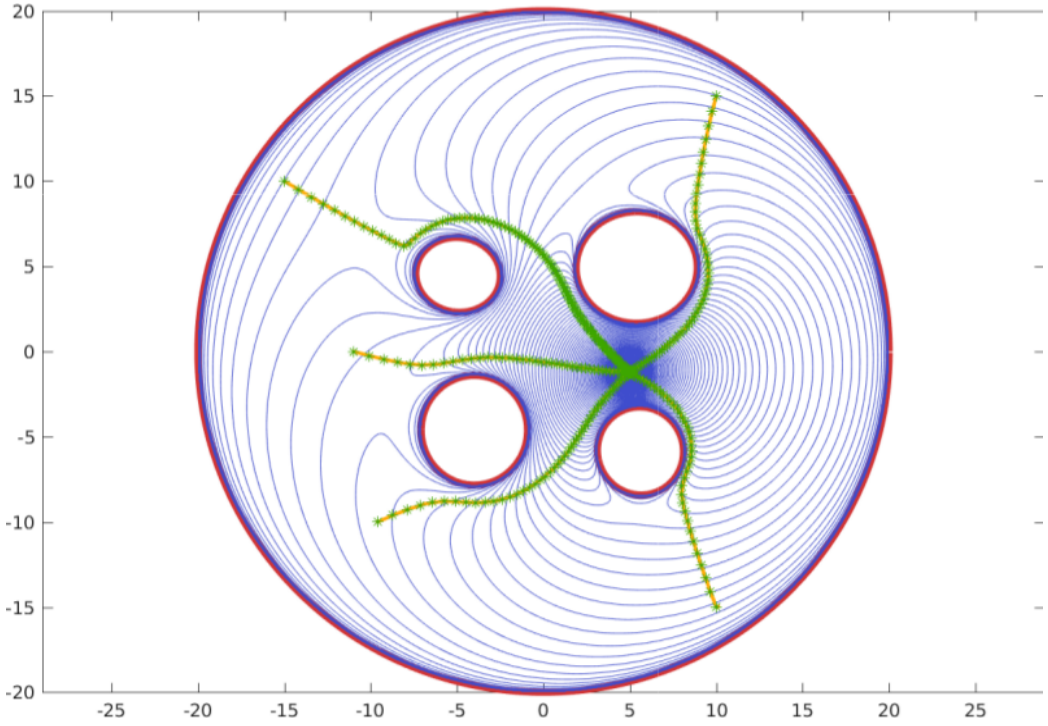
$$\mathcal{O}_i = \left\{ x \in \mathcal{X} \middle| \beta_i(x) < 0 \right\}. \tag{3.42}$$

In these experiments we place the center of each ellipsoid in a different orthant. In particular, each center is set to be in the position $L(\pm 1, \pm 1)$ and then we add a random variation drawn uniformly from $[-\Delta, \Delta]^2$, where $0 < \Delta < L$. The maximum axis of the ellipse – r_i– is drawn uniformly from $[r_0/10, r_0/5]$ and the matrices A_i for $i = 1 \dots m$ are such that they are orthogonal and their eigenvalues are random and uniformly selected from the interval $[1, 2]$. We verify that the obstacles resulting of the previous process do not intersect. If they do, we re draw all previous parameters. For the objective function we consider a quadratic cost given by $f_0(x) = (x - x^*)^\top Q(x - x^*)$, where x^* is drawn uniformly over $[-r_0/2, r_0/2]^2$ and we verify that it is in the free space. The matrix $Q \in \mathcal{M}^{2\times 2}$ is a random positive definite symmetric matrix whose eigenvalues are selected as follows. For each obstacle we compute the maximum condition number that Q could have in order to satisfy condition (2.18). Let N_{cond} be the maximum among these admissible condition numbers. Then, the eigenvalues of Q are selected randomly from $[1, N_{cond} - 1]$, hence ensuring that (2.18) is satisfied. For the estimates of the objective function, its gradient, the distance to the obstacles, the direction defined by the position of the agent and its projection onto the obstacles and their curvature we consider independent gaussian additive noise with mean zero and standard deviation σ_q. The step size selected for the update (3.10) is of the form $\eta_t = \eta_0/(1 + \zeta t)$ and the initial position is selected randomly over $[-r_0, r_0]^2$.

For this experiment we set the parameters to be $c_0 = 0$, $r_0 = 20$, $L = 6$, $\Delta = 1$, $\sigma_{f_0} = \sigma_{\nabla f_0} = 1$ and $\sigma_{d_i} = \sigma_{R_i} = \sigma_{\mathbf{n}_i} = d_i(x)/10$. The selection of a variance that depends on the the distance is done so to ensure that the closer the agent is to the boundary of the free space the better the estimation of the obstacle is. In particular, at the boundary we have that $\sigma_{d_i} = \sigma_{R_i} = \sigma_{\mathbf{n}_i} = 0$. We set the constant at which the agent is able to



(a) Trajectories with $k = 7$



(b) Trajectories with $k = 12$.

Figure 3.1: The trajectories resulting of the navigation function approach – solid line– and its stochastic approximation given in (3.10) –stars–succeed in driving the agent to the goal configuration for five different initial positions as expected in virtue of Theorem 6. We observe that for the same world (cf., Figures 3.1(a) and 3.1(b)) the larger the order parameter k is, the closer the trajectory resulting from stochastic approximation is to the trajectory resulting of descending along the gradient of the navigation function (2.17).

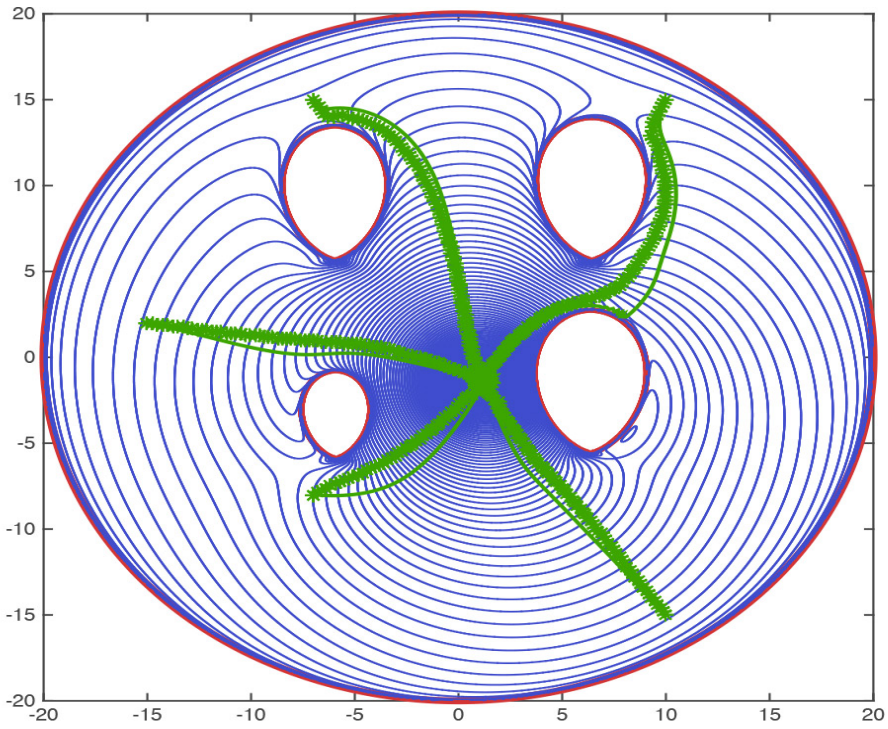
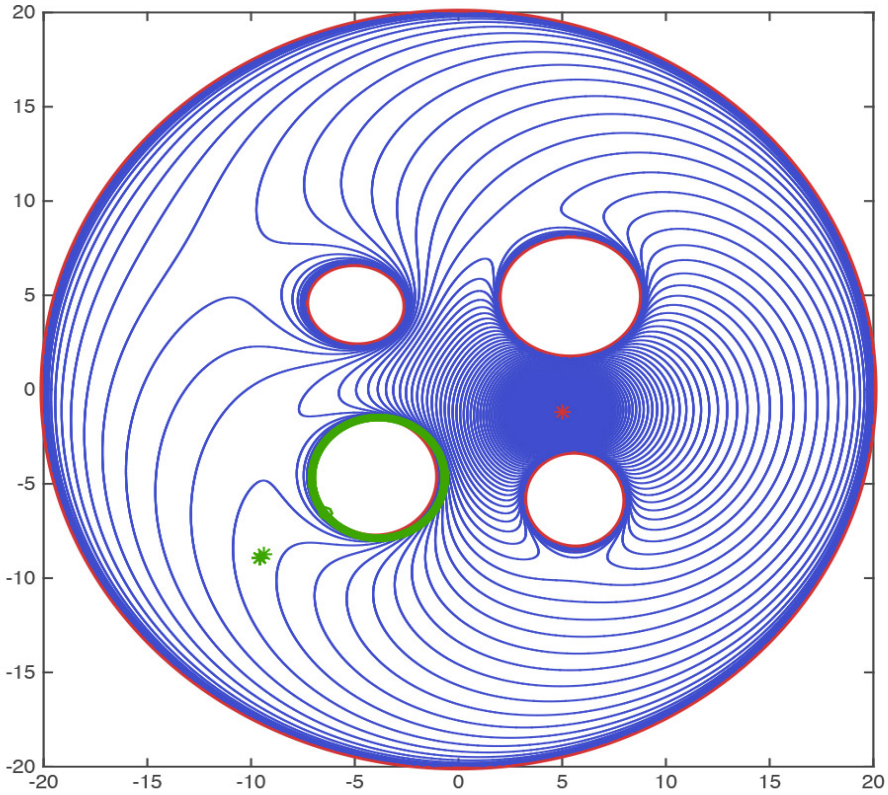
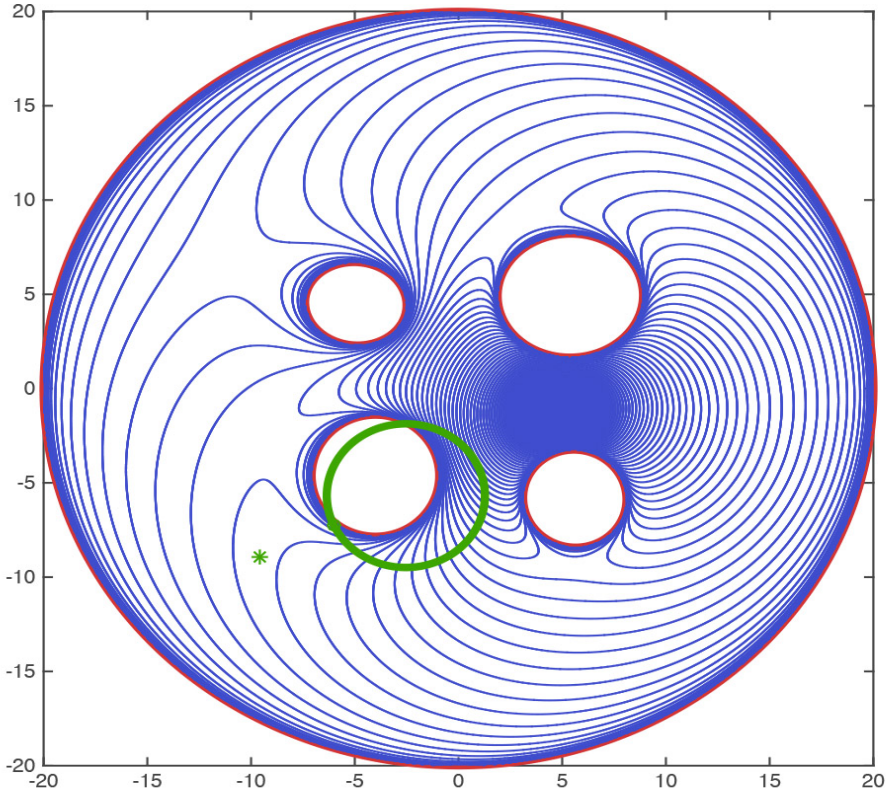


Figure 3.2: The trajectories resulting of the navigation function approach with $k = 15$ – solid line– and its stochastic approximation given in (3.10) –stars–succeed in driving the agent to the goal configuration for five different initial positions as expected in virtue of Theorem 6.



(a) Local estimation of the obstacle with perfect measures.



(b) Stochastic estimation of the obstacle.

Figure 3.3: Estimation of the obstacles by the hallucinated osculating circle for a particular position in the free space with exact and stochastic information. Obstacles are sensed if $d_i(x) < 7$. Noise is Gaussian, additive, mean zero and with variance $\sigma_{d_i} = \sigma_{R_i} = \sigma_{\mathbf{n}_i} = d_i(x)/10$.

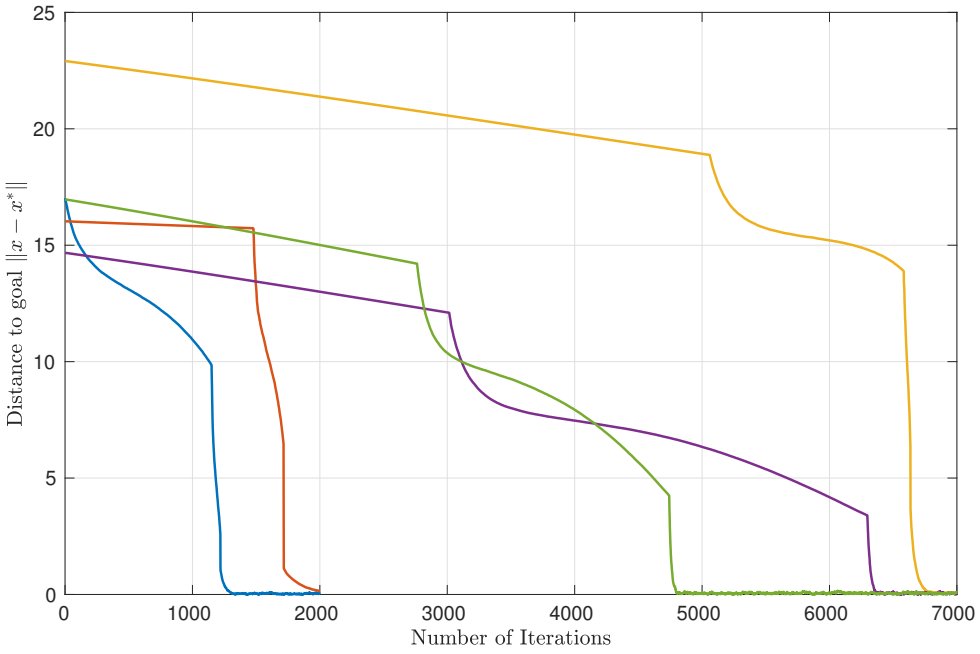


Figure 3.4: Evolution of the distance to the goal in a world with elliptical obstacles. We set the order parameter of the navigation function to $k = 12$, and the step size to satisfy Assumption 5 with the following parameters $\eta_0 = 1 \times 10^{-7}$, $\zeta = 5 \times 10^{-5}$.

measure an obstacle [cf., (A.46)] to be $c = 7$. Finally, the parameters of the step size are $\eta_0 = 5 \times 10^{-2}$ and $\zeta = 5 \times 10^{-3}$ and we run each simulation 100 steps with a normalized estimate.

In Figure 3.1 we observe the behavior of the system that follows the local and stochastic update (3.10) – marked with stars – and that of the system following the gradient dynamical system $\dot{x} = -\nabla\varphi_k(x)$– solid lines – for five different initial conditions. In Figure 3.1(a) the order parameter is set to be $k = 7$ while in 3.1(b) it is set to be 12. In both cases it can be observed that the local and stochastic update succeeds in generating a sequence that remains in the free space and that converges to the minimum of the objective function. It is also observed that the direction in which the agent moves while following the local update differs from that of the agent following the gradient of the navigation function. This result is not surprising in virtue of the fact that as discussed in Section 2.2 the model selected results in a biased estimate of the gradient of the navigation function. However notice that by increasing k the two trajectories become closer to each other. This effect can be observed by comparing the trajectories depicted in figures 3.1(a) and 3.1(b) where the order parameter k is set to be 7 and 12 respectively. This result is expected because the norm of the bias decreases with $1/k$. This is an Assumption in Section 3.2.1 but in Appendix A.2.1 we show that it is indeed the case for circular estimates of the obstacles. In particular by selecting k large enough the bias could be reduced arbitrarily. Another effect of having larger k is that of diminishing the relative weight of the $\nabla\beta(x)$ as compared to $\nabla f_0(x)$ in the gradient of the navigation function. Hence in a sense having large k is equivalent to follow only the direction $-\nabla f_0(x)$ and neglect the obstacles. Thus yielding shorter paths. Since in the stochastic approximation we only consider nearby obstacles a similar effect is expected. This is what we observed in Figure 3.1(b).

The effect of the standard deviations of the noise in the estimation of the obstacles is illustrated in Figure 3.3. In particular, for the initial position of one of the trajectories depicted in Figure 3.1(a) we observe the estimation of the closest obstacle to that position in the noiseless case 3.3(a) and the estimate with noise 3.3(b). The fact that even for noiseless cases the estimation is not perfect is what yields a biased estimate.

In Figure 3.4 we consider the evolution of the distance between the agent and the destination with $k = 12$ for the same five initial conditions than in figures 3.1(a) and 3.1(b). For this simulation, we do not consider a normalized gradient and we take smaller step sizes. In particular we set $\eta_0 = 1 \times 10^{-7}$, $\zeta = 5 \times 10^{-5}$. Observe that the speed at which the agent advances differs considerably depending on its position. The main reason for this to happen is that the number of obstacles considered for the estimate is not constant and in depends on the position of the agent. This results in a piece-wise continuous scaling $\alpha(x)$ with large differences of its value at the points of discontinuity (cf., (3.1)).

3.6.2 Egg shaped world obstacles

In this section we consider egg shaped obstacles as an example of convex obstacles other than ellipses. We draw the center of each obstacle, c_i, from a uniform distribution over $[-L/2, L/2] \times [-L/2, L/2]$. The distance between the "tip" and the "bottom" of the egg, r_i, is drawn uniformly over $[r_0/10; r_0/5]$ and with equal probability the egg is horizontal or vertical. The obstacle being horizontal translates into the fact that the function $\beta_i(x)$ representing the obstacle takes the following form

$$\beta_i(x) = \|x - c_i\|^4 - 2r_i \left(x^{(1)} - c_i^{(1)}\right)^3, \tag{3.43}$$

where the superscript (1) refers to first component of a vector. Likewise, for vertical eggs the function $\beta_i(x)$ takes the form

$$\beta_i(x) = \|x - c_i\|^4 - 2r_i \left(x^{(2)} - x_c^{(2)}\right)^3. \tag{3.44}$$

Notice that the functions β_i as defined above are not convex on $\mathbb{R}^2$, however since their Hessians are positive definite outside the obstacles one could define a convex extension of β_i inside the obstacles. Yet, this is not needed because the agent operates in the free space. In particular, for this experiment we set $r_0 = 20$ and $L = 6$. The selection of the noises standard deviations σ_q and the distance at which the obstacles can be measured are the same as in Section 3.6.1.

In Figure 3.2 we observe the level sets of the navigation function (2.17) and the trajectories resulting from the stochastic approximation (3.10) –marked with stars– and from following the negative gradient of the navigation function for $k = 15$. It can be observed that the update (3.10) succeeds in driving the agent to the goal configuration given by the minimum of the objective function $f_0(x)$ while remaining in the free space at all times.

3.6.3 Logarithmic barrier

In this section we evaluate the performance of the descent along the direction of the negative gradient of the logarithmic barrier artificial potential in (3.39). For these experiments the obstacles and the boundary of the workspace are selected as in Section 3.6.1 with the following values of the parameters $c_0 = 0$, $r_0 = 20$, $L = 6$, $\Delta = 1$, $\sigma_{f_0} = \sigma_{\nabla f_0} = 1$, $\sigma_{d_i} = \sigma_{R_i} = \sigma_{\mathbf{n}_i} = d_i(x)/10$ and $k = 10$. In Figure 3.5 we depict the trajectory of an agent starting at different initial positions. As it can be observed the agent succeeds in reaching the minimum of the objective function $f_0(x)$ while avoiding the obstacles. By comparing these trajectories to those in figures 3.1(a) and 3.1(b) –coming from Rimon-Koditschek potentials– we observe that the logarithmic barrier artificial potential results in paths that

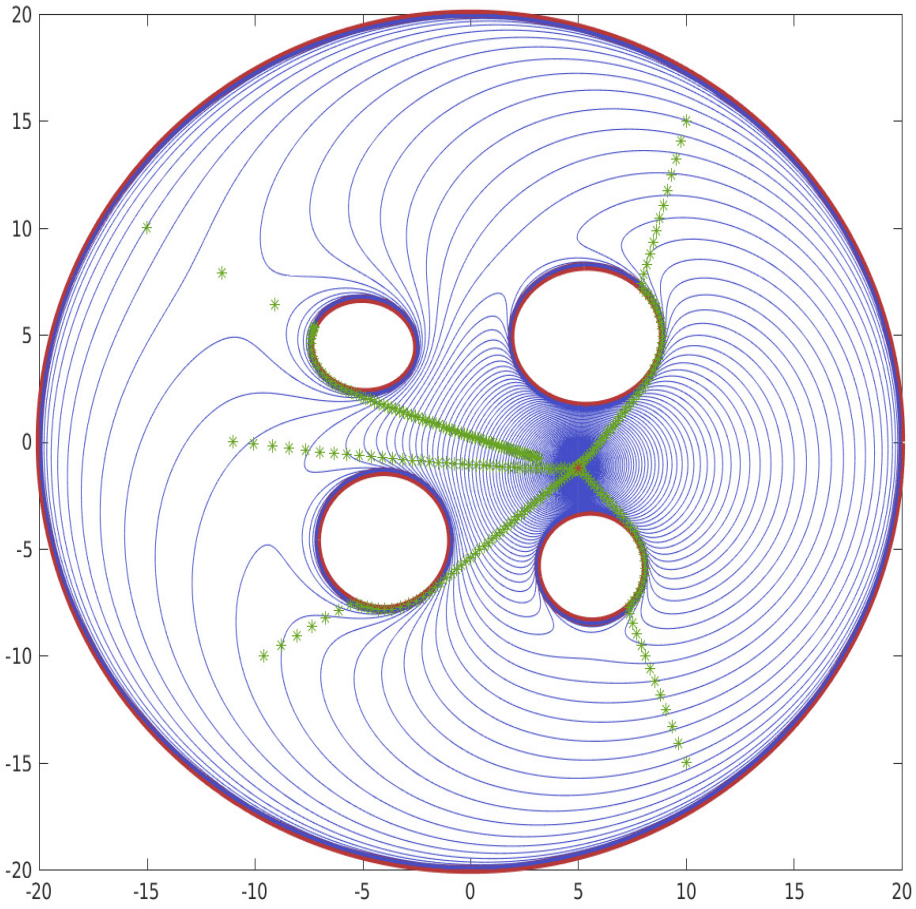


Figure 3.5: Trajectories resulting of following the negative gradient of the logarithmic barrier given in (3.37) for $k = 10$ in an elliptical world. The trajectories resulting from the update (3.10) succeed in driving the agent to the goal configuration for five different initial positions as expected in virtue of Theorem 7.

pass closer to the obstacles.

3.7 Conclusions

We considered a set with convex holes in which an agent must navigate to the minimum of a convex function. The objective function and the obstacles are unknown to the agent and the only information available to him about these is gathered through sensors. Thus, making the available information local and stochastic. With this information at hand, the robot is able to construct an estimate of the gradient of a navigation function of the Rimon-Koditschek form. In the case where the the full gradient of the navigation function can be constructed without noise it has been shown that; by following said gradient a robot can converge to the desired goal while avoiding the obstacles present in the workspace as

long as the following conditions are satisfied. (i) The obstacles are not too flat and the condition number of the Hessian of the convex potential is not large. (ii) The distance from the obstacles' boundary to the minimum of the convex potential is large relative to the size of the obstacles.

We show that, by following the negative of the estimate of the gradient of the navigation function and under the same conditions than in the deterministic case – even when the estimate constructed is biased – the agent succeeds in avoiding the obstacles and in converging to a arbitrarily small neighborhood of the goal with probability one. The origin of the bias is in the mismatch between the real world and the belief the agent has about it, in particular navigation is possible as long as the bias is small as compared to the gradient of the navigation function. We extend the previous result to the case of an artificial potential based on a logarithmic barrier and to arbitrary geometries of the free space and suitable navigation functions as long as the estimates are unbiased. Numerical experiments support the theoretical results.

Chapter 4

Online Learning of Feasible Strategies

Define an environment as a set of convex constraint functions that vary arbitrarily over time and consider a cost function that is also convex and arbitrarily varying. Agents that operate in this environment intend to select actions that are feasible for all times while minimizing the cost's time average. Such action is said optimal and can be computed *offline* if the cost and the environment are known a priori. An *online* policy is one that depends causally on the cost and the environment. To compare online policies to the optimal offline action define the fit of a trajectory as a vector that integrates the constraint violations over time and its regret as the cost difference with the optimal action accumulated over time. Fit measures the extent to which an online policy succeeds in learning feasible actions while regret measures its success in learning optimal actions. In this chapter we learn online policies computed from a saddle point controller which are shown to have fit and regret that are either bounded or grow at a sublinear rate. These properties provide an indication that the controller finds trajectories that are feasible and optimal in a relaxed sense. Concepts are illustrated throughout with the problem of a shepherd that wants to stay close to all sheep in a herd. Numerical experiments show that the saddle point controller allows the shepherd to do so.

4.1 Introduction

In this chapter the objective is for an agent to succeed in adapting to a time varying convex environment defined as a set of convex constraints that an agent must satisfy at all times. The constraints are unknown a priori, vary arbitrarily in time in a possibly discontinuous manner, and are observed locally in space and causally in time. The goal of the agent is to find a feasible strategy that satisfies all of these constraints. This chapter shows that

an online version of the saddle point algorithm of Arrow and Hurwicz [4] executed by the agent succeeds in finding such a strategy. If the agent wants to further minimize a convex cost, we show that the same algorithm succeeds in finding an strategy that is feasible at all times and optimal on average.

To understand the contribution presented in this chapter it is important to observe that the navigation problem outlined above can be mathematically formulated as the solution of a convex program whose solution is progressively more challenging when we progress from deterministic settings to stochastic and online settings. Indeed, in a deterministic setting the cost and constraints are fixed. This yields a canonical convex optimization problem that can be solved with extremum seeking controllers based on gradient descent [3, 43, 63, 123], primal-dual methods [4, 32, 84, 87, 127], or interior point methods [16, Chapter 11]. In a stochastic setting cost and constraints vary randomly according to a stationary distribution. The agent's goal is then expressed as the selection of an action that minimizes the expected value of the objective function while satisfying constraints in an average sense [8,9,76]. This problem is more complicated than its deterministic counterpart but it can be solved using, e.g., stochastic gradient descent [58, 107, 111].

Here, we consider online formulations in which cost and constraints can vary arbitrarily, perhaps strategically, and where the goal is to find an action that is good on average and that satisfies the constraints at all times – assuming such an action exists, which, when functions change strategically, restricts adversarial actions. In this case, *unconstrained* cost minimization can be formulated in the language of regret [14, 113, 128] whereby agents operate online by selecting plays that incur a cost selected by nature. The cost functions are revealed to the agent ex post and used to adapt subsequent plays. The goodness of these *online* policies are determined by comparing to the optimal action chosen *offline* by a clairvoyant agent that has prescient access to the cost. Regret is defined as the difference of the accumulated cost attained online and the optimal offline cost. It is a remarkable fact that an online version of gradient descent is able to find plays whose regret grows at a sublinear rate when the cost is a convex function [42, 138] – therefore suggesting vanishing per-play penalties of online plays with respect to the clairvoyant play.

The constrained optimization equivalent of gradient descent is the saddle point method applied to the determination of a saddle point of the Lagrangian function [4]. This method interprets each constraint as a separate potential and descends on a linear combination of their gradients. The coefficients of this linear combination are multipliers that adapt dynamically so as to push the agent to the optimal solution in the feasible region. Saddle point algorithms and variations have been widely studied [18, 32, 84, 87, 127] and used in various domains such as decentralized control [21, 79], power systems [72, 137] and image processing, see e.g. [17]. Our observation is that since an online version of gradient descent

succeeds in achieving small regret, it is not unreasonable to expect an online saddle point method to succeed in finding feasible actions with small regret. Inded in [83] an agent that is subject to constraints that do not evolve over time is able to find trajectories with sublinear regret and constraint violation by considering a saddle point algorithm of the augmented Lagrangian.

The main contribution of this chapter is to prove that the latter holds as well when constraints evolve over time. We show that an online saddle point algorithm that observes costs and constraints ex post succeeds in finding policies that are feasible and have small regret. Central to this development is the definition of a viable environment as one in which there exist an action that satisfies the time varying constraints at all times and the introduction of the notion of fit (Section 4.2). The latter is defined as a vector that contains the time integrals of the constraints evaluated across the trajectory and is the analogous of regret for the satisfaction of constraints. In the same way in which the accumulated payoff of the online trajectory is compared with the payoff of the offline trajectory, fit compares the accumulation of the constraints along the trajectory with the feasibility of an offline viable strategy. As such, a trajectory can achieve small fit by becoming feasible at all times or by alternating periods in which the constraints are violated with periods in which the constraints are satisfied with slack. This notion of fit is appropriate for constraints that have a cumulative nature. For cases where this is not appropriate we introduce the notion of saturated fit in which only violations of the constraint are accumulated. A trajectory with small saturated fit is one in which the constraints are violated by a significant amount only for a short period of time.

Technical developments begin with the derivation of a projected gradient controller to limit the growth of regret in an environment without constraints (Section 4.3). The purpose of this section is to introduce tools and to clarify connections with existing literature in discrete time [42,138] and continuous time regret minimization [67,116,130]. An important conclusion here is that regret in continuous time can be bounded by a constant that is independent of the time horizon, as opposed to the sublinear growth that is observed in discrete time.

We then move onto the main part of the chapter in which we propose to control fit and regret growth with the use of an online saddle point controller that moves along a linear combination of the negative gradients of the instantaneous constraints and the objective function. The coefficients of this linear combination are adapted dynamically as per the instantaneous constraint functions (Section 4.4). This online saddle point controller is a generalization of (offline) saddle point in the same sense that an online gradient controller generalizes (offline) gradient descent. We show that if there exists an action that satisfies the environmental constraints at all times, the online saddle point controller achieves bounded

fit if optimality is not of interest (Theorem 9). When optimality is considered, the controller achieves bounded regret and a fit that grows sublinearly with the time horizon (Theorem 10). Analogous results are derived for saturated fit. I.e., it is bounded by a constant when optimality is not of interest and grows sublinearly otherwise (corollaries 5 and 6). Throughout this chapter we illustrate concepts with the problem of a shepherd that has to stay close to his herd (Section 4.2.2). A numerical analysis of this problem closes the chapter (Section 4.5) except for concluding remarks (Section 4.6).

4.2 Viability, feasibility and optimality

We consider a continuous time environment in which an agent selects actions that result in a time varying set of penalties. Use t to denote time and let $X \subseteq \mathbb{R}^n$ be a closed convex set from which the agent selects action $x \in X$. The penalties incurred at time t for selected action x are given by the value $f(t,x)$ of the vector function $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^m$. We interpret the vector penalty function f as a definition of the environment. Our interest is in situations where the agent is faced with an environment f and must choose an action $x \in X$ – or perhaps a trajectory $x(t)$ – that guarantees nonpositive penalties $f(t,x(t)) \preceq 0$ for all times t not exceeding a time horizon T. Since the existence of this trajectory depends on the specific environment we define a viable environment as one in which it is possible to select an action with nonpositive penalty for times $0 \leq t \leq T$ as we formally specify next.

Definition 4 (**Viable environment**). *We say that an environment $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^m$ is viable over the time horizon T for an agent that selects actions $x \in X$ if there exists a feasible action $x^\dagger \in X$ such that*

$$f(t,x^\dagger) \leq 0, \quad \textit{for all } t \in [0,T]. \tag{4.1}$$

The set $X^\dagger := \{x^\dagger \in X : f(t,x^\dagger) \preceq 0, \textit{ for all } t \in [0,T]\}$ is termed the feasible set of actions.

Since for a viable environment it is possible to have multiple feasible actions it is desirable to select one that is optimal with respect to some criterion of interest. Introduce then the objective function $f_0 : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$, where for a given time $t \in [0,T]$ and action $x \in X$ the agent suffers a loss $f_0(t,x)$. The optimal action is defined as the one that minimizes the accumulated loss $\int_0^T f_0(t,x)\,dt$ among all viable actions, i.e.,

$$\begin{aligned} x^* := \operatorname*{argmin}_{x \in X} & \int_0^T f_0(t,x)\,dt \\ \text{s.t. } & f(t,x) \preceq 0, \text{ for all } t \in [0,T]. \end{aligned} \tag{4.2}$$

For the definition in (4.2) to be valid the function $f_0(t,x)$ has to be integrable with respect

to t. In subsequent definitions and analyses we also require integrability of the environment f as well as convexity with respect to x as we formally state next.

AS6. *The functions $f(t,x)$ and $f_0(t,x)$ are integrable with respect to t in the interval $[0,T]$.*

AS7. *The functions $f(t,x)$ and $f_0(t,x)$ are convex with respect to x for all times $t \in [0,T]$.*

If the environment $f(t,x)$ and functions $f_0(t,x)$ are known beforehand, finding the action in a viable environment that minimizes the total aggregate cost is equivalent to solving the convex optimization problem in (4.2) for which a number of algorithms are known. Here, we consider the problem of adapting a strategy $x(t)$ when the functions $f(t,x)$ and $f_0(t,x)$ are *arbitrary* and *revealed causally.* I.e., we want to choose the action $x(t)$ using observations of viability $f(t,x)$ and cost $f_0(t,x)$ in the open interval $[0,t)$. This implies that $f(t,x(t))$ and $f_0(t,x(t))$ are not observed before choosing $x(t)$. The action $x(t)$ is chosen ex ante and the corresponding viability $f(t,x(t))$ and cost $f_0(t,x(t))$ are incurred ex post. Further observe that the constraints and objective functions may change abruptly if the number of discontinuities in these are finite for finite T. This makes the problem different from time varying optimization in which the goal is to track the optimal argument of $f_0(t,x)$ subject to the constraint $f(t,x) \leq 0$ under the assumption that these functions change continuously and at a sufficiently small rate [30, 31, 95, 103, 134].

4.2.1 Regret and fit

We evaluate the performance of trajectories $x(t)$ through the concepts of regret and fit. To define regret we compare the accumulated cost $\int_0^T f_0(t,x(t))\,dt$ incurred by $x(t)$ with the cost incurred by the optimal action x^* defined in (4.2),

$$\mathcal{R}_T := \int_0^T f_0(t,x(t))\,dt - \int_0^T f_0(t,x^*)\,dt. \tag{4.3}$$

Analogously, we define the fit of the trajectory $x(t)$ as the accumulated penalties $f(t,x(t))$ incurred for times $t \in [0,T]$,

$$\mathcal{F}_T := \int_0^T f(t,x(t))\,dt. \tag{4.4}$$

The regret $\mathcal{R}_T$ and fit $\mathcal{F}_T$ can be interpreted as performance losses associated with online causal operation as opposed to offline clairvoyant operation. If $\mathcal{F}_T$ is positive in a viable environment we are in a situation in which, had the environment be known a priori, we could have selected an action $x^\dagger$ with $f(t,x^\dagger) \leq 0$. The fit measures how far the trajectory $x(t)$ comes from achieving that goal. As in the case of the fit, if the regret $\mathcal{R}_T$ is large we are in a situation in which prior knowledge of environment and cost would had resulted in

the selection of the action x^* – and in that sense $\mathcal{R}_T$ indicates how much we regret not having had that information available.

Because of the cumulative nature of fit, it is possible to achieve small fit by alternating between actions for which the constraint functions take positive and negative values. This is valid when cumulative constraints are an appropriate model, which happens for quantities that can be stored or preserved in some sense – such as energy budgets enforced through average power constraints. For situations where this is not appropriate, we define the saturated fit in which constraint slacks are saturated to a small constant δ. Formally, let $\delta > 0$ be a positive constant and define the function $\bar{f}_\delta(t,x)) = \max\{f(t,x), -\delta\}$. Then, the δ-saturated fit is defined as

$$\bar{\mathcal{F}}_T = \int_0^T \bar{f}_\delta(t, x(t))\, dt. \tag{4.5}$$

Since $\bar{f}_\delta(t,x)$ is the pointwise maximum of two convex functions with respect to the actions, it is a convex function itself and $\bar{\mathcal{F}}_T$ is not different than the fit for the environment defined by $\bar{f}_\delta(t,x)$. By taking small values of δ we can reduce the negative portion of the fit to be as small as desired. Observe that it could be desirable to set $\delta = 0$ in order to ensure that the saturated fit is not decreased when the constraints are satisfied. However, constraint qualification conditions prevent the use of such δ since there would not exist any feasible $x \in X$ satisfying the constraint $\tilde{f}(t,x) < 0$.

A good learning strategy is one in which $x(t)$ approaches x^*. In that case,the regret and fit grow for small T but eventually stabilize or, at worst, grow at a sublinear rate. Considering regret $\mathcal{R}_T$ and fit $\mathcal{F}_T$ separately, this observation motivates the definitions of feasible trajectories strongly feasible trajectories, and strong optimal trajectories that we formally state next.

Definition 5. *Given an environment $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^m$, a cost $f_0 : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$, and a trajectory $x(t)$ we say that:*

Feasibility. *The trajectory $x(t)$ is feasible in the environment if the fit $\mathcal{F}_T$ grows sublinearly with T. I.e., if there exist a function $h(T)$ with $\limsup_{T\to\infty} h(T)/T = 0$ and a constant vector C such that for all times T it holds,*

$$\mathcal{F}_T := \int_0^T f(t, x(t))\, dt \leq Ch(T). \tag{4.6}$$

Strong Feasibility. *The trajectory $x(t)$ is strongly feasible in the environment if the fit $\mathcal{F}_T$ is bounded for all T. I.e., if there exists a constant vector C such that for all times T it*

holds,

$$\mathcal{F}_T := \int_0^T f(t, x(t))\, dt \leq C. \tag{4.7}$$

Strong optimality. *The trajectory $x(t)$ is strongly optimal in the environment if the regret $\mathcal{R}_T$ is bounded for all T. I.e., if there exists a constant C such that for all times T it holds,*

$$\mathcal{R}_T := \int_0^T f_0(t, x(t))\, dt - \int_0^T f_0(t, x^*)\, dt \leq C. \tag{4.8}$$

Having the regret satisfy $\mathcal{R}_T \leq C$ irrespectively of T is an indication that $f_0(t, x(t))$ is close to $f_0(t, x^*)$ so that the integral stops growing. This is not necessarily so because we can also achieve small regret by having $f_0(t, x(t))$ oscillate above and below $f_0(t, x^*)$ so that positive and negative values of $f_0(t, x(t)) - f_0(t, x^*)$ cancel out. In general, the possibility of having small regret by a trajectory that does not approach x^* is a limitation of the concept of regret. Alternatively, we can think of the optimal offline policy x^* as fixing a budget for cost accumulated across time. An optimal online policy meets that budget up to a constant C – perhaps by overspending at some times and underspending at some other times.

Likewise, when the fit satisfies $\mathcal{F}_T \leq C$ irrespectively of T, it suggests that $x(t)$ approaches the feasible set. This need not be true as it is possible to achieve bounded fit by having $f(t, x(t))$ oscillate around 0. Thus, as in the case of regret, we can interpret strongly feasible trajectories as meeting the *accumulated* budget $\int_0^T f(t, x(t))\, dt \leq 0$ up to a constant term C. This is in contrast with feasible actions $x^\dagger$ that meet the budget $f(t, x^\dagger) \leq 0$ for all times. Feasible trajectories differ from strongly feasible trajectories in that the fit is allowed to grow at a sublinear rate. This means that feasible trajectories do not meet the *accumulated* budget within a constant C but do meet the *time averaged* budget $(1/T)\int_0^T f(t, x(t))\, dt \leq 0$ within that constant. The notion of optimality – as opposed to strong optimality – could have been defined as a case in which regret is bounded by a sublinear function of T. This is not necessary here because our results state strong optimality.

In this chapter we solve three different problems: (i) Finding strongly optimal trajectories in unconstrained environments, (ii) finding strongly feasible trajectories and (iii) finding feasible, strongly optimal trajectories. We develop these solutions in sections 4.3, 4.4.1, and 4.4.2, respectively. Before that, we present pertinent remarks and we clarify concepts with the introduction of an example.

Remark 6 (Not every trajectory is strongly feasible). *In definition* (4.7) *we consider the integral of a measurable function in a finite interval, hence it is always bounded by a constant. Yet if the latter depends on the time horizon T, the trajectory is not strongly feasible, because it is not uniformly bounded for all time horizons T. The same remark is*

valid for the definitions of strongly optimal and feasible trajectories.

Remark 7 (**Connection with Stochastic Optimization**). *One can think about the online learning framework as a generalization of the stochastic optimization setting (see e.g. [15, 65, 107]). In the latter, the objective and constraint functions depend on a random vector $\theta \in \mathbb{R}^p$. Formally, the cost is a function $f_0 : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ and the constraints are given by a multivalued function $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^m$. The constrained stochastic optimization problem can be then formulated as*

$$\begin{aligned} x^* := &\operatorname{argmin} \mathbb{E}\left[f_0(x, \theta)\right] \\ &s.t. \qquad \mathbb{E}\left[f(x, \theta)\right] \preceq 0, \end{aligned} \tag{4.9}$$

where the above expectations are with respect to the random vector θ. When the process that determines the temporal evolution of the random vector θ_t is stationary, the expectations can be replaced by time averages. In that sense problem (4.9) *is equivalent to the problem of generating trajectories that are feasible and optimal in the sense of Definition 5.*

Remark 8 (**Sleeping Experts**). *Observe that we are considering situations in which there exits a fixed action such that it satisfies the constraints for all times $t \in [0, T]$. An alternative to this problem is to consider situations in which there is no such action, and hence the viable set $X_t^\dagger = \{x \in X : f(t, x) \preceq 0\}$ is time dependent. This is the situation considered in the sleeping-expert framework [47, 48, 53, 90]. The notions of regret considered in this framework are such that they take into account explicitly these hard constraints as opposed with our setting where we accumulate the constraint violation, thus treating them as soft constraints.*

4.2.2 The shepherd problem

Consider a target tracking problem in which an agent – the shepherd – follows a group of m targets – the sheep. Specifically, let $z(t) = [z_1(t), z_2(t)]^\top \in \mathbb{R}^2$ denote the position of the shepherd at time t. To model smooth paths for the shepherd introduce a polynomial parameterization so that each of the position components $z_k(t)$ can be written as

$$z_k(t) = \sum_{j=0}^{n-1} x_{kj} p_j(t), \tag{4.10}$$

where $p_j(t)$ are polynomials that parameterize the space of possible trajectories. The action space of the shepherd is then given by the vector that stacks the coefficients of the parameterization in (4.10), i.e., $x = [x_{10}, \ldots, x_{1,n-1}, x_{20}, \ldots, x_{2,n-1}]^\top \in \mathbb{R}^{2n}$.

Further define $y_i(t) = [y_{i1}(t), y_{i2}(t)]^\top$ as the position of the i–th sheep at time t for $i = 1, \ldots, m$ and introduce a maximum allowable distance r_i between the shepherd and

each of the sheep . The goal of the shepherd is to find a path $z(t)$ that is within distance r_i of sheep i for all sheep. This can be captured by defining an m-dimensional environment f with each component function f_i defined as

$$f_i(t,x) = \|z(t) - y_i(t)\|^2 - r_i^2 \quad \text{for all} \quad i = 1 \dots m. \tag{4.11}$$

That the environment defined by (4.11) is viable means that it is possible to select a vector of coefficients x so that the shepherd's trajectory given by (4.10) stays close to all sheep for all times. To the extent that (4.10) is a loose parameterization – we can approximate arbitrary functions with sufficiently large index n, if the time horizon is fixed and not allowed to tend to infinity –, this simply means that the sheep are sufficiently close to each other at all times. E.g., if $r_i = r$ for all times, viability is equivalent to having a maximum separation between sheep smaller than $2r$.

As an example of a problem with an optimality criterion say that the first target – the black sheep – is preferred in that the shepherd wants to stay as close as possible to it. We can accomplish that by introducing the objective function

$$f_0(t,x) = \|z(t) - y_1(t)\|^2. \tag{4.12}$$

Alternatively, we can require the shepherd to minimize the work required to follow the sheep. This behavior can be induced by minimizing the integral of the acceleration which in turn can be accomplished by defining the optimality criterion [cf. (4.2)],

$$f_0(t,x) = \left\|\ddot{z}(t)\right\| = \left\| \left[\sum_{j=0}^{n-1} x_{1j} \ddot{p}_j(t),\ \sum_{j=0}^{n-1} x_{2j} \ddot{p}_j(t) \right] \right\|. \tag{4.13}$$

Trajectories $x(t)$ differ from actions in that they are allowed to change over time, i.e., the constant values x_{kj} in (4.10) are replaced by the time varying values $x_{kj}(t)$. A feasible or strongly feasible trajectory $x(t)$ means that the shepherd is repositioning to stay close to all sheep. An optimal trajectory with respect to (4.12) is one in which he does so while staying as close as possible to the black sheep. An optimal trajectory with respect to (4.13) is one in which the work required to follow the sheep is minimized. In all three cases we apply the usual caveat that small fit and regret may be achieved with stretches of underachievement following stretches of overachievement.

4.3 Unconstrained regret in continuous time.

Before considering the feasibility problem we consider the following unconstrained minimization problem. Given an unconstrained environment $f(t,x) \equiv 0$ our goal is to generate

strong optimal trajectories $x(t)$ in the sense of Definition 5, selecting actions from a closed convex set X, i.e., $x(t) \in X$ for all $t \in [0, T]$. Given the convexity of the objective function with respect to the action, as per Assumption 7, it is natural to consider a gradient descent controller. To avoid restricting attention to functions that are differentiable with respect to x, we work with subgradients. For a convex function $g : X \to \mathbb{R}$ a subgradient g_x satisfies the inequality

$$g(y) \geq g(x) + g_x(x)^\top (y - x) \quad \text{for all} \quad y \in X. \tag{4.14}$$

In general, subgradients are defined at all points for all convex functions. At the points where the function f is differentiable the subgradient and the gradient coincide. In the case of vector functions $f : \mathbb{R}^n \to \mathbb{R}^m$ we group the subgradients of each component into a matrix $f_x(x) \in \mathbb{R}^{n \times m}$ defined as

$$f_x(x) = \left[\begin{array}{cccc} f_{1,x}(x) & f_{2,x}(x) & \cdots & f_{m,x}(x) \end{array} \right], \tag{4.15}$$

where $f_{i,x}(x)$ is a subgradient of $f_i(x)$. In addition, since the action must always be selected from the set X we define the controller in a way that the actions are the solution of a projected dynamical system over the set X. The solution has been studied in [135] and we define the notion as follow.

Definition 6. *Let X be a closed convex set.*

Projection of a point. *For any $z \in R^n$, there exits a unique element in X, denoted $P_X(z)$ such that*

$$P_X(z) = \underset{y \in X}{\operatorname{argmin}} \|y - z\|. \tag{4.16}$$

Projection of a vector at a point. *Let $x \in X$ and v a vector, the projection of v over the set X at the point x is*

$$\Pi_X(x, v) = \lim_{\delta \to 0^+} \left(P_X(x + \delta v) - x \right) / \delta. \tag{4.17}$$

Projected dynamical system. *Given a closed convex set X and a vector field $F(t, x)$ which takes elements from $\mathbb{R} \times X$ into $\mathbb{R}^n$ the projected differential equation associated with X and F is defined to be*

$$\dot{x}(t) = \Pi_X \left(x, F(t, x) \right). \tag{4.18}$$

In the above projection if the point x is in the interior of X then the projection is not different from the original vector field, i.e., $\Pi_X(x, F(t, x)) = F(t, x)$. On the other hand if the point x is in the border of X, then the projection is just the component of the vector field that is tangential to the set X at the point x. Let's consider for instance the case

where the set X is a box in R^n. Let $X = [a_1, b_1] \times \ldots \times [a_n, b_n]$ where $a_1 \ldots a_n$ and $b_1 \ldots b_n$ are real numbers. Then for each component of the vector field we have that

$$\Pi_X\left(x, F(t,x)\right)_i = \begin{cases} 0 \quad \text{if} \quad x_i = a_i \quad \text{and} \quad F(t,x)_i < 0, \\ 0 \quad \text{if} \quad x_i = b_i \quad \text{and} \quad F(t,x)_i > 0, \\ F(t,x)_i \quad \text{otherwise.} \end{cases} \tag{4.19}$$

Therefore, the proposed controller takes the form of the following projected dynamical system:

$$\dot{x} = \Pi_X\left(x, -K f_{0,x}(t,x)\right), \tag{4.20}$$

where $K > 0$ is the gain of the controller. Before stating the first theorem we need a Lemma concerning the relation between the original vector field and the projected vector field. This lemma is used in the proofs of theorems 8, 9 and 10.

Lemma 11. *Let X be a convex set and let $x_0, x \in X$. Then*

$$(x_0 - x)^\top \Pi_X(x_0, v) \leq (x_0 - x)^\top v. \tag{4.21}$$

Proof. See Appendix A.3.1. □

Let's define an Energy function $V_{\bar{x}} : \mathbb{R}^n \to \mathbb{R}$ as

$$V_{\bar{x}}(x) = \frac{1}{2}(x - \bar{x})^\top (x - \bar{x}). \tag{4.22}$$

Where $\bar{x} \in X \subset \mathbb{R}^n$ is an arbitrary fixed action. We are now in conditions to present the first theorem, which states that the solution of the gradient controller defined in (4.20) is a strongly optimal trajectory, i.e., with bounded regret for all T.

Theorem 8. *Let $f_0 : \mathbb{R} \times X \to \mathbb{R}$ be cost function satisfying assumptions 1 and 2, with $X \subseteq \mathbb{R}^n$ convex. The solution $x(t)$ of the online projected gradient controller in* (4.20) *is strongly optimal in the sense of Definition 5. In particular, the regret $\mathcal{R}_T$ can be bounded by*

$$\mathcal{R}_T \leq \frac{V_{x^*}\left(x(0)\right)}{K}, \quad \textit{for all } T \tag{4.23}$$

where $V_{\bar{x}}$ is the Energy function in (4.22).

Proof. Consider an action trajectory $x(t)$, an arbitrary given action $\bar{x} \in X$, and the corresponding energy function $V_{\bar{x}}(x(t))$ as per (4.22). The time derivative $\dot{V}_{\bar{x}}(x(t))$ is given by

$$\dot{V}_{\bar{x}}(x(t)) = (x(t) - \bar{x})^\top \dot{x}(t). \tag{4.24}$$

If the trajectory $x(t)$ follows from the online projected gradient dynamical system in (4.20) we can substitute the trajectory derivative $\dot{x}$ by the vector field value and reduce (4.24) to

$$\dot{V}_{\bar{x}}(x(t)) = (x(t) - \bar{x})^{\top} \Pi_X \left(x(t), -K f_{0,x}(t, x(t)) \right). \tag{4.25}$$

Use now the result in Lemma 11 with $v = -K f_{0,x}(t, x(t))$ to remove the projection operator from (4.25) and write

$$\dot{V}_{\bar{x}}(x(t)) \leq -K(x(t) - \bar{x})^{\top} f_{0,x}(t, x(t)). \tag{4.26}$$

Using the defining equation of a subgradient (4.14), we can upper bound the inner product $-(x(t) - \bar{x})^{\top} f_{0,x}(t, x(t))$ by the difference $f_0(t, \bar{x}) - f_0(t, x(t))$ and transform (4.26) into

$$\dot{V}_{\bar{x}}(x(t)) \leq K \left(f_0(t, \bar{x}) - f_0(t, x(t)) \right). \tag{4.27}$$

Rearranging and integrating the above inequality yields

$$\int_0^T f_0(t, x(t))\, dt - \int_0^T f_0(t, \bar{x})\, dt \leq -\frac{1}{K} \int_0^T \dot{V}_{\bar{x}}(x(t))\, dt. \tag{4.28}$$

Since the primitive of $\dot{V}_{\bar{x}}(x(t))$ is $V_{\bar{x}}(x(t))$ we can evaluate the integral on the right hand side of (4.28) and further use the fact that $V_{\bar{x}}(x) \geq 0$ for all $x \in \mathbb{R}^n$ to conclude that

$$-\int_0^T \dot{V}_{\bar{x}}(x(t)) dt \;=\; V_{\bar{x}}(x(0)) - V_{\bar{x}}(x(T)) \;\leq\; V_{\bar{x}}\left(x(0)\right). \tag{4.29}$$

Combining the bounds in (4.28) and (4.29) we have that

$$\int_0^T f_0(t, x(t))\, dt - \int_0^T f_0(t, \bar{x})\, dt \leq V_{\bar{x}}(x(0))/K. \tag{4.30}$$

Since the above inequality holds for an arbitrary point $\bar{x} \in \mathbb{R}^n$ it holds for $\bar{x} = x^*$ in particular. When making $\bar{x} = x^*$ in (4.30) the left hand side reduces to the regret $\mathcal{R}_T$ associated with the trajectory $x(t)$ [cf. (4.3)] and in the right hand side we have $V_{\bar{x}}(x(0))/K = V_{x^*}(x(0))/K$. Eq. (4.23) follows because (4.30) is true for all times T. This implies that the trajectory is strongly optimal according to (4.8) in Definition 5. □

The strong optimality of the online projected gradient controller in (4.20) that we claim in Theorem 8 is not a straightforward generalization of the optimality of gradient controllers in constant convex potentials. The functions f_0 are allowed to change arbitrarily over time and are not observed until after the cost $f_0(t, x(t))$ has been incurred.

Since the initial value of the Energy function $V_{x^*}(x(0))$ is the square of the distance

between $x(0)$ and x^*, the bound on the regret in (4.23) shows that the closer we start to the optimal point the smaller the accumulated cost is. Likewise, the larger the controller gain K, the smaller the bound on the regret is. Theoretically, we can make this bound arbitrarily small. This is not possible in practice because larger K entails trajectories with larger derivatives which cannot be implemented in systems with physical constraints. In the example in Section 4.2.2 the derivatives of the state $x(t)$ control the speed and acceleration of the shepherd. The physical limits of these quantities along with an upper bound on the cost gradient $f_{0,x}(t,x)$ can be used to estimate the largest allowable gain K.

Another observation regarding the bound on the regret is that it does not depend on the function that we are minimizing –except for the location of the point x^*. For instance by scaling a function the bound on the regret is kept constant if the same gain K can be selected. This is not surprising since a scaling in the function implies a bigger cost but it also entails a larger action derivative, which allows to track better changes on the function. However, if a bound on the maximum allowed gain exists then the regret bound cannot be invariant to scalings.

Remark 9. In discrete time systems where t is a natural variable and the integrals in (4.3) are replaced by sums, online gradient descent algorithms can be used to reduce regret; see e.g. [42, 138]. The online gradient controller in (4.20) is a direct generalization of online gradient descent to continuous time. This similarity notwithstanding, the result in Theorem 8 is stronger than the corresponding bound on the regret in discrete time which states a sublinear growth at a rate not faster than $\sqrt{T}$ if the cost function is convex [138], and $\log T$ if the cost function is strictly convex [42]. The key where this difference lies is in the fact that discrete time algorithms have to "pay" to switch from the action at time t to the action at time $t+1$. In the proofs of [42, 138] a term related to the norm square of the gradient is present in the upper bound on the regret while in continuous time this term is absent. The bound on the norm of the gradient is related to the selecting a different action. As in the case of fictitious plays that lead to no regret in the continuous time but not in discrete time (see e.g. [41, 130, 133]) the bounds on the regret in continuous time are tighter than in its discrete counterpart for online gradient descent.

4.4 Saddle point algorithm

Given an environment $f(t,x)$ and an objective function $f_0(t,x)$ verifying assumptions 6 and 7 we set our attention towards two different problems: design a controller whose solution is a strongly feasible trajectory and a controller whose solution is a feasible and strongly optimal trajectory. As already noted, when the environment is known beforehand the problem of finding such trajectories is a constrained convex optimization problem, which

we can solve using the saddle point algorithm of Arrow and Hurwicz [4]. Following this idea, let $\lambda \in \Lambda = \mathbb{R}^m_+$, be a multiplier and define the time-varying Lagrangian associated with the online problem as

$$\mathcal{L}(t, x, \lambda) = f_0(t, x) + \lambda^\top f(t, x). \tag{4.31}$$

Saddle point methods rely on the fact that for a constrained convex optimization problem, a pair is a primal-dual optimal solution if and only if it is a saddle point of the Lagrangian associated with the problem; see e.g. [16]. The idea of the algorithm is then to generate trajectories that descend in the opposite direction of the gradient of the Lagrangian with respect to x and that ascend in the direction of the gradient with respect to λ. Since the Lagrangian is differentiable with respect to λ, we denote by $\mathcal{L}_\lambda(t, x, \lambda) = f(t, x)$ the derivative of the Lagrangian with respect to λ. On the other hand, since the functions $f_0(\cdot, x)$ and $f(\cdot, x)$ are convex, the Lagrangian is also convex with respect to x. Thus, its subgradient with respect to x always exist, let us denote it by $\mathcal{L}_x(t, x, \lambda)$. Let K be the gain of the controller, then following the ideas in [4] we define a controller that descends in the direction of the subgradient with respect to the action x

$$\begin{aligned} \dot{x} &= \Pi_X\left(x, -K\mathcal{L}_x(t, x, \lambda)\right) \\ &= \Pi_X\left(x, -K(f_{0,x}(t, x) + f_x(t, x)\lambda)\right), \end{aligned} \tag{4.32}$$

and that ascends in the direction of the subgradient with respect to the multiplier λ

$$\dot{\lambda} = \Pi_\Lambda\left(\lambda, K\mathcal{L}_\lambda(t, x, \lambda)\right) = \Pi_\Lambda\left(\lambda, Kf(t, x)\right). \tag{4.33}$$

The projection over the set X in (4.32) is done to assure that the trajectory is always in the set of possible actions. The operator $\Pi_\Lambda(\lambda, f)$ is a projected dynamical system in the sense of Definition 6 over the set Λ. This projection is done to assure that $\lambda(t) \in \mathbb{R}^m_+$ for all times $t \in [0, T]$. An important observation regarding (4.32) and (4.33) is that the environment is observed locally in space and causally in time. The values of the environment constraints and its subgradients are observed at the current trajectory position $x(t)$ and the values of $f(t, x(t))$ and $f_x(t, x(t))$ affect the derivatives of $x(t)$ and $\lambda(t)$ only. Notice that if the environment function satisfies $f(t, x) \equiv 0$ we recover the algorithm defined in (4.20) as a particular case of the saddle point controller.

A block diagram for the controller in (4.32) - (4.33) is shown in Figure 4.1. The controller operates in an environment to which it inputs at time t an action $x(t)$ that results in a penalty $f(t, x(t))$ and cost $f_0(t, x(t))$. The value of these functions and their subgradients $f_x(t, x(t))$ and $f_{0,x}(t, x(t))$ are observed and fed to the multiplier and action feedback loops.

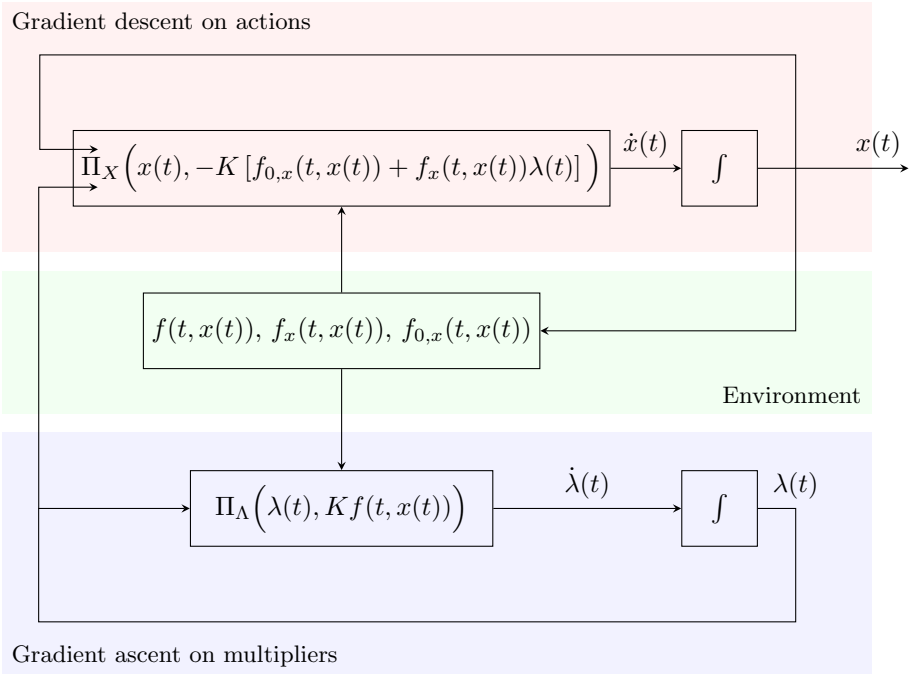


Figure 4.1: Block diagram of the saddle point controller. Once that action $x(t)$ is selected at time t, we measure the corresponding values of $f(t, x(t))$, $f_x(t, x(t))$ and $f_{0,x}(t, x(t))$. This information is fed to the two feedback loops. The action loop defines the descent direction by computing weighted averages of the subgradients $f_x(t, x(t))$ and $f_{0,x}(t, x(t))$. The multiplier loop uses $f(t, x(t))$ to update the corresponding weights.

The action feedback loop behaves like a weighted gradient descent controller. We move in the direction given by a linear combination of the the gradient of the objective function $f_{0,x}(t, x(t))$ and the constraint subgradients $f_{i,x}(t, x(t))$ weighted by their corresponding multipliers $\lambda_i(t)$. Intuitively, this pushes $x(t)$ towards satisfying the constraints and to the minimum of the objective function in the set where constraints are satisfied. However, the question remains of how much weight to give to each constraint. This is the task of the multiplier feedback loop. When constraint i is violated we have $f_i(t, x(t)) > 0$. This pushes the multiplier $\lambda_i(t)$ up, thereby increasing the force $\lambda_i(t) f_{i,x}(t, x(t))$ pushing $x(t)$ towards satisfying the constraint. If the constraint is satisfied, we have $f_i(t, x(t)) < 0$, the multiplier $\lambda_i(t)$ being decreased, and the corresponding force decreasing. The more that constraint i is violated, the faster we increase the multiplier, and the more we increase the force that pushes $x(t)$ towards satisfying $f_i(t, x(t)) < 0$. If the constraint is satisfied, the force is decreased and may eventually vanish altogether if we reach the point of making $\lambda_i(t) = 0$.

4.4.1 Strongly feasible trajectories

We begin by studying the saddle point controller defined by (4.32) and (4.33) in a problem in which optimality is *not* taken into account, i.e., $f_0(t,x) \equiv 0$. In this case the action descent equation of the controller (4.32) takes the form

$$\dot{x} = \Pi_X \left(x, -K\mathcal{L}_x(t,x,\lambda)\right) = \Pi_X \left(x, -Kf_x(t,x)\lambda\right), \tag{4.34}$$

while the multiplier ascent equation (4.33) remains unchanged. The bounds to be derived for the fit ensure that the trajectories $x(t)$ are strongly feasible in the sense of Definition 5. To state the result consider an arbitrary fixed action $\bar{x} \in X$ and an arbitrary multiplier $\bar{\lambda} \in \Lambda$ and define the energy function

$$V_{\bar{x},\bar{\lambda}}(x,\lambda) = \frac{1}{2}\left(||x-\bar{x}||^2 + ||\lambda - \bar{\lambda}||^2\right). \tag{4.35}$$

We can then bound fit in terms of the initial value $V_{\bar{x},\bar{\lambda}}(x(0),\lambda(0))$ of the energy function for properly chosen $\bar{x}$ and $\bar{\lambda}$ as we formally state next.

Theorem 9. *Let $f : \mathbb{R} \times X \to \mathbb{R}^m$, satisfying assumptions 6 and 7, where $X \subseteq \mathbb{R}^n$ is a convex set. If the environment is viable, then the solution $x(t)$ of the dynamical system defined by* (4.34) *and* (4.33) *is strongly feasible for all $T > 0$. Specifically, the fit is bounded by*

$$\mathcal{F}_{T,i} \leq \min_{x^\dagger \in X^\dagger} \frac{1}{K} V_{x^\dagger,e_i}(x(0),\lambda(0)), \tag{4.36}$$

where e_i with $i = 1 \ldots m$ form the canonical base of $\mathbb{R}^m$.

Proof. Consider action trajectories $x(t)$ and multiplier trajectories $\lambda(t)$ and the corresponding energy function $V_{\bar{x},\bar{\lambda}}(x(t),\lambda(t))$ in (4.35) for arbitrary given action $\bar{x} \in X$ and multiplier $\bar{\lambda} \in \Lambda$. The derivative $\dot{V}_{\bar{x},\bar{\lambda}}(x(t),\lambda(t))$ of the energy with respect to time is then given by

$$\dot{V}_{\bar{x},\bar{\lambda}}(x(t),\lambda(t)) = (x(t)-\bar{x})^\top \dot{x}(t) + (\lambda(t)-\bar{\lambda})^\top \dot{\lambda}(t). \tag{4.37}$$

Substitute the action and multiplier derivatives by their corresponding values given in (4.34) and (4.33) to reduce (4.37) to

$$\begin{aligned}\dot{V}_{\bar{x},\bar{\lambda}}(x(t),\lambda(t)) =& (x(t)-\bar{x})^\top \Pi_X \left(x, -Kf_x(t,x(t))\lambda(t)\right) \\ &+ (\lambda(t)-\bar{\lambda})^\top \Pi_\Lambda \left(\lambda, Kf(t,x(t))\right).\end{aligned} \tag{4.38}$$

Then, using the result of Lemma 11 for both X and Λ, the following inequality holds

$$\begin{aligned}\dot{V}_{\bar{x},\bar{\lambda}}(x(t),\lambda(t)) \leq K(\bar{x}-x(t))^\top f_x(t,x(t))\lambda(t) \\ + K(\lambda(t)-\bar{\lambda})^\top f(t,x(t)).\end{aligned} \tag{4.39}$$

Notice that $f(t,x)\lambda(t)$ is a convex function with respect to the action, therefore we can upper bound the inner product $(\bar{x}-x(t))^\top f_x(t,x(t))\lambda(t)$ by the quantity $f(t,\bar{x})^\top\lambda(t) - f(t,x(t))^\top\lambda(t)$ and transform (4.39) into

$$\begin{aligned}\dot{V}_{\bar{x},\bar{\lambda}}(x(t),\lambda(t)) \leq K\left(f(t,\bar{x})-f(t,x(t))\right)^\top\lambda(t) \\ + K(\lambda(t)-\bar{\lambda})^\top f(t,x(t)).\end{aligned} \tag{4.40}$$

Further note that in the above equation the second and the third term are opposite. Thus, it reduces to

$$\dot{V}_{\bar{x},\bar{\lambda}}(x(t),\lambda(t)) \leq K\left[\lambda(t)^\top f(t,\bar{x}) - \bar{\lambda}^\top f(t,x(t))\right]. \tag{4.41}$$

Observe that the integral of the left hand side of the above equation can be written as

$$\int_0^T \dot{V}_{\bar{x},\bar{\lambda}}(x(t),\lambda(t))dt = V_{\bar{x},\bar{\lambda}}(x(T),\lambda(T)) - V_{\bar{x},\bar{\lambda}}(x(0),\lambda(0)). \tag{4.42}$$

Then using the fact that $V_{\bar{x},\bar{\lambda}}(x(t)),\lambda(t)) \geq 0$ for all t, yields

$$\int_0^T \dot{V}_{\bar{x},\bar{\lambda}}(x(t),\lambda(t))dt \geq -V_{\bar{x},\bar{\lambda}}\left(x(0),\lambda(0)\right). \tag{4.43}$$

Then, integrating both sides of (4.42) and using the bound in (4.43), we have that

$$\int_0^T \bar{\lambda}^\top f(t,x(t)) - \lambda(t)^\top f(t,\bar{x})dt \leq \frac{V_{x^\dagger,\bar{\lambda}}(x(0),\lambda(0))}{K}. \tag{4.44}$$

Since the environment is viable, there exist a fixed action $x^\dagger$ such that $f(t,x^\dagger) \preceq 0$ for all $t \geq 0$. Then choosing $\bar{x} = x^\dagger$, since $\lambda(t) \succeq 0$ for all t, we have that

$$\lambda(t)^\top f(t,x^\dagger) \leq 0 \; \forall t \in [0,T]. \tag{4.45}$$

Therefore the left hand side of (4.44) can be lower bounded by

$$\bar{\lambda}^\top \int_0^T f(t,x(t))dt \leq \frac{V_{x^\dagger,\bar{\lambda}}(x(0),\lambda(0))}{K}. \tag{4.46}$$

Choosing $\bar{\lambda} = e_i$ where e_i is the i–th element of the canonical base of $\mathbb{R}^m$, we have that for

all $i = 1 \dots m$:

$$\int_0^T f_i(t, x(t))dt \leq \frac{V_{x^\dagger, e_i}(x(0), \lambda(0))}{K}. \tag{4.47}$$

Notice that since the above inequality holds for any $x^\dagger \in X^\dagger$ it is also true for the particular $x^\dagger$ that minimizes the right hand side. The left hand side of the above inequality is the i–th component of the fit. Thus, since the m components of the fit of the trajectory generated by the saddle point algorithm are bounded for all T, the trajectory is strongly feasible with the specific upper bound stated in (4.36). □

Theorem 9 assures that if an environment is viable for an agent that selects actions over a set X, the solution of the dynamical system given by (4.34) and (4.33) is a trajectory $x(t)$ that is strongly feasible in the sense of Definition 5. This result is not trivial, since the function f that defines the environment is observed causally and can change arbitrarily over time. In particular, the agent could be faced with an adversarial environment that changes the function f in a way that makes the value of $f(t, x(t))$ larger. The caveat is that the choice of the function f must respect the viability condition that there exists a feasible action $x^\dagger$ such that $f(t, x^\dagger) \preceq 0$ for all $t \in [0, T]$. This restriction still leaves significant leeway for strategic behavior. E.g., in the shepherd problem of Section 4.2.2 we can allow for strategic sheep that observe the shepherd's movement and respond by separating as much as possible. The strategic action of the sheep are restricted by the condition that the environment remains viable, which in this case reduces to the not so stringent condition that the sheep stay in a ball of radius $2r$ if all $r_i = r$.

Since the initial value of the energy function $V_{x^\dagger, e_i}(x(0), \lambda(0))$ is the square of the distance between $x(0)$ and $x^\dagger$ added to a term that depends on the distance between the initial multiplier and e_i, the bound on the fit in (4.36) shows that the closer we start to the feasible set the smaller the accumulated constraint violation becomes. Likewise, the larger the gain K, the smaller the bound on the fit is. As in section 4.3 we observe that increasing K can make the bound on the fit arbitrarily small, yet for the same reasons discussed in that section this can't be done.

Further notice that for the saddle point controller defined by (4.34) and (4.33) the action derivatives are proportional not only to the gain K but to the value of the multiplier λ. Thus, to select gains that are compatible with the system's physical constraints we need to determine upper bounds in the multiplier values $\lambda(t)$. An upper bound follows as a consequence of Theorem 9 as we state in the following corollary.

Corollary 4. *Given the controller defined by* (4.34) *and* (4.33) *and assuming the same hypothesis of Theorem 9, if the set of actions X is bounded in norm by R, then the multipliers*

λ are bounded for all times by

$$0 \leq \lambda_i(t) \leq \left(4R^2 + 1\right), \text{ for all } i = 1, \ldots, m. \tag{4.48}$$

Proof. First of all notice that according to (4.33) a projection over the positive orthant is performed for the multiplier update. Therefore, for each component of the multiplier we have that $\lambda_i(t) \geq 0$ for all $t \in [0, T]$. On the other hand, since the trajectory of the multipliers is defined by $\dot{\lambda}(t) = \Pi_\Lambda(\lambda(t), Kf(t, x(t))$, while $\lambda(t) > 0$ we have that $\dot{\lambda}(t) = Kf(t, x(t))$. Let t_0 be the first time instant for which $\lambda_i(t) > 0$ for a given $i \in \{1, 2, \ldots, m\}$, i.e.,

$$t_0 = \inf \{t \in [0, T], \lambda_i(t) > 0\}. \tag{4.49}$$

In addition, let T_0^* be the first time instant greater than t_0 where $\lambda_i(t) = 0$, if this time is larger than T we set $T_0^* = T$

$$T_0^* = \max \{\inf \{t \in (t_0, T], \lambda_i(t) > 0\}, T\}. \tag{4.50}$$

Further define $t_{s+1} = \inf \{t \in [T_s^*, T], \lambda_i(t) > 0\}$, and

$$T_s^* = \max \{\inf \{t \in (t_s, T], \lambda_i(t) > 0\}, T\}. \tag{4.51}$$

From the above definition it holds that in any time in the interval $(T_s^*, t_{s+1}]$, $\lambda_i(t) = 0$. And therefore in those intervals the multipliers are bounded. In the case where $\tau \in (t_s, T_s^*]$

$$\int_{t_s}^{\tau} \dot{\lambda}_i(t)dt = \int_{t_s}^{\tau} Kf_i(t, x(t))dt. \tag{4.52}$$

Notice that the right hand side of the above equation is, proportional to the i–th component of the fit restricted to the time interval $[t_0, \tau]$. In Theorem 9 it was proved that the i–th component of the fit is bounded for all time horizons by $V_{x^\dagger, e_i}(x(t_s), 0)/K$. In this particular case we have that

$$V_{x^\dagger, e_i}(x(t_s), 0) = \frac{1}{2}\left((x(t_s) - x^\dagger)^2 + (0 - e_i)^2\right), \tag{4.53}$$

and since for any $x \in X$ we have that $\|x\| \leq R$, we conclude

$$V_{x^\dagger, e_i}(x(t_s), 0) \leq \frac{1}{2}\left((2R)^2 + 1^2\right). \tag{4.54}$$

Therefore, for all $\tau \in (t_s T_s^*]$ $\lambda_i(\tau) \leq \frac{1}{2}\left(4R^2 + 1^2\right)$. This completes the proof that the multipliers are bounded. □

The bound in Corollary 4 ensures that action derivatives $\dot{x}(t)$ remain bounded if the subgradients are. This means that action derivatives increase, at most, linearly with K and is not compounded by an arbitrary increase of the multipliers. Observe as well, tat the cumulative nature of the fit does not guarantee that the constraint violation is controlled. This is because time intervals of constraint violations can be compensated by time intervals where the constraints are negative. To overcome this issue, we next show that the saddle point controller archives bounded saturated fit for all time horizon.

Corollary 5. *Let the hypothesis of Theorem 9 hold. Let $\delta > 0$ and let $\bar{\mathcal{F}}_T$ be the saturated fit defined in* (4.5)*. Then, the solution of the dynamical system* (4.34) *and* (4.33) *when $f(t,x)$ is replaced by $\bar{f}_\delta(t,x)) = \max\{f(t,x), -\delta\}$ archives a bounded saturated fit. Furthermore the bound is given by*

$$\bar{\mathcal{F}}_{T,i} \leq \min_{x^\dagger \in X^\dagger} \frac{1}{K} V_{x^\dagger, e_i}(x(0), \lambda(0)), \tag{4.55}$$

where e_i with $i = 1 \dots m$ form the canonical base of $\mathbb{R}^m$.

Proof. Since $\bar{f}_\delta(t,x)$ is the pointwise maximum of two convex functions, it is a convex function itself. As a consequence of Theorem 9 the fit for the environment $\bar{f}_\delta(t,x)$ satisfies

$$\int_0^T \bar{f}_{\delta,i}(t, x(t))\, dt \leq \min_{x^\dagger \in X^\dagger} \frac{1}{K} V_{x^\dagger, e_i}(x(0), \lambda(0)). \tag{4.56}$$

The fact that the left hand side of the above equation corresponds to the saturated fit [cf., (4.5)] completes the proof. □

The above result establishes that a trajectory that follows the saddle point dynamics for the environment defined by $\bar{f}_\delta(t,x)$ achieves bounded saturated fit. This means that it is possible to adapt the controller (4.34) and (4.33), so that the fit is bounded while not alternating between periods of large under and over satisfaction of the constraints.

4.4.2 Strongly optimal feasible trajectories

This section presents bounds on the growth of the fit and the regret of the trajectories $x(t)$ that are solutions of the saddle point controller defined by (4.32) and (4.33). These bounds ensure that the trajectory is feasible and strongly optimal in the sense of Definition 5. To derive these bounds we need the following assumption regarding the objective function.

AS8. *There is a finite constant γ independent of the time horizon T such that for all t in the interval $[0, T]$.*

$$\gamma \geq f_0(t, x^*) - \min_{x \in X} f_0(t, x), \tag{4.57}$$

where x^* is the solution of the offline problem (4.2)*.*

The existence of the bound in (4.57) is a mild requirement. Since the function $f_0(t,x)$ is convex, for any time t it is lower bounded if the action space is bounded, as is the case in most applications of practical interest. The only restriction imposed is that $\min_{x\in X} f_0(t,x)$ does not become progressively smaller with time so that a uniform bound γ holds for all times t. The bound can still hold if X is not compact as long as the span of the functions $f_0(t,x)$ is not unbounded below. A consequence of Assumption 8 is that the regret cannot *decrease* faster than a linear rate as we formally state in the following lemma.

Lemma 12. *Let $X \subset \mathbb{R}^n$ be a convex set. If Assumption 8 holds, then the regret defined in* (4.3) *is lower bounded by $-\gamma T$ where γ is the constant defined in* (4.57)*, i.e.,*

$$\mathcal{R}_T \geq -\gamma T. \tag{4.58}$$

Proof. Let $x(t)$ be the action at time t when the agent follows the dynamics defined by (4.32) and (4.33), because of Assumption 8, we have that

$$f_0(t,x(t)) - f_0(t,x^*) \geq -\gamma, \tag{4.59}$$

Integrating both sides of the above equation yields

$$\int_0^T f_0(t,x(t))dt - \int_{t=0}^T f_0(t,x^*)dt \geq -\gamma T. \tag{4.60}$$

Since the left hand side of the above equation is the regret up to time T defined in (4.3), the proof is completed. □

Observe that regret is a quantity that we want to make small and, therefore, having negative regret is a desirable outcome. The result in Lemma 12 puts a floor on how much we can succeed in making regret negative. Using the bound in (4.58) and the definition of the energy function in (4.35) we can formalize bounds on the regret and the fit, for an action trajectory $x(t)$ that follows the saddle point dynamics in (4.32) and (4.33).

Theorem 10. *Let $X \subset \mathbb{R}^n$ be a compact convex set and let $f : \mathbb{R} \times X \to \mathbb{R}^m$ and $f_0 : \mathbb{R} \times X \to \mathbb{R}$, be functions satisfying assumptions 6, 7 and 8. If the environment is viable, then the solution of the system defined by* (4.32) *and* (4.33) *is a trajectory $x(t)$ that is feasible and strongly optimal for all time horizons $T > 0$ if the gain $K > 1$. In particular, the fit is bounded by*

$$\mathcal{F}_{T,i} \leq \mathcal{O}\left(\sqrt{\gamma T}, K^0\right), \tag{4.61}$$

and the regret is bounded by

$$\mathcal{R}_T \leq \frac{1}{K} V_{x*,0}\left(x(0), \lambda(0)\right), \tag{4.62}$$

where $V_{\bar{x},\bar{\lambda}}(x,\lambda)$ is the energy function defined in (4.35)*, x^* is the solution to the problem* (4.2) *and γ is the constant defined in* (4.57)*. The notation $\mathcal{O}\left(K^0\right)$ refers to a function that is constant with respect to the gain K.*

Proof. Consider action trajectories $x(t)$ and multiplier trajectories $\lambda(t)$ and the corresponding energy function $V_{\bar{x},\bar{\lambda}}(x,\lambda)$ in (4.35), for arbitrary given action $\bar{x} \in \mathbb{R}^n$ and multiplier $\bar{\lambda} \in \Lambda$. The derivative $\dot{V}_{\bar{x},\bar{\lambda}}(x(t),\lambda(t))$ is given by

$$\dot{V}_{\bar{x},\bar{\lambda}}(x(t),\lambda(t)) = (x(t)-\bar{x})^\top \dot{x}(t) + (\lambda(t)-\bar{\lambda})^\top \dot{\lambda}(t). \tag{4.63}$$

If the trajectories $x(t)$ and $\lambda(t)$ follow from the saddle point dynamical system defined by (4.32) and (4.33) respectively we can substitute the action and multiplier derivatives by their corresponding values and reduce (4.63) to

$$\begin{aligned}\dot{V}_{\bar{x},\bar{\lambda}}(x(t),\lambda(t)) = (x(t)-\bar{x})^\top \Pi_X(x, -K\mathcal{L}_x(t,x(t),\lambda(t))) \\ + (\lambda(t)-\bar{\lambda})^\top \Pi_\Lambda(\lambda, K\mathcal{L}_\lambda(t,x(t),\lambda(t))).\end{aligned} \tag{4.64}$$

Then, use Lemma 11 for both X and Λ to write

$$\begin{aligned}\dot{V}_{\bar{x},\bar{\lambda}}(x(t),\lambda(t)) \leq -K(x(t)-\bar{x})^\top \mathcal{L}_x(t,x(t),\lambda(t)) \\ + K(\lambda(t)-\bar{\lambda})^\top \mathcal{L}_\lambda(t,x(t),\lambda(t)).\end{aligned} \tag{4.65}$$

Since $\mathcal{L}(t,x(t),\lambda(t))$ is a convex function, (4.14) takes the form

$$-(x(t)-\bar{x})^\top \mathcal{L}_x(t,x(t),\lambda(t)) \leq \mathcal{L}(t,\bar{x},\lambda(t)) - \mathcal{L}(t,x(t),\lambda(t)). \tag{4.66}$$

From the linearity of the Lagrangian with respect to λ we have

$$(\lambda(t)-\bar{\lambda})^\top \mathcal{L}_\lambda(t,x(t),\lambda(t)) = \mathcal{L}(t,x(t),\lambda(t)) - \mathcal{L}(t,x(t),\bar{\lambda}). \tag{4.67}$$

Combine expressions (4.66) and (4.67) to reduce (4.65) to

$$\dot{V}_{\bar{x},\bar{\lambda}}(x(t),\lambda(t)) \leq K\left(\mathcal{L}(t,\bar{x},\lambda(t)) - \mathcal{L}(t,x(t),\bar{\lambda})\right). \tag{4.68}$$

Substituting the Lagrangians by the expression (4.31)

$$\begin{aligned}\dot{V}_{\bar{x},\bar{\lambda}}(x(t),\lambda(t)) \leq K[f_0(t,\bar{x}) + \lambda^\top(t) f(t,\bar{x}) \\ - f_0(t,x(t)) - \bar{\lambda}^\top f(t,x(t))].\end{aligned} \tag{4.69}$$

Rewriting the above inequality and integrating both sides with respect to the time from

time $t = 0$ to $t = T$, we obtain

$$\int_0^T f_0(t, x(t)) - f_0(t, \bar{x}) + \bar{\lambda}^\top f(t, x(t)) - \lambda(t)^\top f(t, \bar{x}) dt \\ \leq -\frac{1}{K} \int_0^T \dot{V}_{\bar{x},\bar{\lambda}}(x(t), \lambda(t)) dt. \tag{4.70}$$

Using the result (4.43) the above equation reduces to

$$\int_0^T f_0(t, x(t)) - f_0(t, \bar{x}) + \bar{\lambda}^\top f(t, x(t)) - \lambda(t)^\top f(t, \bar{x}) dt \\ \leq \frac{1}{K} V_{\bar{x},\bar{\lambda}}(x(0), \lambda(0)). \tag{4.71}$$

Since (4.71) holds for any $\bar{x} \in X$ and any $\bar{\lambda} \in \Lambda$, it holds for $\bar{x} = x^*$, $\bar{\lambda} = 0$. Since $\lambda(t)^\top f(t, x^*)\, dt \leq 0 \;\; \forall t \in [0, T]$ we can lower bound the left hand side of (4.71) to obtain:

$$\int_0^T f_0(t, x(t)) - f_0(t, x^*) dt \leq \frac{1}{K} V_{x^*,0}(x(0), \lambda(0)). \tag{4.72}$$

Notice that the left hand side of the above equation is the definition of regret given in (4.3). Thus, we have showed that (4.62) holds and since the right hand side of the above equation is a constant for all T we proved that the trajectory generated by the saddle point controller is strongly optimal. It remains to prove that the trajectory generated is feasible. Choosing $\bar{x} = x^*$ in (4.71) and using the result of Lemma 12 yields

$$\int_0^T \bar{\lambda}^\top f(t, x(t)) - \lambda(t)^\top f(t, x^*)\, dt \\ \leq \frac{1}{K} V_{x^*,\bar{\lambda}}(x(0), \lambda(0)) + \gamma T. \tag{4.73}$$

Since $\lambda(t)^\top f(t, x^*)\, dt \leq 0 \;\; \forall t \in [0, T]$ the left hand side of the above equation is lower bounded by $\bar{\lambda}^\top \int_0^T f(t, x(t))$, yielding

$$\bar{\lambda}^\top \int_0^T f(t, x(t)) dt \leq \frac{V_{x^*,\bar{\lambda}}(x(0), \lambda(0))}{K} + \gamma T. \tag{4.74}$$

Now let's choose $\bar{\lambda} = [\mathcal{F}_T]^+ = \left[\int_0^T f(t, x(t))\, dt\right]^+$ and define the following set of indices

$$I = \{i = 1 \dots m | \int_0^T f_i(t, x(t))\, dt \geq 0)\}. \tag{4.75}$$

Notice that if $i \notin I$, then $\bar{\lambda}_i \int_0^T f_i(t, x(t))\, dt = 0$. On the other hand, if $i \in I$, $\bar{\lambda}_i \int_0^T f_i(t, x(t))\, dt =$

$\left(\int_0^T f_i(t,x(t))\,dt\right)^2 \geq 0$. Thus,

$$\bar{\lambda}^\top \int_0^T f(t,x(t))dt = \left\|[\mathcal{F}_T]^+\right\|^2. \tag{4.76}$$

Write then inequality (4.74) for the particular choice of $\bar{\lambda}$ as

$$\left\|[\mathcal{F}_T]^+\right\|^2 \leq \frac{1}{K} V_{x^*,[\mathcal{F}_T]^+}(x(0),\lambda(0)) + \gamma T. \tag{4.77}$$

Use the definition of the energy function $V_{\bar{x},\bar{\lambda}}(x,\lambda)$ given in (4.35) to write the above inequality as

$$\left\|[\mathcal{F}_T]^+\right\|^2 \leq \frac{1}{K}\left(\|x(0)-x^*\|^2 + \left\|[\mathcal{F}_T]^+ - \lambda(0)\right\|^2\right) + \gamma T. \tag{4.78}$$

Expand the second square in the right hand side of the above expression and re arrange terms to write

$$\left\|[\mathcal{F}_T]^+\right\|^2 + \lambda(0)^\top [\mathcal{F}_T]^+ \frac{2}{K-1} \leq \frac{1}{K-1}\left(\|x(0)-x^*\|^2 + \|\lambda(0)\|^2\right) + \gamma T \frac{K}{K-1}. \tag{4.79}$$

Adding in both sides of the above inequality $\|\lambda(0)\|^2 \left(\frac{1}{K-1}\right)^2$, then factorizing the left hand side the above inequality yields

$$\left\|[\mathcal{F}_T]^+ + \lambda(0)\frac{1}{K-1}\right\|^2 \leq \frac{1}{K-1}\|x(0)-x^*\|^2 + \gamma T \frac{K}{K-1} + \frac{\|\lambda(0)\|^2}{K-1}\left(1+\frac{1}{K-1}\right). \tag{4.80}$$

Since the term $\lambda(0)/(K-1)$ is constant with respect to T it is the case that the norm of $[\mathcal{F}_T]^+$ is bounded by a function that grows like $\sqrt{T}$. On the other hand it also holds that $\|[\mathcal{F}_T]^+\|$ is bounded by a constant function of the gain K. These observations lead to the conclusion that

$$\|[\mathcal{F}_T]^+\| \leq \mathcal{O}\left(\sqrt{\gamma T}, K^0\right). \tag{4.81}$$

The above inequality implies that for any $i \in I$ it is the case that $\mathcal{F}_{T,i} \leq \mathcal{O}\left(\sqrt{\gamma T}, K^0\right)$. If $i \notin I$ it means that $\mathcal{F}_{T,i} < 0$ and it trivially satisfies (4.61). Which proves that the trajectories that are solution of the saddle point controller defined by (4.32) and (4.33) are feasible since they are bounded by a sublinear function of the time horizon for all T. □

Theorem 10 assures that if the environment is viable for an agent selecting actions from a bounded set X, the solution of the saddle point dynamics defined in (4.32)-(4.33) is a trajectory that is feasible and strongly optimal. The bounds on the fit in theorems 9 and 10 prove a trade off between optimality and feasibility. If optimality of the trajectory is

not of interest it is possible to get strongly feasible trajectories with fit that is bounded by a constant independent of the time horizon T (cf. Theorem 9). When an optimality criterion is added to the problem, its satisfaction may come at the cost of a fit that may increase as $\sqrt{T}$. An important consequence of this difference is that even if we could set the gain K to be arbitrarily large, the bound on the fit cannot be made arbitrarily small. This bound would still grow as $\sqrt{\gamma T}$. The result in Theorem 10 also necessitates Assumption 8 as opposed to Theorem 9.

As in the cases of theorems 8 and 9 it is possible to have the environment and objective function selected strategically. Further note that, again, the initial value of the energy function used to bound regret is related with the square of the distance between the initial action and the optimal offline solution of problem (4.2). It also follows from the proof that this distance is related to the bound on the fit. Thus, the closer we start from this action the tighter the bounds will be. We next show that similar results holds for the saddle point dynamics if we consider the notion of saturated fit in lieu of fit.

Corollary 6. *Let the hypothesis of Theorem 10 hold. Let $\delta > 0$ and let $\bar{\mathcal{F}}_T$ be the saturated fit defined in* (4.5)*. Then, the solution of the dynamical system* (4.32) *and* (4.33)*, when $f(t,x)$ is replaced by $\bar{f}_\delta(t,x)) = \max\{f(t,x), -\delta\}$ achieves a regret satisfying* (4.62) *and saturated fit that is bounded by*

$$\bar{\mathcal{F}}_{T,i} \leq \mathcal{O}\left(\sqrt{KT}, K^0\right). \tag{4.82}$$

Proof. Same as Corollary 5. □

The above result establishes that a trajectory that follows the saddle point dynamics for the environment defined by $\bar{f}_\delta(t,x)$ achieves bounded saturated fit. This means that it is possible to adapt the controller (4.32) and (4.33), so that the growth of the fit is controlled while not alternating between periods of large under and over satisfaction of the constraints. In the next section we evaluate the performance of the saddle point controller, after a pertinent remark on the selection of the gain.

Remark 10 (**Gain depending on the Time Horizon).** *If it were possible to select the gain as a function of the time horizon T, fit could be bounded by a constant that does not grow with T. Take* (4.74) *and choose $\bar{\lambda} = e_i T$, where e_i is the i-th component of the canonical base of $\mathbb{R}^m$ we have that*

$$T \int_0^T f_i(t, x(t))dt \leq \frac{V_{x^*, Te_i}(x(0), \lambda(0))}{K} + KT. \tag{4.83}$$

With this selection of $\bar{\lambda}$ the function $V_{x^*,Te_i}(x(0), \lambda(0)))$ grows like T^2. Dividing both sides

of the above equation by T we have that the i-th component of the fit is bounded by

$$\mathcal{F}_{T,i} \leq \frac{\mathcal{O}(T)}{K} + K. \tag{4.84}$$

If the gain is set to have order $\Omega(T)$, the right hand side of (4.84) becomes of order $\mathcal{O}(T^0)$. This means that fit can be bounded by a constant that does not depend on T.

4.5 Numerical experiments

We evaluate performance of the saddle point algorithm defined by (4.32)-(4.33) in the solution of the shepherd problem introduced in Section 4.2.2. We determine sheep paths using a perturbed polynomial characterization akin to the one in (4.10). Specifically, letting $p_j(t)$ be elements of a polynomial basis, the path $y_i(t) = [y_{i,1}(t), y_{i,2}(t)]^\top$ of the i–th sheep is given by

$$y_{i,k}(t) = \sum_{j=0}^{n_i-1} y_{i,k,j} p_j(t) + w_{i,k}(t), \tag{4.85}$$

where $k = 1, 2$ denotes different path components, n_i the dimension of the base that parameterizes the path followed by sheep i, and $y_{i,k,j}$ represent the corresponding n_i coefficients. The noise terms $w_{i,k}(t)$ are Gaussian white with zero mean, standard deviation σ and independent across components and sheep. Their purpose is to obtain more erratic paths.

To determine $y_{i,k,j}$ we make $w_{i,k}(t) = 0$ in (4.85) and require all sheep to start at $y_i(0) = [0, 0]^\top$ and finish at $y_i(T) = [1, 1]^\top$. A total of L random points $\{\tilde{y}_l\}_{l=1}^L$ are then drawn independently and uniformly at random in the unit box $[0, 1]^2$. Sheep $i = 1$ is required to pass through points $\tilde{y}_l$ at times $lT/(L+1)$, i.e., $y_1(lT/(L+1)) = \tilde{y}_l$. For each of the other sheep $i \neq 1$ we draw L random offsets $\{\Delta\tilde{y}_{i,l}\}_{l=1}^L$ uniformly at random from the box $[-\Delta, \Delta]^2$ and require the i–th sheep path to satisfy $y_i(lT/(L+1)) = \tilde{y}_l + \Delta\tilde{y}_{i,l}$. Paths $y_i(t)$ are then chosen as those that minimize the path integral of the acceleration squared subject to the constraints of each path

$$\begin{aligned} y_i^* = &\operatorname{argmin} \int_0^T \|\ddot{y}_i(t)\|^2 dt, \\ &\text{s.t.} \quad y_i(0) = [0, 0]^\top, \quad y_i(T) = [1, 1]^\top, \\ &\qquad y_i(lT/(L+1)) = \tilde{y}_l + \Delta\tilde{y}_{i,l}, \end{aligned} \tag{4.86}$$

where, by construction $\Delta\tilde{y}_{1,l} = 0$. The paths in (4.86) can be computed as solutions of a quadratic program [85]. Let $y_i^*(t)$ be the trajectory given by (4.85) when we set $y_{i,k,j} = y_{i,k,j}^*$. We obtain the paths $y_{i,k}(t)$ by adding $w_{i,k}(t)$ to $y_i^*(t)$.

In subsequent numerical experiments we consider $m = 5$ sheep, a time horizon $T = 1$,

and set the proximity constraint in (4.11) to $r_i = 0.3$. We use the polynomial basis $p_j(t) = t^j$ in both, (4.10) and (4.85). The number of basis elements in both cases is set to $n = n_i = 30$. To generate sheep paths we consider a total of $L = 3$ randomly chosen intermediate points, set the variation parameter to $\Delta = 0.1$, and the perturbation standard deviation to $\sigma = 0.1$. These problem parameters are such that the environment is most likely viable in the sense of Definition 4. We check that this is true by solving the offline feasibility problem. If the environment is not viable a new one is drawn before proceeding to the implementation of (4.32)-(4.33).

We emphasize that while the path of the sheep is known to us, the information is not used by the controller. The latter is only fed information of the position of the sheep at the current time, which it uses to evaluate the environment functions $f_i(t, x)$, their gradients $f_{ix}(t, x)$ and the gradient of $f_0(t, x)$. In this example we do not assume any constraints on the maximum speed that the agent can achieve, therefore the gain K in (4.32)-(4.33) can be set to have any value.

4.5.1 Strongly feasible trajectories

We consider a problem without optimality criterion in which case (4.32)-(4.33) simplifies to (4.34)-(4.33) and the strong feasibility result in Theorem 9 applies. The system's behavior is illustrated in Figure 4.2 when the gain is set to $K = 50$. In this problem the average and maximal speed of the sheep is $5.1km/h$ and $14.8km/h$ respectively while for the shepherd these are $6.1km/h$ and $18.3km/h$ for the selected gain. This speeds are in he range of reasonable velocities for this particular problem. A qualitative examination of the sheep and shepherd paths shows that the shepherd succeeds in following the herd. A more quantitative evaluation is presented in Figure 4.3 where we plot the instantaneous constraint violation $f_i(t, x(t))$ with respect to each sheep for the trajectories $x(t)$. Observe the oscillatory behavior that has the constraint violations $f_i(t, x(t))$ hovering at around $f_i(t, x(t)) = 0$. When the constraints are violated, i.e., when $f_i(t, x(t)) > 0$, the saddle point controller drives the shepherd towards a position that makes him stay within r_i of all sheep. When a constraint is satisfied we have $f_i(t, x(t)) < 0$. This drives the multiplier $\lambda_i(t)$ towards 0 and removes the force that pushes the shepherd towards the sheep (cf,. Figure 4.3). The absence of this force makes the constraint violation grow and eventually surpass the maximum tolerance $f_i(t, x(t)) = 0$. At this point the multipliers start to grow and, as a consequence, to push the shepherd back towards proximity with the sheep.

The behavior observed in Figure 4.3 does not contradict the result in Theorem 9 which gives us a guarantee on fit, not on instantaneous constraint violations. The components of the fit are shown in Figure 4.4(a) where we see that they are indeed bounded. Thus, the trajectory is feasible in the sense of Definition 5, even if the instantaneous problem's

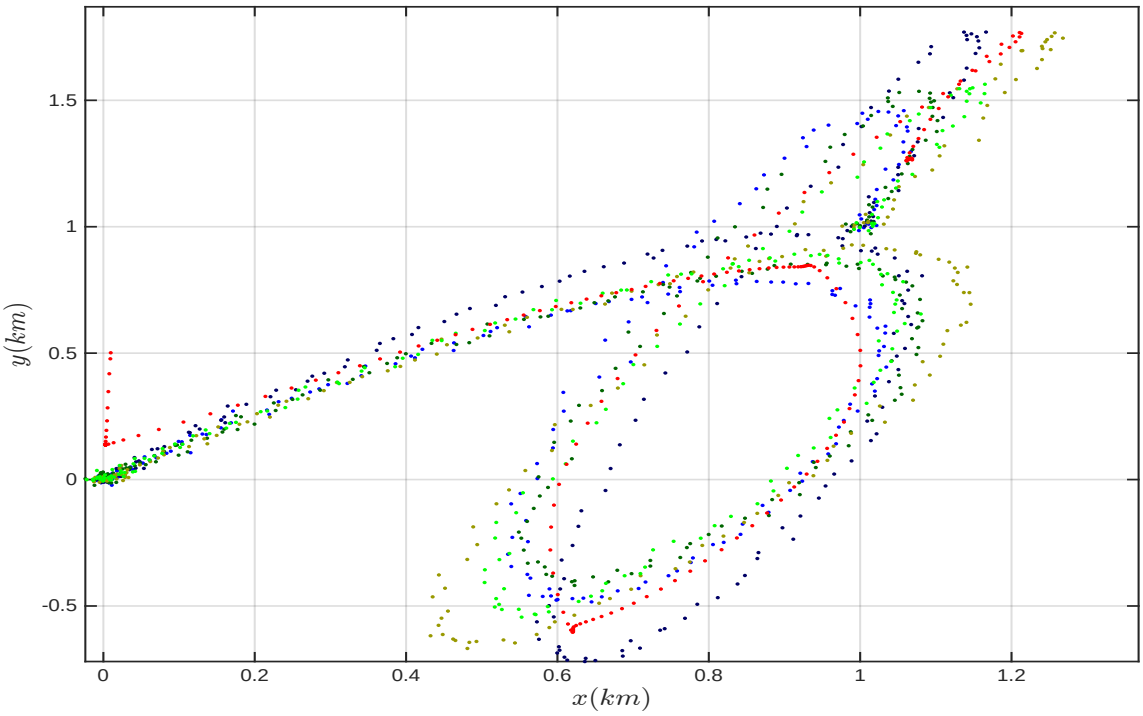


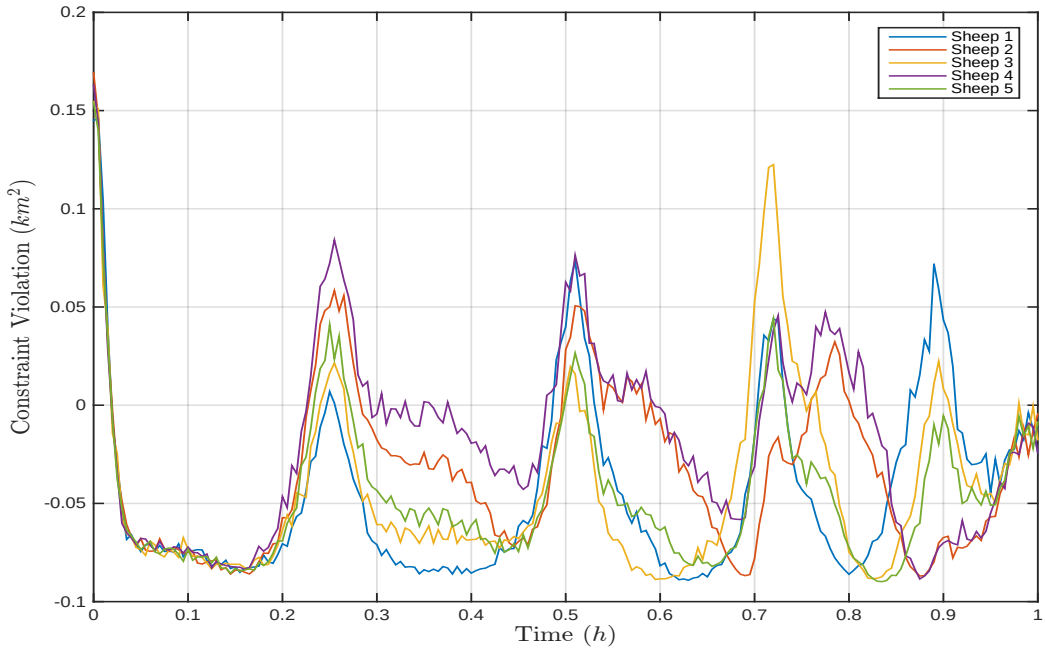
Figure 4.2: Path of the sheep and the shepherd for the feasibility-only problem (Section 4.5.1) when the gain of the saddle point controller is set to be $K = 50$. The shepherd succeed in following the herd since its path – in red – is close to the path of all sheep.

constraints are being violated at specific time instances. Further note that the fit is not only bounded but actually becomes negative. This is a consequence of the relatively large gain $K = 50$ which helps the shepherd to respond quickly to the sheep movements. The fit for a second experiment in which the gain is reduced to $K = 5$ is shown in Figure 4.4(b). In this case the fit stabilizes at a positive value. This behavior is expected because reducing K decreases the speed with which the shepherd can adapt to changes in the sheep paths. More to the point, the bound on the fit in Theorem 9 is inversely proportional to the gain K. The paths and instantaneous constraints violations for $K = 5$ are not shown but they are qualitatively similar to the ones shown for $K = 50$ in figures 4.2 and 4.3.

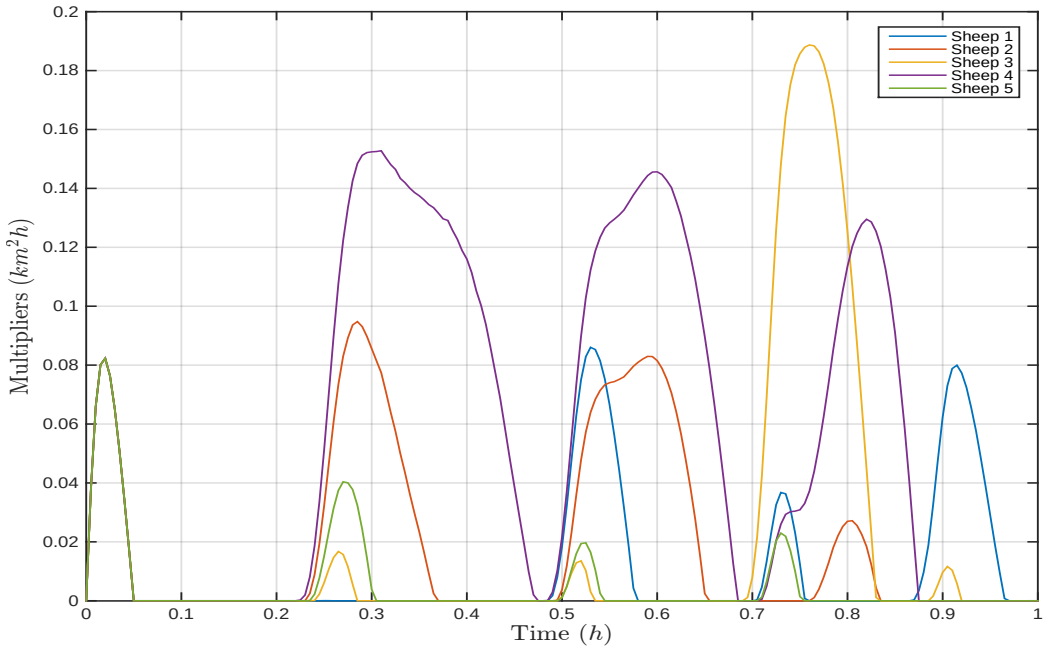
4.5.2 Preferred sheep problem

Besides satisfying the constraints in (4.11), the shepherd wishes to follow the first (black) sheep as close as possible. This translates into the optimality criterion (4.12). Since the sheep trajectories are viable the hypotheses of Theorem 10 hold. Thus, for a shepherd following the dynamics (4.32) and (4.33), the resulting trajectory is feasible and strongly optimal.

Given that the trajectory is guaranteed to be feasible, we expect to have the fit bounded

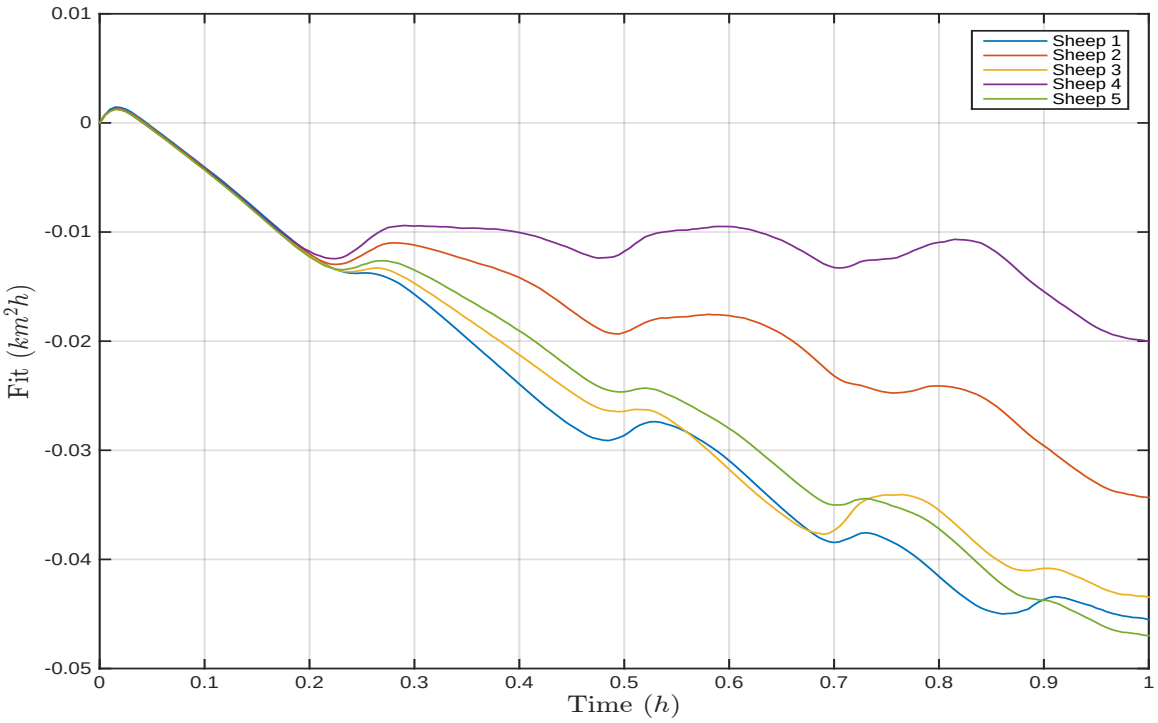


(a) Instantaneous constraint value.

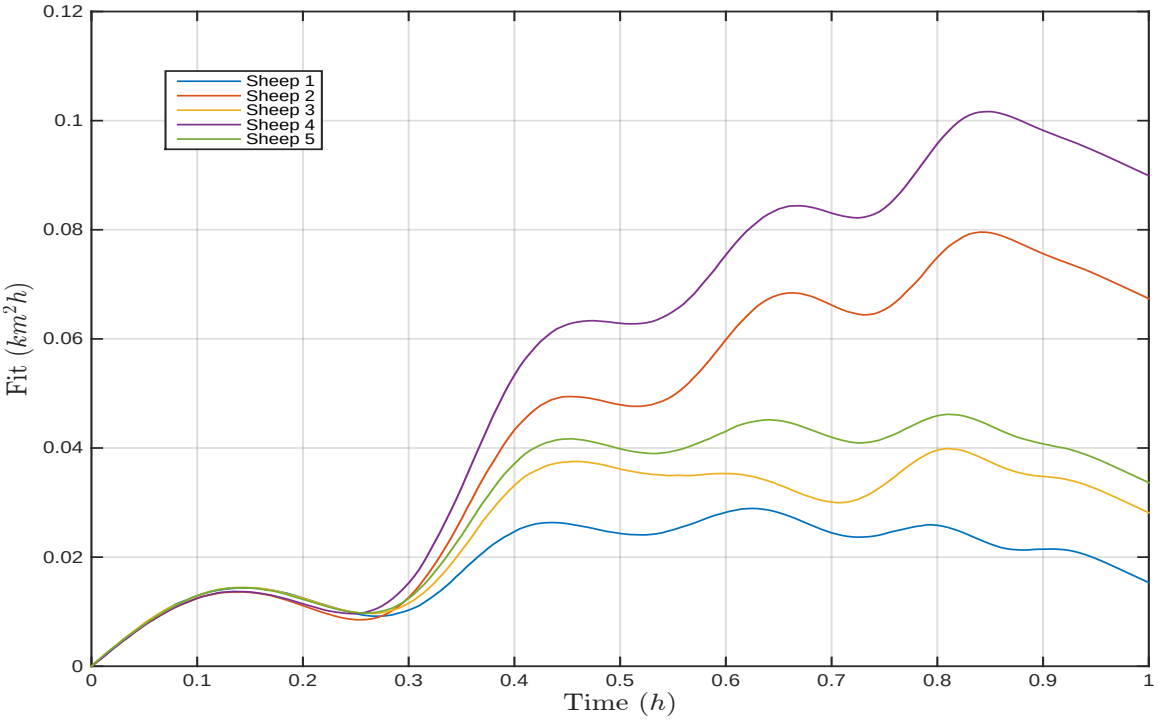


(b) Temporal evolution of the multipliers.

Figure 4.3: Relationship between the instantaneous value of the constraints and their corresponding multipliers for the feasibility-only problem (Section 4.5.1). At the times in which the value of a constraint is positive, its corresponding multiplier increases. When the value of the multipliers is large enough a decrease of the value of the constraint function is observed. Once the constraint function is negative the corresponding multiplier decreases until it reaches zero.



(a) Experiment with gain $K = 50$.



(b) Experiment with gain $K = 5$.

Figure 4.4: Fit $\mathcal{F}_T$ for two different controller gains in the feasibility-only problem (Section 4.5.1). Fit is bounded in both cases as predicted by Theorem 9. As is also predicted by Theorem 9, the larger the value of the gain K the smaller the bound on the fit of the shepherd's trajectory.

by a sublinear function of T. This does happen, as can be seen in the fit trajectories illustrated in Figure 4.5 where a gain $K = 50$ is used. In fact, the fit does not grow and is bounded by a constant for all time horizons T. The trajectory is therefore not only feasible but strongly feasible. This does not contradict Theorem 10 because strong feasibility implies feasibility. The reason why it's reasonable to see bounded fit here is that the objective function pushing the shepherd closer to the sheep is, in a sense, redundant with the constraints that push the shepherd to stay closer to all sheep. This redundancy can be also observed in the fact that the fit in this problem (cf., Figure 4.5) is smaller than the fit in the problem of Section 4.5.1 (cf., Figure 4.4(a)). To explain why this may happen, focus on the value of the multipliers in Figure 4.3(b) between, e.g., times $0.07\text{h} < t < 0.21\text{h}$. During this time the multipliers are equal to zero because all constraints are satisfied. As a consequence, the Lagrangian subgradient with respect to the action is identically zero in the time interval. In turn, this implies that the action is constant and no effort is made to reduce the value of the constraints. If the optimality criterion was present, the shepherd would be pushed towards the black sheep and fit would be further reduced.

The regret in this experiment when $K = 50$ is shown in Figure 4.6. Since the trajectory is strongly optimal as per Theorem 10, we expect regret to be bounded. This is the case in Figure 4.6 The path of the shepherd is not shown for this experiment as it is qualitatively analogous to the one in Figure 4.2 for the feasibility-only problem considered in Section 4.5.1.

4.5.3 Minimum acceleration problem

We consider, an environment defined by the distances between the shepherd and the sheep given by (4.11), with the minimum acceleration objective defined in (4.13). Since the construction of the target trajectories gives a viable environment we satisfy, again, the hypotheses of Theorem 10. Hence, for a shepherd following the dynamics given by (4.32) and (4.33), the action trajectory is feasible and strongly optimal. In this section the gain of the controller is set to $K = 50$.

A feasible trajectory implies that the fit must be bounded by a function that grows sublinearly with the time horizon T. Notice that this is the case in Figure 4.8. Periods of growth of the fit are observed, yet the presence of inflection points is an evidence of the growth being controlled. The fit in this problem is larger than the one in problem 4.5.2 (cf., figures 4.5 and 4.8). This result is predictable since the constraints and the objective function push the action in different directions. For instance, suppose that all constraints are satisfied and that the Lagrange multipliers are zero. Then, the subgradient of the Lagrangian is equal to the subgradient of the objective function. Hence the action will be modified trying to minimize the acceleration without taking the constraints (distance with

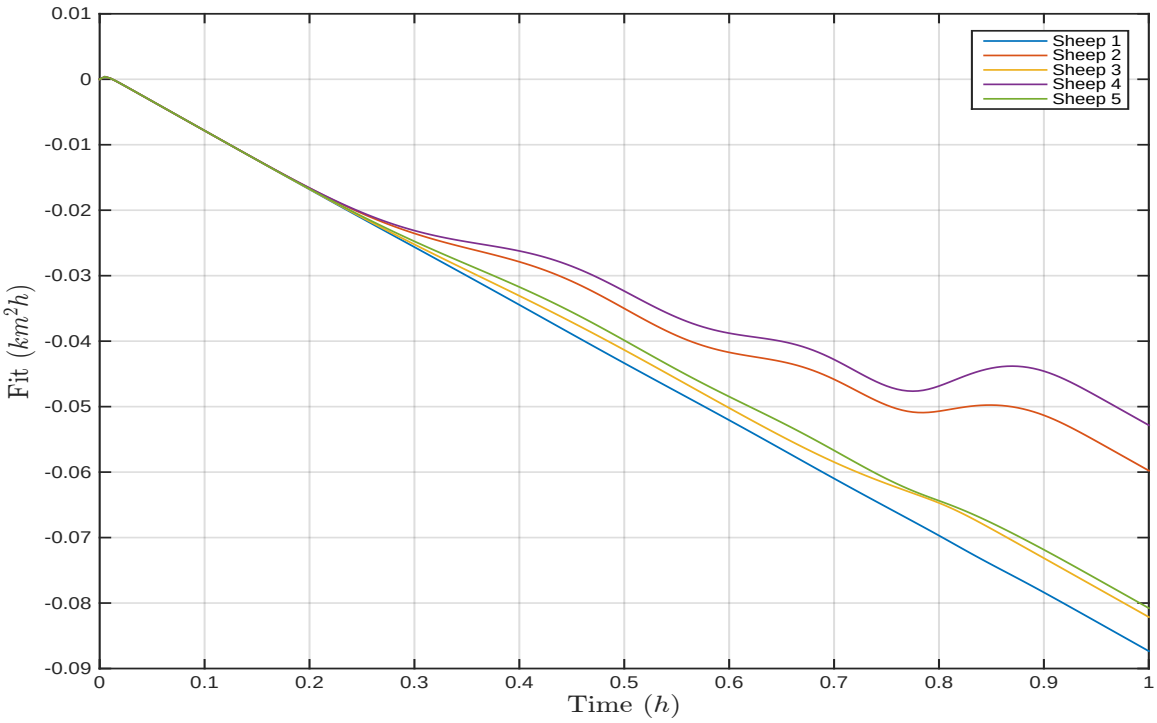


Figure 4.5: Fit $\mathcal{F}_T$ for the preferred sheep problem (Section 4.5.2) when the gain of the saddle point controller is set to be $K = 50$. As predicted by Theorem 10 the trajectory is feasible since the fit is bounded, and, in fact, appears to be strongly feasible. Since the subgradient of the objective function is the same as the subgradient of the first constrain the fit is smaller than in the pure feasibility problem (cf., Figure 4.4).

the sheep) into account. Hence, pushing the action to the boundary of the feasible set. In this problem, this translates into the fact that the shepherd does not follow the sheep as closely as in the problems in sections 4.5.1 and 4.5.2 (cf., Figure 4.7).

Since the trajectory is strongly optimal, we should observe a regret bounded by a constant. This is the case in Figure 4.9, where in fact we observe negative regret for some time intervals. Negative regret implies that the trajectory of the shepherd is incurring a total cost that is smaller than the one associated with the optimal solution. Notice that while the optimal fixed action minimizes the total cost as defined in (4.2) it does not minimize the objective at all times. Thus, by selecting different actions the shepherd can suffer smaller instantaneous losses than the ones associated with the optimal fixed action. If this is the case, regret – which is the integral of the difference between these two losses – can be negative.

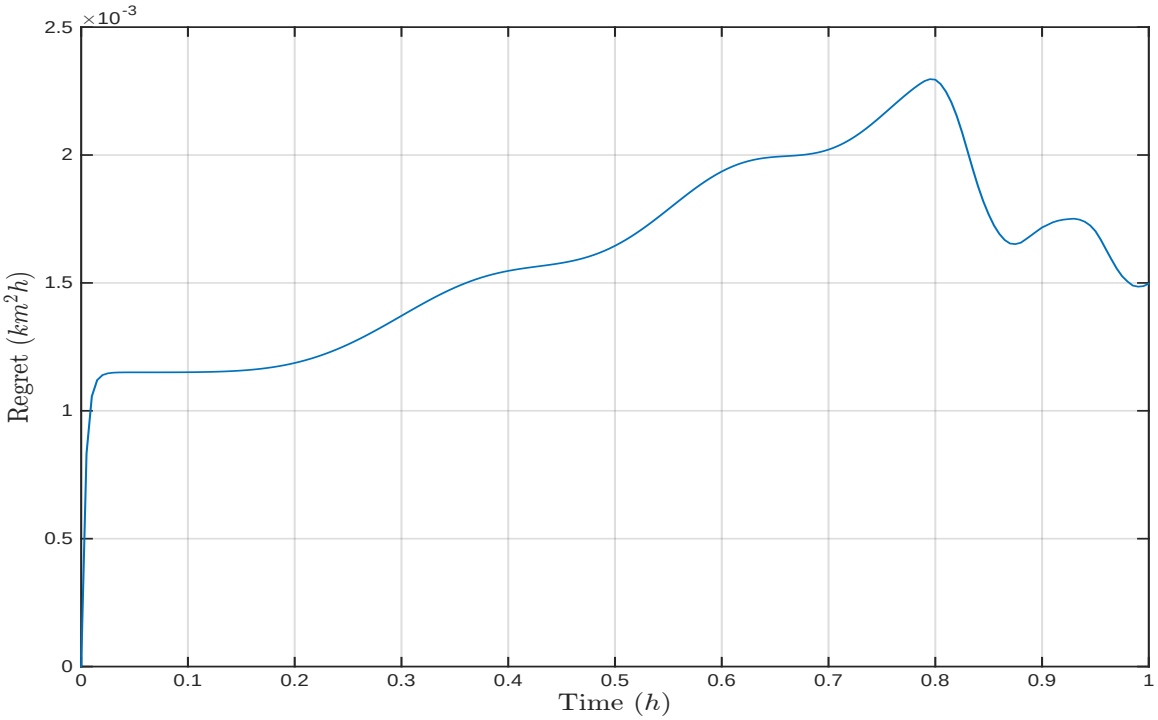


Figure 4.6: Regret $\mathcal{R}_T$ for the preferred sheep problem (Section 4.5.2) when the gain of the saddle point controller is set to be $K = 50$. The trajectory is strongly optimal, as predicted by Theorem 10, since the regret is bounded by a constant. The initial increment in the regret is due to the fact that the shepherd starts away from the first sheep while in the optimal offline trajectory would start close to it.

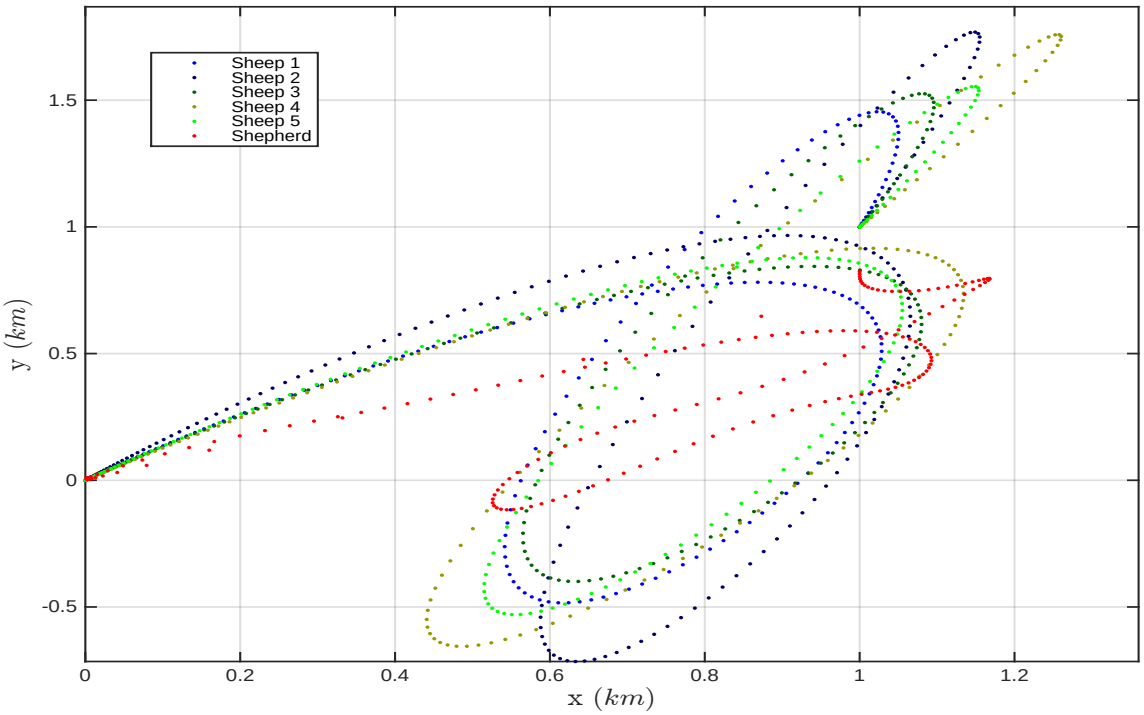


Figure 4.7: Path of the sheep and the shepherd for the minimum acceleration problem (Section 4.5.3) when the gain of the saddle point controller is set to be $K = 50$. Observe that the shepherd path – in red – is not as close to the path of the sheep as in Figure 4.2. This is reasonable because the objective function and the constraints push the shepherd in different directions.

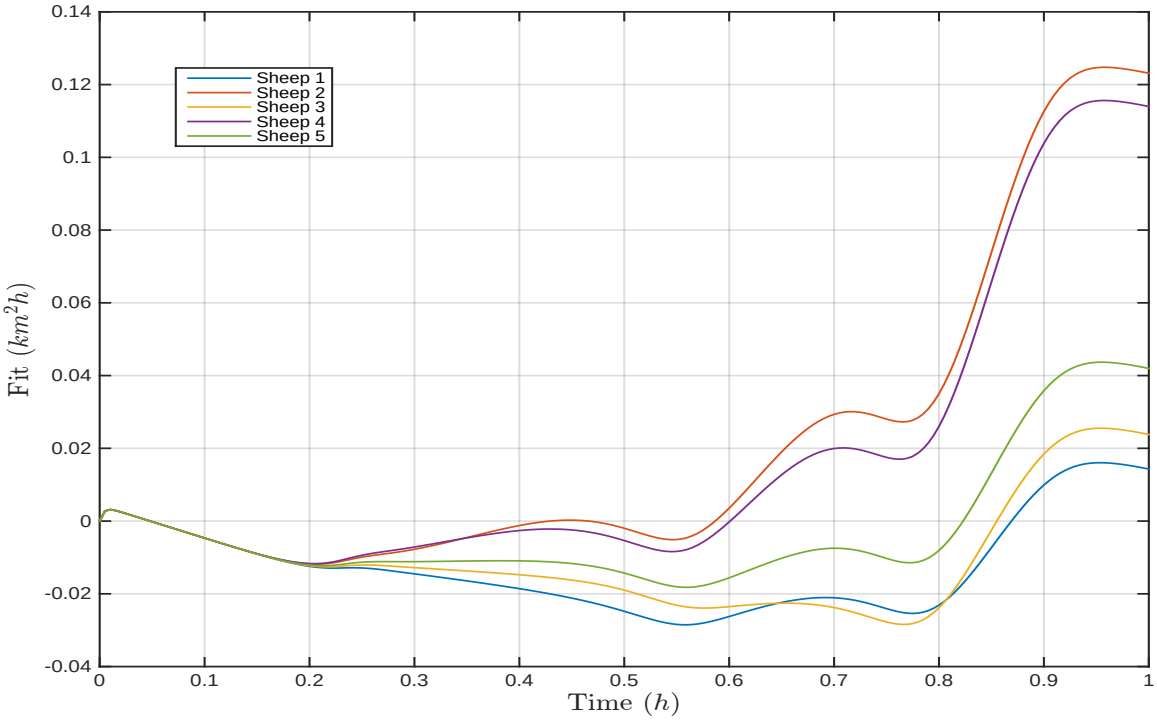


Figure 4.8: Fit $\mathcal{F}_T$ for the minimum acceleration problem (Section 4.5.3) when the gain of the saddle point controller is set to $K = 50$. Since the fit is bounded, the trajectory is feasible in accordance with Theorem 10. Since the gradient of the objective function and the gradient of the feasibility constraints tend to point in different directions, the fit is larger than in the preferred sheep problem (cf., Figure 4.5).

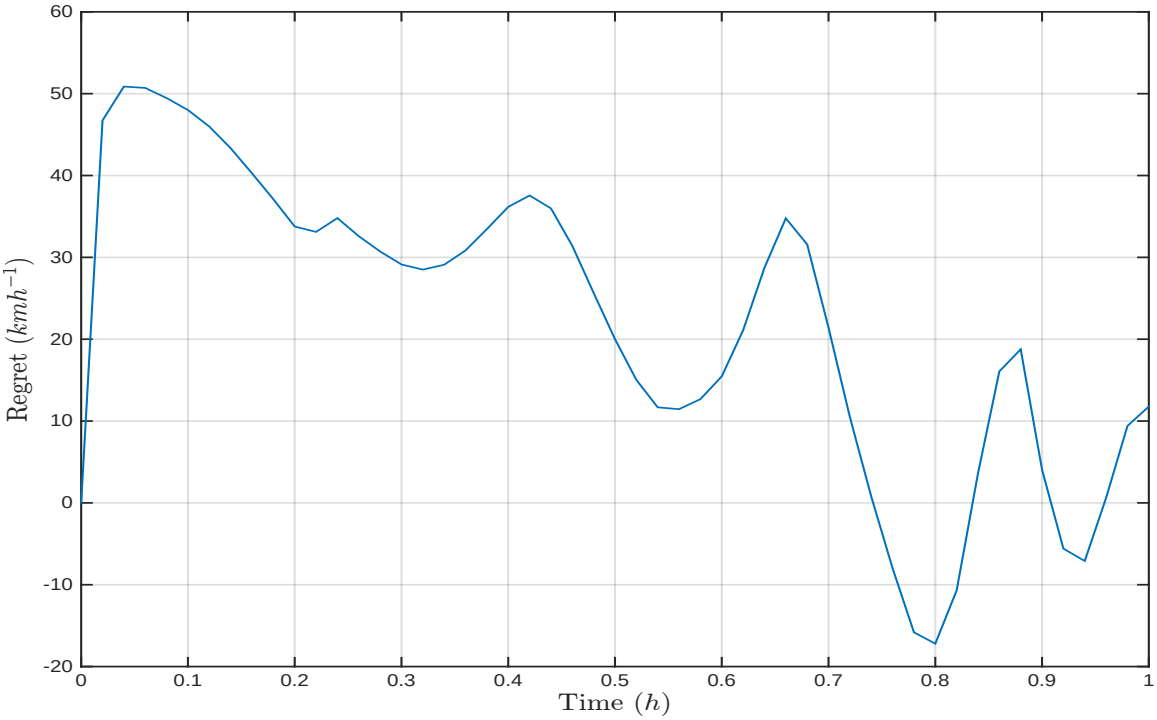


Figure 4.9: Regret $\mathcal{R}_T$ for the minimum acceleration problem (Section 4.5.3) when the gain of the saddle point controller is set to be $K = 50$. The trajectory is strongly optimal as predicted by Theorem 10. Observe that regret is negative due to the fact that the agent is allowed to select different actions at different times as opposed to the clairvoyant player that is allowed to select a fixed action.

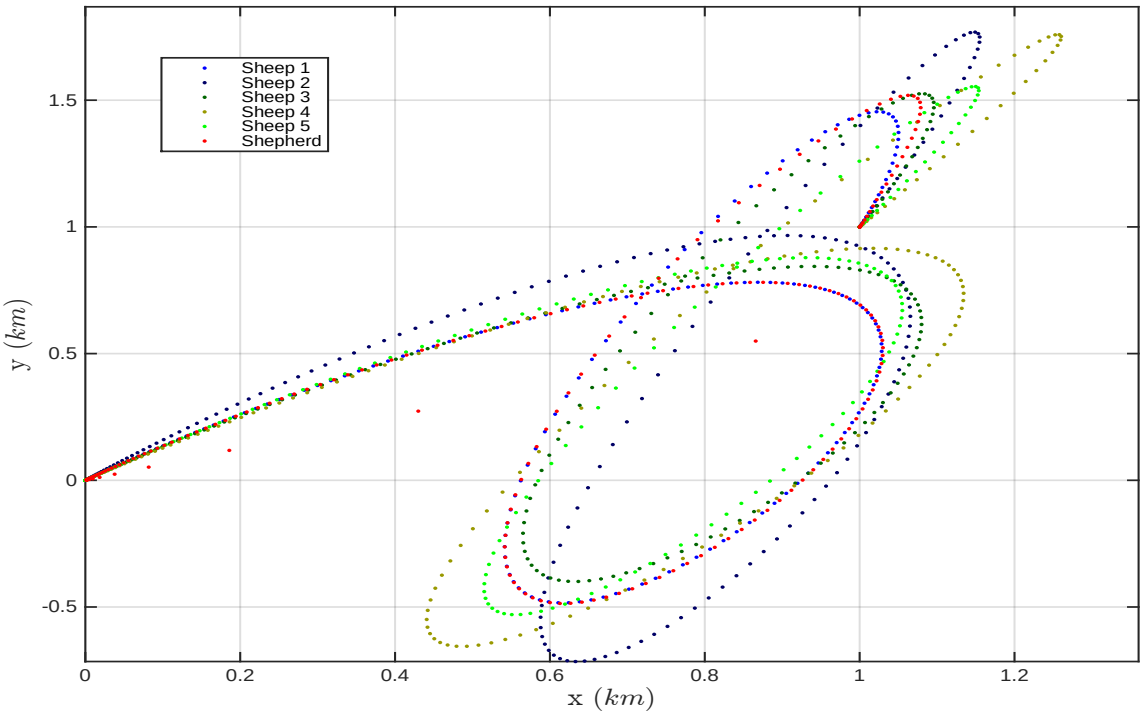


Figure 4.10: Path of the sheep and the shepherd for preferred sheep problem when saturated fit is considered (Section 4.5.4) and the gain of the saddle point controller is set to be $K = 50$. The shepherd succeed in following the herd since its path – in red – is close to the path of all sheep.

4.5.4 Saturated Fit

We apply the modified saddle point algorithm in the setting of Section 4.5.2 so to consider the saturated fit [cf., (4.5)] in lieu of the fit. Since the construction of the target trajectories gives a viable environment the hypotheses of Corollary 6 are satisfied. Hence for a shepherd following the dynamics given by (4.32) and (4.33), the trajectories are such that have saturated fit bounded by a function that grows sub linearly and bounded regret. For the simulation in this section the gain of the controller is set to $K = 50$. Observe that the shepherd succeeds in following the herd, since his path remains close to the sheep (cf., Figure 4.10). As predicted by the Corollary 6 the fit of the trajectory is bounded by a function that grows sub linearly and the regret is bounded by a constant as it can be observed in figures 4.11 and 4.12 respectively. Further notice that the regret in this scenario is similar to the regret of the trajectory in the preferred sheep problem (cf., Section 4.5.2).

4.6 Conclusion

We considered a continuous time environment in which an agent must select actions to satisfy a set of constraints that are time varying and unknown a priori. We defined a viable

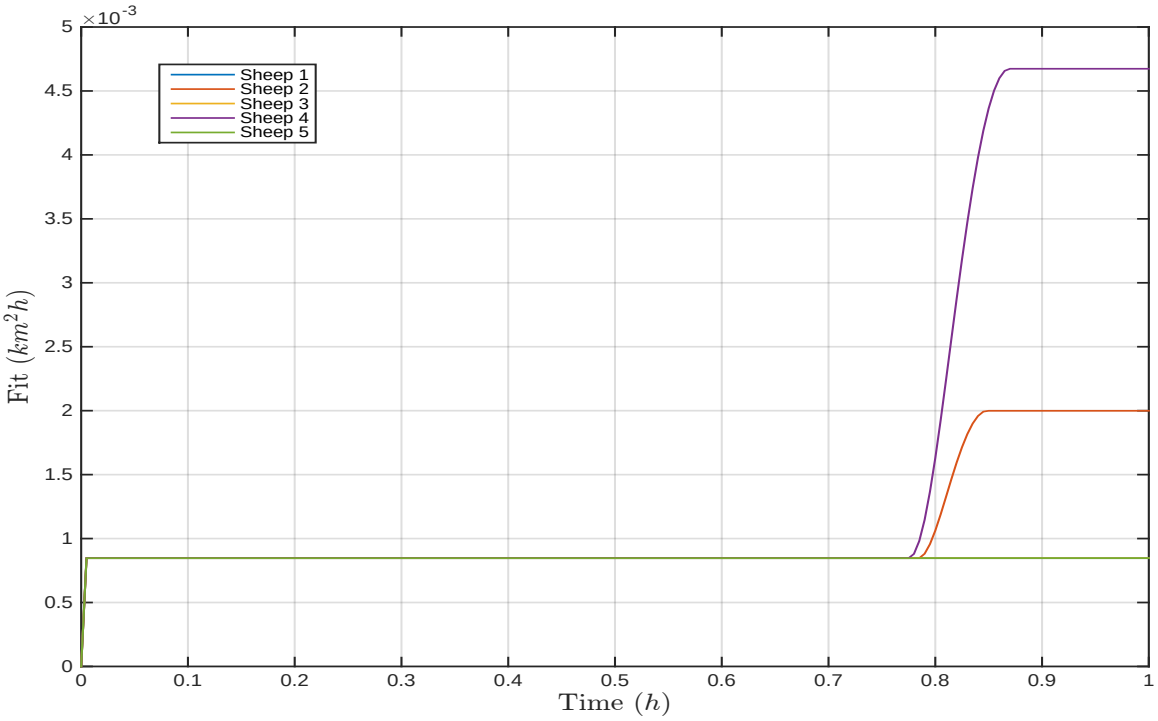


Figure 4.11: Saturated fit $\mathcal{F}_T^{sat}$ for the preferred sheep problem (Section 4.5.4) when the gain of the saddle point controller is set to $K = 50$. Since the saturated fit grows sublinearly in accordance with Corollary 6, the trajectory is feasible.

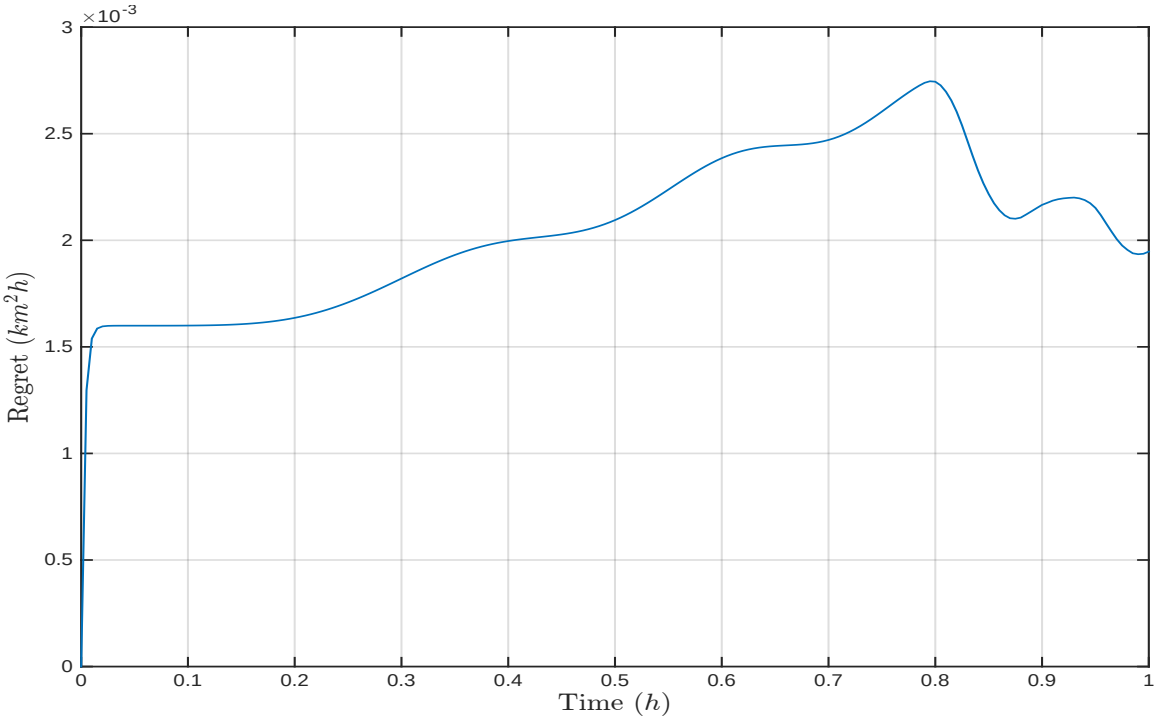


Figure 4.12: Regret $\mathcal{R}_T$ for the preferred sheep problem when saturated fit is considered (Section 4.5.4)and the gain of the saddle point controller is set to be $K = 50$. The regret is bounded as predicted by Corollary 6 and therefore the trajectory is strongly optimal. Notice that regret in this case is identical to regret in the preferred sheep problem when regular fit is considered (cf., Figure 4.6).

environment as one in which there is a fixed action that satisfies the constraints at all times. We defined the fit as the cumulated constraint violation and the notions of feasible and strongly feasible trajectories. Feasible trajectories are such that the fit is bounded by a constant independent of the time horizon, and strongly feasible trajectories are such that the fit is bounded by a sublinear function of the time horizon. An objective function was considered to select a strategy that meets an optimality criterion and we defined regret in continuous time as the difference between the cumulative costs of the agent and the best clairvoyant agent. We then defined strongly optimal trajectories as those for which the regret is bounded by a constant that is independent of the time horizon.

We proposed an online version of the saddle point controller of Arrow-Hurwicz to generate trajectories with small fit and regret. We showed that for any viable environment the trajectories that follow the dynamics of this controller are: (i) Strongly feasible if no optimality criterion is considered. (ii) Feasible and strongly optimal when an optimality criterion is considered. Numerical experiments on a shepherd that tries to follow a herd of sheep support these theoretical results. Algorithms inspired in the online saddle point have extended the applicability of such concepts to distributed settings [19, 20, 70].

Chapter 5

Lagrange Multipliers as price interfaces

Define an environment as a set of convex constraint functions and a cost function that is also convex. An agent operating in such environment intend to select optimal actions that are feasible. In cases where the problem is feasible, such action can be found via the Arrow Hurwicz algorithm, that consists in finding the saddle point of the Lagrangian associated to the optimization problem. This controller and its variations – stochastic models or viability (Chapter 4) operate by computing the gradient of all the constraints and updating the action along the negative of a weighted combination of these gradients. The coefficients of this linear combination are increased when the constraints are violated and decreased when they are satisfied. If a constraint is far from being satisfied it means that its satisfaction is relatively difficult and the corresponding multiplier will be large. In that sense, weights can be thought of as prices for satisfying a given constraint. In this chapter, we consider the situation where the optimization problem is not feasible and hence, the multipliers for such algorithm would diverge. To overcome this limitation, we modify the saddle point algorithm by introducing a slack variable that is increased when the constraints are being violated and reduced if the slack grows too much. We show that this modification converges to a point for which the limit of the slack gives us a measure of the relative hardness of satisfying each constraint.

5.1 Introduction

As we discussed in the previous chapter, saddle point algorithms [4] and their stochastic versions [65] allows us to solve convex constrained optimization problems in cases where the agent can measure the constraints and the objective functions exactly or when it there is a probabilistic model for such functions. The main contribution of Chapter 4 is to show

that an online version of such algorithm succeeds in doing the same in settings where the only information available to the agent is that there exists an action for which the problem is solvable. In this chapter, we consider the situation in which the latter information is not available to the agent and it is his task to identify whether the problem is feasible or not. In cases where it is not the case, we would like to identify which of the constraints are harder to satisfy so that the agent can remove it and fall back into a laxer notion of feasibility.

In all three cases, the algorithms are such that they compute gradients for all of them. The coefficients of this linear combination are adapted according to how far from being satisfied the respective constraint is. In that sense, the weights can be thought of as prices. If a constraint is far from being satisfied it means that its satisfaction is relatively difficult and that a large coefficient, i.e., a large price, is justified in the corresponding element of the linear combination. For instance, let us consider the surveillance problem in which we are interested in tracking several obstacles. Suppose that there is no way of being close to all of the targets, then at least one of the multipliers will increase for all times. The logical reasoning part of the system can use this information to decide a different policy, for instance it could change the problem of being at a given distance of all the targets for a new problem stated as being at a given distance of the target whose multipliers are bounded and adding an optimality criteria given by being as close as possible to the remaining targets. The problem of deciding the policy that must be accomplished is the task of the logical reasoning part of the system, and as discussed the information arising from the low-level control is a fundamental piece of information to effectively chose the strategy to follow.

In particular, we propose a modification to the saddle point algorithm, where we introduce a slack for every constraint. The slack is updated in the following way; they are increased when the multipliers are positive, i.e., when constraints are violated and they are decreased if the slack increases much (Section 5.2) The algorithm is such that it converges to the primal-dual optimal solution for a relaxed problem, this slack is such that proportional to the gradient of the optimal cost with respect to the slack. Larger slacks mean then, that if we try to reduce the slack, the cost in which we incur is large and hence it is a measure of the difficulty in satisfying such constraints (Section 5.3). In Section 5.4 we instead of having deterministic objective functions and constraints a stochastic model is available and we are interested in solving the problem in expectation. The solution proposed in Section 5.5 is a stochastic approximation of the deterministic case and the same convergence guarantees can be provided with probability one.

5.2 Problem Formulation

Let $f_0 : \mathbb{R}^n \to \mathbb{R}$ and $f : \mathbb{R}^n \to \mathbb{R}^m$ and let us define the following optimization problem.

$$\begin{aligned} p^* := \min_{x \in \mathbb{R}^n} \; & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f(\mathbf{x}) \preceq 0. \end{aligned} \tag{5.1}$$

The objective of this work is to determine whether the previous problem is feasible or not, i.e., if there exists $\mathbf{x}^\dagger \in \mathbb{R}^n$ such that $f(\mathbf{x}^\dagger) \preceq 0$. In cases where the latter does not hold we would like to solve a relaxed version of the problem, where we can allow for some constraint violation. But most importantly, we want to identify which of the constraints is the hardest to satisfy, so the agent can decide which constraints should be removed from the problem and fall back into a laxer notion of feasibility. A possibility to understand the relative difficulty of satisfying different constraints is trhough Duality Theory. Each dual variable can be interpreted as "cost" or "price" associated to satisfying a given constraint and hence, the larger the value of the dual variable associated to a constraint, the harder it is to satisfy it. To formalize these ideas, introduce the following slack variable $\mathbf{s} \in \mathbb{R}^m_+$ and consider the following relaxation of the problem (5.1)

$$\begin{aligned} p^*(\mathbf{s}) := \min_{x \in \mathbb{R}^n} \; & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_0(\mathbf{x}) - \mathbf{s} \preceq 0, \end{aligned} \tag{5.2}$$

and its associated Lagrangian

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) := f_0(\mathbf{x}) + \boldsymbol{\lambda}^\top \left(f(\mathbf{x}) - \mathbf{s} \right), \tag{5.3}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^m_+$. Likewise, let us define the dual function $g(\boldsymbol{\lambda}, \mathbf{s})$

$$g(\boldsymbol{\lambda}, \mathbf{s}) := \min_{\mathbf{x} \in \mathbb{R}^n} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}). \tag{5.4}$$

The dual function is a lower bound for the primal function [16, Section 5.1.3], this is, for all $\boldsymbol{\lambda}$ and $\mathbf{s}$, we have that

$$g(\boldsymbol{\lambda}, \mathbf{s}) \leq p^*(\mathbf{s}). \tag{5.5}$$

The dual problem is then defined as the best lower bound for the previous problem

$$d^*(\mathbf{s}) := \max_{\boldsymbol{\lambda} \in \mathbb{R}^m_+} g(\boldsymbol{\lambda}, \mathbf{s}). \tag{5.6}$$

Notice that for $\mathbf{s} = 0$ we recover the original primal problem (5.1). Duality Theory can allow us to establish whether a problem is feasible or not by looking at the dual problem. Indeed, if the the dual problem is unbounded above, i.e., $d^* = \infty$ it implies that $p^*(\mathbf{s}) = \infty$, hence the primal problem is infeasible. Because the dual function is concave – it is the point-wise minimum of linear functions– when the dual problem is unbounded it means that the dual solution $\lambda^*(\mathbf{s})$

$$\boldsymbol{\lambda}^*(\mathbf{s}) := \underset{\boldsymbol{\lambda}\mathbb{R}^m_+}{\operatorname{argmax}}\, g(\boldsymbol{\lambda}, \mathbf{s}). \tag{5.7}$$

is also unbounded. The converse holds when strong duality does, i.e., when $d^*(\mathbf{s}) = p^*(\mathbf{s})$. Conditions for strong duality to hold are that $f_0(\mathbf{x})$ and $f(\mathbf{x})$ are convex functions and that there exists a strictly feasible point (see e.g., [16, Section 5.3.2]). We formalize this assumptions next for future reference.

AS9. *We assume $f : \mathbb{R}^n \to \mathbb{R}^m$ is convex and $f_0 : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex.*

AS10. *There exists $\mathbf{x}^\dagger \in \mathbb{R}^n$ and $\mathbf{s}^\dagger \in \mathbb{R}^m_+$ such that $f(\mathbf{x}^\dagger) - \mathbf{s}^\dagger \prec 0$.*

Under Assumptions 9 and 10, for any $\mathbf{s} \succeq \mathbf{s}^\dagger$, it also holds that the primal-dual solution $(\mathbf{x}^*(\mathbf{s}), \boldsymbol{\lambda}^*(\mathbf{s}))$ is a saddle point of the Lagrangian (5.3) [16, Section 5.4.2]. The latter means, that for all $\mathbf{x} \in \mathbb{R}^n$ and $\boldsymbol{\lambda} \in \mathbb{R}^m_+$ it holds that

$$\mathcal{L}(\mathbf{x}^*(\mathbf{s}), \boldsymbol{\lambda}, \mathbf{s}) \leq \mathcal{L}(\mathbf{x}^*(\mathbf{s}), \boldsymbol{\lambda}^*(\mathbf{s}), \mathbf{s}) \leq \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*(\mathbf{s}), \mathbf{s}). \tag{5.8}$$

The latter can be found via the Arrow-Hurwicz algorithm [4]. For a fixed $\mathbf{s}$, the algorithm is such that it descends in $\mathbf{x}$ along the direction of the negative gradient of the Lagrangian with respect to $\mathbf{x}$

$$\dot{\mathbf{x}} = -\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) = -\left(\nabla f_0(\mathbf{x}) + \sum_{i=1}^{m} \boldsymbol{\lambda}_i \nabla f(\mathbf{x}) \right), \tag{5.9}$$

and it ascends in $\boldsymbol{\lambda}$ along the direction of the gradient of the Lagrangian with respect to $\boldsymbol{\lambda}$

$$\dot{\boldsymbol{\lambda}} = \Pi_{\mathbb{R}^m_+} \left(\boldsymbol{\lambda}, \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) \right) = \Pi_{\mathbb{R}^m_+} \left(\boldsymbol{\lambda}, f(\mathbf{x}) - \mathbf{s} \right), \tag{5.10}$$

where $\Pi_{\mathbb{R}^m_+}(\cdot, \cdot)$ refers to a projected dynamical system over the positive orthant of $\mathbb{R}^m$. This projection is introduced to ensure that the Lagrange multipliers are always non-negative. The intuition behind the previous algorithm is that as long as a constraint i is satisfied, its corresponding Lagrangian multiplier is zero, i.e., $\boldsymbol{\lambda}_i = 0$. However, if said constraint is being violated, then $f_i(\mathbf{x}) - \mathbf{s} > 0$ and the value of the corresponding multiplier is increased. The intuition behind the update of the primal variable is that it descends along a weighted combination of the gradients of the objective function and the constraints, so to reduce the

value of all the functions. The specific values of the weights are given by each $\boldsymbol{\lambda}_i$. Hence, the relative strength that each gradient has is related with how much the constraint is being violated.

The main drawback with Arrow-Hurwciz algorithm in this context is that the value $\mathbf{s}^\dagger$ that makes the problem (5.2) feasible, is not known beforehand. To overcome this limitation, we propose to update $\mathbf{x}$ and $\boldsymbol{\lambda}$ as in the classic Arrow-Hurwicz algorithm (5.9)–(5.10), with the following update in the slack variable $\mathbf{s}$

$$\dot{\mathbf{s}} = \mathbf{K}\left(\mathbf{K}\boldsymbol{\lambda} - \mathbf{s}\right), \tag{5.11}$$

where $\mathbf{K} \succ 0$ is a matrix gain. The intuition behind the previous update is that as long as the constraints in the relaxed problem (5.2) are satisfied, i.e. $\boldsymbol{\lambda}_i = 0$ the value of the slack can be reduced. However, if a constraint is no longer satisfied, we will have $\boldsymbol{\lambda}_i > 0$ which will increase the slack of the corresponding constraint. In the next section we show that the solutions of (5.9)–(5.11) are such that $\lim_{t\to\infty} \mathbf{s}(t) = \mathbf{s}_\infty$ and such that $\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{x}^*(\mathbf{s}_\infty)$ and $\lim_{t\to\infty} \boldsymbol{\lambda}(t) = \boldsymbol{\lambda}^*(\mathbf{s}_\infty)$. For the slack variable to converge it must hold that $\dot{\mathbf{s}} = 0$, which can only happen if (cf., (5.11))

$$\boldsymbol{\lambda} = \mathbf{K}^{-1}\mathbf{s}. \tag{5.12}$$

To understand the importance of the previous condition, we need to refer back to the idea that dual variables are costs associated to satisfying a constraint. Formally, we have that (cf., [16, Section 5.6.2.])

$$\nabla_{\mathbf{s}} p^*(\mathbf{s})\Big|_{\mathbf{s}=\mathbf{s}_\infty} = -\boldsymbol{\lambda}^*(\mathbf{s}_\infty). \tag{5.13}$$

The latter relationship, combined with the equilibrium condition (5.12) implies that the slack variable in the limit satisfies

$$\nabla_{\mathbf{s}} p^*(\mathbf{s})\Big|_{\mathbf{s}=\mathbf{s}_\infty} = -\mathbf{K}^{-1}\mathbf{s}_\infty. \tag{5.14}$$

The latter condition allows us to analyze the relative hardness of satisfying given constraints. Notice that the gain Matrix $\mathbf{K}$ can be used to assign relative importance to the different constraints, however, if they are all equally important, we could think of having $\mathbf{K}$ being the identity matrix. In this case, the larger the slack it means that the derivative is larger in absolute value. Hence, a reduction of the slack produces a higher increase in the optimal cost.

In the next section we formalize the convergence results outlined here and in Section 5.4 we generalize these results to settings in which we are not able to evaluate the functions and their gradients, but we have access to a stochastic model about them.

5.3 Convergence of the modified saddle point algorithm

We start the convergence analysis by showing that the solutions of the modified saddle point (5.9)–(5.11) are bounded for all time.

Proposition 1. *Let $f_0 : \mathbb{R}^n \to \mathbb{R}$ and $f : \mathbb{R}^n \to \mathbb{R}^p$ satisfy assumptions 9 and 10. Then, the solutions of the dynamics* (5.9)–(5.11) *are bounded for all $t \in [0, \infty)$.*

Proof. From Lemma 18 it follows that it is possible to chose $\mathbf{s}^*$ with bounded norm such that the optimal dual variable for the optimization problem (5.2) with slack variable $\mathbf{s}^*$ satisfies

$$\boldsymbol{\lambda}^*(\mathbf{s}^*) = \mathbf{K}^{-1}\mathbf{s}^*. \tag{5.15}$$

Let $\mathbf{x}^*(\mathbf{s}^*)$ be the optimal primal variable for said problem and define the following function

$$U(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) = \frac{1}{2}\left(\|\mathbf{x} - \mathbf{x}^*(\mathbf{s}^*)\|^2 + \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*(\mathbf{s}^*)\|^2 + \|\mathbf{s} - \mathbf{s}^*\|^2_{\mathbf{K}^{-2}}\right), \tag{5.16}$$

where for a vector $\boldsymbol{\nu}$ the norm is $\|\boldsymbol{\nu}\|_{\mathbf{K}^{-2}}$ is defined as

$$\|\boldsymbol{\nu}\|_{\mathbf{K}^{-2}} = \left\|\mathbf{K}^{-1}\boldsymbol{\nu}\right\|^2. \tag{5.17}$$

Because $U(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})$ is radially unbounded, to show that the solutions are bounded, it suffices to show that $U(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})$ is non-increasing along the dynamics (5.9)–(5.11). To do so, take the time derivative of (5.16) with respect to time

$$\dot{U}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) = \left(\mathbf{x} - \mathbf{x}^*(\mathbf{s}^*)\right)^\top \dot{\mathbf{x}} + \left(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*(\mathbf{s}^*)\right)^\top \dot{\boldsymbol{\lambda}} + \left(\mathbf{s} - \mathbf{s}^*\right)^\top \mathbf{K}^{-2}\dot{\mathbf{s}}. \tag{5.18}$$

Substituting $\dot{\mathbf{x}}$, $\dot{\boldsymbol{\lambda}}$ and $\dot{\mathbf{s}}$ for their respective expressions (5.9)–(5.11) in the previous derivative yields

$$\begin{aligned}\dot{U}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) = &- \left(\mathbf{x} - \mathbf{x}^*(\mathbf{s}^*)\right)^\top \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) + \left(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*(\mathbf{s}^*)\right)^\top \Pi_{\mathbb{R}^m_+}\left(\boldsymbol{\lambda}, f(\mathbf{x}) - \mathbf{s}\right) \\ &+ \left(\mathbf{s} - \mathbf{s}^*\right)^\top \mathbf{K}^{-1}\left(\mathbf{K}\boldsymbol{\lambda} - \mathbf{s}\right).\end{aligned} \tag{5.19}$$

Because both, $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^*(\mathbf{s}^*)$ belong to $\mathbb{R}^m_+$ it follows from Lemma 1 [98] that the inner product $\left(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*(\mathbf{s}^*)\right)^\top \Pi_{\mathbb{R}^m_+}\left(\boldsymbol{\lambda}, f(\mathbf{x}) - \mathbf{s}\right)$ is upper bounded by $\left(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*(\mathbf{s}^*)\right)^\top \left(f(\mathbf{x}) - \mathbf{s}\right)$. Thus, the previous derivative can be in turn upper bounded by

$$\begin{aligned}\dot{U}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) \leq &- \left(\mathbf{x} - \mathbf{x}^*(\mathbf{s}^*)\right)^\top \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) + \left(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*(\mathbf{s}^*)\right)^\top \left(f(\mathbf{x}) - \mathbf{s}\right) \\ &+ \left(\mathbf{s} - \mathbf{s}^*\right)^\top \mathbf{K}^{-1}\left(\mathbf{K}\boldsymbol{\lambda} - \mathbf{s}\right).\end{aligned} \tag{5.20}$$

Notice that the gradient of the Lagrangian with respect to $\mathbf{x}$ is independent of the slack variable $\mathbf{s}$ (cf., (5.9)), hence it holds that $\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s}) = \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s}^*)$. Then, adding and subtracting $(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*(\mathbf{s}^*))^\top \mathbf{s}^*$ to the previous expression yields

$$\begin{aligned}\dot{U}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s}) = &-(\mathbf{x} - \mathbf{x}^*(\mathbf{s}^*))^\top \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s}^*) + (\boldsymbol{\lambda} - \boldsymbol{\lambda}^*(\mathbf{s}^*))^\top (f(\mathbf{x}) - \mathbf{s}^*) \\ &+ (\mathbf{s} - \mathbf{s}^*)^\top \left(\boldsymbol{\lambda} - \mathbf{K}^{-1}\mathbf{s} - (\boldsymbol{\lambda} - \boldsymbol{\lambda}^*(\mathbf{s}^*))\right).\end{aligned} \tag{5.21}$$

Because the Lagrangian is convex in $\mathbf{x}$ the inner product $-(\mathbf{x} - \mathbf{x}^*(\mathbf{s}^*))^\top \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s})$ can be upper bounded by

$$-(\mathbf{x} - \mathbf{x}^*(\mathbf{s}^*))^\top \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s}) \leq \mathcal{L}(\mathbf{x}^*(\mathbf{s}^*),\boldsymbol{\lambda},\mathbf{s}^*) - \mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s}^*). \tag{5.22}$$

Likewise, from (5.3) it follows that

$$(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*(\mathbf{s}^*))^\top (f(\mathbf{x}) - \mathbf{s}^*) = \mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s}^*) - \mathcal{L}(\mathbf{x},\boldsymbol{\lambda}^*(\mathbf{s}^*),\mathbf{s}^*). \tag{5.23}$$

Substituting (5.22) and (5.23) in (5.21) yields the following upper bound

$$\dot{U}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s}) \leq \mathcal{L}(\mathbf{x}^*(\mathbf{s}^*),\boldsymbol{\lambda},\mathbf{s}^*) - \mathcal{L}(\mathbf{x},\boldsymbol{\lambda}^*(\mathbf{s}),\mathbf{s}^*) + (\mathbf{s} - \mathbf{s}^*)^\top \left(\boldsymbol{\lambda}^*(\mathbf{s}^*) - \mathbf{K}^{-1}\mathbf{s}\right). \tag{5.24}$$

Because $\mathbf{x}^*(\mathbf{s}^*)$ and $\boldsymbol{\lambda}^*(\mathbf{s}^*)$ are primal dual solutions of the optimization problem (5.2) with slack $\mathbf{s}^*$, it follows from the saddle point property (cf., (5.8)) that $\mathcal{L}(\mathbf{x}^*(\mathbf{s}^*),\boldsymbol{\lambda},\mathbf{s}^*) - \mathcal{L}(\mathbf{x},\boldsymbol{\lambda}^*(\mathbf{s}),\mathbf{s}^*) \leq 0$. Hence, we have that

$$\dot{U}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s}) \leq (\mathbf{s} - \mathbf{s}^*)^\top \left(\boldsymbol{\lambda}^*(\mathbf{s}^*) - \mathbf{K}^{-1}\mathbf{s}\right). \tag{5.25}$$

Substituting $\boldsymbol{\lambda}^*(\mathbf{s}^*)$ for its expression (5.15), the previous inequality can be re written as

$$\dot{U}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s}) \leq (\mathbf{s} - \mathbf{s}^*)^\top \mathbf{K}^{-1} (\mathbf{s}^* - \mathbf{s}) \leq 0. \tag{5.26}$$

The latter shows that $U(\mathbf{x},\boldsymbol{\lambda},\mathbf{s})$ is non-increasing along the solutions of (5.9)–(5.11), hence completing the proof of the proposition. □

To show convergence of the dynamics to the point satisfying $(\mathbf{x}^*(\mathbf{s}^*),\boldsymbol{\lambda}^*(\mathbf{s}^*))$ with $\boldsymbol{\lambda}^*(\mathbf{s}^*) = \mathbf{K}^{-1}\mathbf{s}^*$ we definite the following function $V(\mathbf{x},\boldsymbol{\lambda},\mathbf{s})$ inspired on the analysis in [32]

$$V(\mathbf{x},\boldsymbol{\lambda},\mathbf{s}) = \frac{1}{2}\left(\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s})\|^2 + \sum_{i\notin\sigma} |f_i(\mathbf{x}) - \mathbf{s}_i|^2\right) + \frac{1}{2}\|\mathbf{s} - \mathbf{K}\boldsymbol{\lambda}\|^2. \tag{5.27}$$

where the set σ includes the inactive constraints

$$\sigma = \left\{ i = 1 \dots m \middle| \boldsymbol{\lambda}_i = 0, f_i(\mathbf{x}) - \mathbf{s}_i < 0 \right\}. \tag{5.28}$$

In the next proposition we show that the function defined in (5.27) is also nonincreasing along the dynamics (5.9)–(5.11).

Proposition 2. *Let $f_0 : \mathbb{R}^n \to \mathbb{R}$ and $f : \mathbb{R}^n \to \mathbb{R}^p$ satisfy assumptions 9 and 10. Then, the function $V(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})$ defined in* (5.27) *is non-increasing along the solutions of* (5.9)–(5.11).

Proof. Observe that the function $V(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})$ is not always differentiable due to the presence of the projection in $\dot{\boldsymbol{\lambda}}$. However, as long as there are no changes in the set σ the previous function is differentiable. We start by considering this case. Taking the derivative of $V(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})$ along the dynamics (5.9)–(5.11) yields

$$\begin{aligned} \dot{V} = & \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})^\top \left(\nabla^2_{\mathbf{xx}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) \dot{\mathbf{x}} + \nabla^2_{\mathbf{x}\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) \dot{\boldsymbol{\lambda}} \right) \\ & + \sum_{i \notin \sigma} \nabla_{\boldsymbol{\lambda}_i} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) \left(\nabla^2_{\boldsymbol{\lambda}_i \mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})^\top \dot{\mathbf{x}} + \nabla^2_{\boldsymbol{\lambda}_i \mathbf{s}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})^\top \dot{\mathbf{s}} \right) \\ & + (\mathbf{s} - \mathbf{K}\boldsymbol{\lambda})^\top (\dot{\mathbf{s}} - \mathbf{K}\dot{\boldsymbol{\lambda}}). \end{aligned} \tag{5.29}$$

Notice that in the previous expression we have used the fact that $\nabla^2_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) = 0$ and that $\nabla^2_{\mathbf{xs}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) = 0$ (cf., (5.3)). We will next show that

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})^\top \nabla^2_{\mathbf{x}\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) \dot{\boldsymbol{\lambda}} + \sum_{i \notin \sigma} \nabla_{\boldsymbol{\lambda}_i} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) \nabla^2_{\boldsymbol{\lambda}_i \mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})^\top \dot{\mathbf{x}} = 0. \tag{5.30}$$

Notice that the product $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})^\top \nabla^2_{\mathbf{x}\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})$ yields

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})^\top \nabla^2_{\mathbf{x}\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) \dot{\boldsymbol{\lambda}} = \sum_{i=1}^{m} \dot{\lambda}_i \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})^\top \nabla f_i(\mathbf{x}). \tag{5.31}$$

Replacing $\dot{\boldsymbol{\lambda}}$ in the previous expression for that in (5.10) yields

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})^\top \nabla^2_{\mathbf{x}\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) \dot{\boldsymbol{\lambda}} = \sum_{i \notin \sigma} (f_i(\mathbf{x}) - \mathbf{s}) \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})^\top \nabla f_i(\mathbf{x}). \tag{5.32}$$

Notice that the second term in (5.30) can be written as

$$\sum_{i \notin \sigma} \nabla_{\boldsymbol{\lambda}_i} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) \nabla^2_{\boldsymbol{\lambda}_i \mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})^\top \dot{\mathbf{x}} = \sum_{i \notin \sigma} (f_i(\mathbf{x}) - \mathbf{s}_i) \nabla f_i(\mathbf{x})^\top \dot{\mathbf{x}} \tag{5.33}$$

Replacing $\dot{\mathbf{x}}$ in (5.33) for the expression in (5.9) yields the negative of (5.32). Hence, (5.30)

holds. Likewise, replacing $\dot{\boldsymbol{\lambda}}$ and $\dot{\mathbf{s}}$ by their expressions in (5.10) and (5.11) follows that

$$\sum_{i\notin\sigma} \nabla_{\boldsymbol{\lambda}_i}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s})\nabla^2_{\boldsymbol{\lambda}_i\mathbf{s}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s})^\top\dot{\mathbf{s}} - (\mathbf{s}-\mathbf{K}\boldsymbol{\lambda})^\top\mathbf{K}\dot{\boldsymbol{\lambda}} = 0. \tag{5.34}$$

Taking into account the previous cancellations (5.29) reduces to

$$\dot{V} = -\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s})^\top\nabla^2_{\mathbf{xx}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s})\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\mathbf{s}) - (\mathbf{s}-\mathbf{K}\boldsymbol{\lambda})^\top\mathbf{K}\,(\mathbf{s}-\mathbf{K}\boldsymbol{\lambda}) \leq 0. \tag{5.35}$$

The latter shows that $V(\mathbf{x},\boldsymbol{\lambda},\mathbf{s})$ is non-increasing as long as there are no changes in the set of inactive constraints σ. We are left to analyze the cases where one constraint is either added or removed from the set σ. Observe that if a constraint is added to the set, the sum in (5.27) loses one term, and hence $V(\mathbf{x},\boldsymbol{\lambda},\mathbf{s})$ cannot increase. We will next show that if a constraint leaves the set σ at time t it must be the case that $f_i(\mathbf{x}(t)) - \mathbf{s}_i(t) = 0$ and thus $V(\mathbf{x}(t^+),\boldsymbol{\lambda}(t^+),\mathbf{s}(t^+)) = V(\mathbf{x}(t^-),\boldsymbol{\lambda}(t^-),\mathbf{s}(t^-))$, where the times t^- and t^+ correspond to the directional limits of the time before and after the discontinuity happens. By definition of the set σ (cf., (5.28)) a constraint can only leave the set if either λ_i goes from zero to positive or if $f_i(\mathbf{x}) - \mathbf{s}_i$ goes from negative to positive. Observe that as long as $\lambda_i = 0$ and $f_i(\mathbf{x}) - \mathbf{s}_i < 0$ from the dual dynamics (5.10) we have that $\dot{\boldsymbol{\lambda}}_i = 0$. Hence, no constraint can leave the set σ by λ_i becoming positive. Hence, it must be the case that $f_i(\mathbf{x}) - \mathbf{s}_i$ becomes positive. At the precise moment of the constraint leaving the set, we have that $f_i(\mathbf{x}) - \mathbf{s}_i = 0$ and therefore even if there is one more term in the summation $\sum_{i\notin\sigma}|f_i(\mathbf{x}(t^+) - \mathbf{s}_i(t^+)$ its value is zero. Thus $V(\mathbf{x}(t^+),\boldsymbol{\lambda}(t^+),\mathbf{s}(t^+)) = V(\mathbf{x}(t^-),\boldsymbol{\lambda}(t^-),\mathbf{s}(t^-))$. The latter completes the proof that $V(\mathbf{x},\boldsymbol{\lambda},\mathbf{s})$ is non-increasing along the dynamics (5.10)–(5.11). □

Based on the previous proposition, where we established that the function $V(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\mu})$ (5.27), is non-increasing along the modified saddle point dynamics, we claim convergence of the algorithm to the optimal solution for the relaxed problem (5.2) with slack variable satisfying $\boldsymbol{\lambda}^*(\mathbf{s}^*) = \mathbf{K}^{-1}\mathbf{s}^*$.

Proposition 3. *Let $f_0 : \mathbb{R}^n \to \mathbb{R}$ and $f : \mathbb{R}^n \to \mathbb{R}^m$ satisfy assumptions 9 and 10. Then, there exists $\mathbf{s}^*$ such that the dual optimal* (5.7) *satisfies*

$$\boldsymbol{\lambda}^*(\mathbf{s}^*) = \mathbf{K}^{-1}\mathbf{s}^*. \tag{5.36}$$

In addition, the dynamics (5.9)–(5.11) *converge to $(\mathbf{x}^*(\mathbf{s}^*),\boldsymbol{\lambda}^*(\mathbf{s}^*),\mathbf{s}^*)$, where $\mathbf{x}^*(\mathbf{s}^*)$ is the solution of* (5.2) *with slack $\mathbf{s}^*$.*

Proof. The proof of (5.36) follows from the result of Lemma 18. To show convergence observe that, from proposition 1 and 2 that there exists a compact positively invariant set

Ω for the dynamics (5.9)–(5.11) and a function $V(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})$ that decreases along trajectories in Ω. Then, the LaSalle invariance principle for hybrid systems [82] establishes, that every trajectory in Ω converges to M, the largest positively invariant set within Ω with trajectories satisfying $\dot{V}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})$ if σ is constant and such that $V(\mathbf{x}(t^-), \boldsymbol{\lambda}(t^-), \mathbf{s}(t^-)) = V(\mathbf{x}(t^+), \boldsymbol{\lambda}(t^+), \mathbf{s}(t^+))$ if σ changes. Formally, the previous conditions define the following set for which $V(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})$ is constant

$$E_1 = \left\{ \mathbf{x} \in \mathbb{R}^n, \boldsymbol{\lambda} \in \mathbb{R}^m_+, \mathbf{s} \in \mathbb{R}^m \middle| \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) = 0, \mathbf{K}\boldsymbol{\lambda} = \mathbf{s} \right\}, \tag{5.37}$$

for fixed σ and

$$E_2 = \left\{ \mathbf{x} \in \mathbb{R}^n, \boldsymbol{\lambda} \in \mathbb{R}^m_+, \mathbf{s} \in \mathbb{R}^m \middle| f_i(\mathbf{x}) - \mathbf{s}_i = 0 \right\}, \tag{5.38}$$

if i enters or leaves the set σ. Observe that $\boldsymbol{\lambda}, \mathbf{s} \in E_1$ means that $\mathbf{K}\boldsymbol{\lambda} = \mathbf{s}$. Hence, $\dot{\mathbf{s}} = 0$ (cf., (5.11)) at intervals in which σ is constant. The latter implies that $\mathbf{s}$ is in equilibrium, hence there exists $\mathbf{s}_\infty$ such that

$$\lim_{t \to \infty} \mathbf{s}(t) = \mathbf{s}_\infty. \tag{5.39}$$

Likewise we have that

$$\lim_{t \to \infty} \boldsymbol{\lambda}(t) = \boldsymbol{\lambda}_\infty = \mathbf{K}^{-1}\mathbf{s}_\infty. \tag{5.40}$$

To complete the proof, we need to show that the limit of the primal and dual variable indeed converge to the primal dual solution of (5.2) with slack $\mathbf{s}_\infty$. Notice that $(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) \in E_1$ implies as well that $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}) = 0$. From (5.9) it follows then that the limit of $\mathbf{x}(t)$ exists and moreover and it is to the minimizer of the Lagrangian for $(\boldsymbol{\lambda}_\infty, \mathbf{s}_\infty)$, i.e.,

$$\lim_{t \to \infty} \mathbf{x}(t) = \operatorname*{argmin}_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mathbf{K}^{-1}\mathbf{s}_\infty, \mathbf{s}_\infty) := \mathbf{x}_\infty. \tag{5.41}$$

From the KKT conditions [16, Section 5.5.3] it remains to be shown that the point $\mathbf{x}_\infty$ is feasible and that complementary slackness holds. Notice that there cannot be any $i = 1 \dots m$ for which $f_i(\mathbf{x}_\infty) - (\mathbf{s}_\infty)_i > 0$. If that were the case, $\boldsymbol{\lambda}_i$ would diverge (cf., (5.10)). Therefore, $\mathbf{x}_\infty$ is a feasible point. Notice that because $f_i(\mathbf{x}_\infty) - (\mathbf{s}_\infty)_i \leq 0$ for all $i = 1 \dots m$, if $(\boldsymbol{\lambda}_\infty)_i = 0$, then there cannot be a change in the set σ. If there is a change in the set it has to be the case where $(\boldsymbol{\lambda}_\infty)_i > 0$ and $(\mathbf{x}_\infty, \mathbf{s}_\infty) \in E_2$. Then complementary slackness holds and the proof of the proposition is completed. □

The previous result implies that the modified saddle point controller is such that it converges to the primal dual solution of the relaxed optimization problem (5.2) with slack $\mathbf{s}^*$ satisfying the relationship $\mathbf{K}^{-1}\mathbf{s}^* = \boldsymbol{\lambda}^*(\mathbf{s}^*)$. As previously discussed, the implication of the result is that it allows us to evaluate which one of the constraints is the hardest to satisfy by observing that the dual optimum is the derivative of the optimal value $p^*(\mathbf{s})$.

And in that sense, large multipliers, means that reducing the value of the corresponding slack entails a large increase in the cost of the problem. In the next section we generalize the previous result to settings in which the objective function and the constraints are not known but only a probabilistic model of them is available.

5.4 Stochastic Formulation

The main difference with the previous scenario is that instead of having access to the constraints and the objective function we assume some probabilistic model for them. Assume now, that we are given objective and constraint functions $f_0 : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ and $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^m$ then the problem of interest, is to to minimize the objective while satisfying the set of constraints in expectation

$$\begin{aligned} p^* := \min_{x \in \mathbb{R}^n} & \mathbb{E}_{\boldsymbol{\theta}} \left[f_0(\mathbf{x}, \boldsymbol{\theta}) \right] \\ \text{s.t.} \quad & \mathbb{E}_{\boldsymbol{\theta}} \left[f(\mathbf{x}, \boldsymbol{\theta}) \right] \preceq 0, \end{aligned} \tag{5.42}$$

where $\boldsymbol{\theta}$ is a random vector. As in the deterministic scenario (cf., Section 2.2) we are interested in determine whether the previous problem is feasible or not, i.e., if there exists $\mathbf{x}^\dagger \in \mathbb{R}^n$ such that $\mathbb{E}_{\boldsymbol{\theta}} \left[f(\mathbf{x}^\dagger, \boldsymbol{\theta}) \right] \preceq 0$ and to identify which of the constraints are harder to satisfy. Notice that, if the functions $f_0(\mathbf{x}, \boldsymbol{\theta})$ and $f(\mathbf{x}, \boldsymbol{\theta})$ are convex in the first argument, then problem (5.42) is not different than (5.1). And thus the methodology proposed here will be very similar to that employed in Section 2.2. We start by defining a slack variable $\mathbf{s}$ and the following relaxation of (5.42)

$$\begin{aligned} p^*(\boldsymbol{s}) := \min_{x \in \mathbb{R}^n} & \mathbb{E}_{\boldsymbol{\theta}} \left[f_0(\mathbf{x}, \boldsymbol{\theta}) \right] \\ \text{s.t.} \quad & \mathbb{E}_{\boldsymbol{\theta}} \left[f(\mathbf{x}, \boldsymbol{\theta}) \right] - \mathbf{s} \preceq 0, \end{aligned} \tag{5.43}$$

and its associated Lagrangian

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{s}) := \mathbb{E}_{\boldsymbol{\theta}} \left[f_0(\mathbf{x}, \boldsymbol{\theta}) \right] + \boldsymbol{\lambda}^\top \left(\mathbb{E}_{\boldsymbol{\theta}} \left[f(\mathbf{x}, \boldsymbol{\theta}) \right] - \boldsymbol{s} \right). \tag{5.44}$$

As in the previous section the objective is to find the saddle point of the Lagrangian for a given slack that results interesting in the sense that can allows to understand which constraints are more difficult to satisfy. The main difference, is that in the stochastic setting, the Lagrangian – and its gradients– cannot be estimated directly and at each iteration we can only sample from the underlying distribution of $\boldsymbol{\theta}$. Hence, what we are doing is a stochastic approximation [66, 107] of the algorithm (5.9)–(5.11).

In such settings is custumary to assume that the estimates available are unbiased and

with bounded moments. To be formal define a probability space $(\Omega, \mathcal{G}, P)$ and the following filtrarion as a sequence of increasing sigma algebras $\{\emptyset, \Omega\} = \mathcal{G}_0 \subset \ldots \subset \mathcal{G}_t \subset \ldots \mathcal{G}_\infty = \mathcal{G}$. We next formalize the assumptions regarding the estimate of the gradient of the Lagrangian along with convexity assumptions and constraint qualifications as in Section 2.2.

AS11. *The estimates of the gradient of the Lagrangian are unbiased, i.e., for all* $\mathbf{x}, \boldsymbol{\lambda}$ *and* $\mathbf{s}$ *it holds that*

$$\mathbb{E}\left[\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}, \boldsymbol{\theta}_t) \middle| \mathcal{G}_t\right] = \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}), \tag{5.45}$$

and

$$\mathbb{E}\left[\nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}, \boldsymbol{\theta}_t) \middle| \mathcal{G}_t\right] = \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}). \tag{5.46}$$

In addition the estimates have second moments bounded, i.e., there exist constants $\sigma_{\mathbf{x}}$ *and* $\sigma_{\boldsymbol{\lambda}}$ *such that*

$$\mathbb{E}\left[\left\|\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}, \boldsymbol{\theta}_t) - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})\right\|^2 \middle| \mathcal{G}_t\right] \leq \sigma_{\mathbf{x}}^2, \tag{5.47}$$

and

$$\mathbb{E}\left[\left\|\nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s}, \boldsymbol{\theta}_t) - \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})\right\|^2 \middle| \mathcal{G}_t\right] \leq \sigma_{\boldsymbol{\lambda}}^2, \tag{5.48}$$

AS12. *We assume* $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^m$ *is convex and* $f_0 : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ *is* μ*-strongly convex with respect to the first argument.*

AS13. *There exists* $\mathbf{x}^\dagger \in \mathbb{R}^n$ *and* $\mathbf{s}^\dagger \in \mathbb{R}^m_+$ *such that* $\mathbb{E}_{\boldsymbol{\theta}}\left[f(\mathbf{x}^\dagger, \boldsymbol{\theta})\right] - \mathbf{s}^\dagger \prec 0$.

We are now in conditions of presenting the stochastic approximation of the modified saddle point (5.9)–(5.10), where we update the primal variable $\mathbf{x}$ by descending along the direction of the negative graident of the Lagrangian with respect to $\mathbf{x}$

$$\mathbf{x}_{t+1} = \left[\mathbf{x}_t - \eta_t \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t, \boldsymbol{\theta}_t\right]_{\mathcal{X}} = \left[\mathbf{x}_t - \eta_t \left(\nabla_{\mathbf{x}} f_0(\mathbf{x}_t, \boldsymbol{\theta}_t) + \nabla_{\mathbf{x}} f(\mathbf{x}_t, \boldsymbol{\theta}_t) \boldsymbol{\lambda}_t\right)\right]_{\mathcal{X}}, \tag{5.49}$$

it ascends in $\boldsymbol{\lambda}$ along the direction of the gradient of the Lagrangian with respect to $\boldsymbol{\lambda}$

$$\boldsymbol{\lambda}_{t+1} = \left[\boldsymbol{\lambda}_t + \eta_t \hat{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t)\right]_{\Lambda} = \left[\boldsymbol{\lambda}_t + \eta_t \left(f(\mathbf{x}_t, \boldsymbol{\theta}_t) - \mathbf{s}_t\right)\right]_{\Lambda}, \tag{5.50}$$

and the slack variable is updated as in the deterministic case

$$\boldsymbol{s}_{t+1} = \left[\mathbf{s}_t + \eta_t \mathbf{K} \left(\mathbf{K} \boldsymbol{\lambda}_t - \mathbf{s}_t\right)\right]_{\Lambda}, \tag{5.51}$$

where the step-size of the algorithm η_t is a decreasing sequence satisfying

$$\sum_{t=0}^{\infty} \eta_t = \infty, \quad \text{and} \quad \sum_{t=0}^{\infty} \eta_t^2 < \infty, \tag{5.52}$$

and $[\cdot]_{\mathcal{X}}$ and $[\cdot]_{\Lambda}$ are projections over the sets $\mathcal{X}$ and Λ respectively. The former is a compact convex set contained in $\mathbb{R}^n$ and Λ is a compact convex subset of $\mathbb{R}^m_+$. The projection of the multipliers over the positive orthant is required to ensure that the multipliers remain non-negative, however the projection over bounded sets are technical requirements for the convergence proof. We formalize this assumptions for future reference.

AS14. *The sets $\mathcal{X}$ and Λ are convex and there exist positive constants $B_{\mathbf{x}}$ and $B_{\boldsymbol{\lambda}}$ such that for all $\mathbf{x} \in \mathcal{X}$ and for all $\boldsymbol{\lambda}, \mathbf{s} \in \Lambda$ it holds that $\|\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})\|^2 \leq B_{\mathbf{x}}$, $\|\nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{s})\|^2 \leq B_{\boldsymbol{\lambda}}$ and $\|\mathbf{K}\boldsymbol{\lambda} - \mathbf{s}\|^2 \leq B_{\mathbf{s}}$.*

In the next section we derive the analogous result to Proposition 3 where we establish convergence to the saddle point of the Lagrangian (5.44) for a slack variable satisfying

$$\boldsymbol{\lambda}^*(\mathbf{s}^*) = \mathbf{K}^{-1}\mathbf{s}^* \tag{5.53}$$

with probability one. In this setting the interpretation of such point is not different than in the deterministic counterpart. The larger the slack, the larger it is the cost in which we incur when trying to force the value of the slack down. Hence, giving us a notion of the relative hardness of satisfying the different constraints in the stochastic setting.

5.5 Stochastic Analysis

Notice that the problem (5.43) is also convex under Assumption 12 and it has zero duality gap when Assumption 13 hodls. Hence, by virtue of Lemma 18 it is possible to chose $\mathbf{s}^*$ with bounded norm such that the optimal dual variable for the optimization problem (5.43) with slack variable $\mathbf{s}^*$ satisfies

$$\boldsymbol{\lambda}^*(\mathbf{s}^*) = \mathbf{K}^{-1}\mathbf{s}^*. \tag{5.54}$$

Let $\mathbf{x}^*(\mathbf{s}^*)$ be the optimal primal variable for said problem and define the following sequence of random variables

$$V_t = \frac{1}{2}\left(\|\mathbf{x}_t - \mathbf{x}^*(\mathbf{s}^*)\|^2 + \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*(\mathbf{s}^*)\|^2 + \|\mathbf{s}_t - \mathbf{s}^*\|^2_{\mathbf{K}^{-2}}\right), \tag{5.55}$$

In order for the aglorithm (5.49)–(5.51) to converge we require that the points $\mathbf{s}^*, \boldsymbol{\lambda}^*(\mathbf{s}^*)$ to be in the set Λ and $\mathbf{x}^*(\mathbf{s}^*)$ to be in the set $\mathcal{X}$. We formalize this assumption for future reference.

AS15. *The gain Matrix $\mathbf{K}$ and the sets $\mathcal{X}$ and Λ are such that $\mathbf{s}^*, \boldsymbol{\lambda}^*(\mathbf{s}^*) \in \Lambda$ and $\mathbf{x}^*(\mathbf{s}^*) \in \mathcal{X}$.*

We start the analysis by defining the following sequence of random variables

$$S_t = V_t + \sum_{s=t}^{\infty} \eta_s^2 B, \tag{5.56}$$

where V_t is the sequence defined in (5.55) and B is defined as

$$B = B_{\mathbf{x}} + \sigma_{\mathbf{x}}^2 + B_{\boldsymbol{\lambda}} + \sigma_{\boldsymbol{\lambda}}^2 + B_{\mathbf{s}}, \tag{5.57}$$

with $\sigma_{\mathbf{x}}^2$ and $\sigma_{\boldsymbol{\lambda}}^2$ being the constants defined in Assumption 11 and $B_{\mathbf{x}}, B_{\boldsymbol{\lambda}}$ and $B_{\mathbf{s}}$ being the constants defined in Assumption 14. We next show that S_t is a non-negative supermartingale.

Lemma 13. *Let $f_0 : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^m$ satisfy assumptions 11–15. Then, the sequence S_t defined in* (5.56) *is a non-negative supermartingale.*

Proof. Notice that $\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t \in \mathcal{G}_t$, hence S_t is adapted to $\mathcal{G}_t$. S_t is also non-negative because is the sum of non-negative terms. To show that S_t is a supermartingale it remains to be shown that $\mathbb{E}\left[S_{t+1}|\mathcal{G}_t\right] \leq S_t$. To do so, use the update rule (5.49) to upper bound the norm

$$\left\|\mathbf{x}_{t+1} - \mathbf{x}^*(\mathbf{s}^*)\right\|^2 \leq \left\|\mathbf{x}_t - \eta \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t, \boldsymbol{\theta}_t) - \mathbf{x}^*(\mathbf{s}^*)\right\|^2, \tag{5.58}$$

where we have used the fact that because $\mathbf{x}^*(\mathbf{s}^*) \in \mathcal{X}$ (cf., Assumption 15) the projection cannot increase the norm. Expanding the squares yields

$$\begin{aligned}\left\|\mathbf{x}_{t+1} - \mathbf{x}^*(\mathbf{s}^*)\right\|^2 \leq \left\|\mathbf{x}_t - \mathbf{x}^*(\mathbf{s}^*)\right\|^2 + \eta_t^2 \left\|\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t, \boldsymbol{\theta}_t)\right\|^2 \\ -2\eta_t \left(\mathbf{x}_t - \mathbf{x}^*(\mathbf{s}^*)\right)^\top \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t, \boldsymbol{\theta}_t).\end{aligned} \tag{5.59}$$

Because $\mathbf{x}_t, \boldsymbol{\lambda}_t$ and $\mathbf{s}_t$ are measurable with respect to $\mathcal{G}_t$ and $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}, \boldsymbol{\theta}_t)$ is an unbiased estimate of the gradient, when conditioning the previous expression with respect to $\mathcal{G}_t$ it follows that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\mathbf{x}_{t+1} - \mathbf{x}^*(\mathbf{s}^*)\right\|^2 \middle| \mathcal{G}_t\right] \leq \left\|\mathbf{x}_t - \mathbf{x}^*(\mathbf{s}^*)\right\|^2 + \eta_t^2 \mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t, \boldsymbol{\theta}_t)\right\|^2\right] \\ -2\eta_t \left(\mathbf{x}_t - \mathbf{x}^*(\mathbf{s}^*)\right)^\top \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t).\end{aligned} \tag{5.60}$$

Using the convexity of the Lagrangian with respect to $\mathbf{x}$ and the fact that $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t) = \nabla f_0(\mathbf{x}_t) + \sum_{i=1}^m \lambda_i \nabla f_i(\mathbf{x}_t) = \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}^*)$, the inner product $\left(\mathbf{x}^*(\mathbf{s}^*) - \mathbf{x}_t\right)^\top \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t)$ can be upper bounded by

$$\left(\mathbf{x}^*(\mathbf{s}^*) - \mathbf{x}_t\right)^\top \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t) \leq \mathcal{L}(\mathbf{x}^*(\mathbf{s}^*), \boldsymbol{\lambda}_t, \mathbf{s}^*) - \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}^*). \tag{5.61}$$

Using assumptions 11 and 14 one can bound the second moment of $\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t, \boldsymbol{\theta}_t)$ by

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t, \boldsymbol{\theta}_t)\|^2\right] \leq B_{\mathbf{x}} + \sigma_{\mathbf{x}}^2. \tag{5.62}$$

Replacing the previous bound and the one in (5.61) in (5.60) yields

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*(\mathbf{s}^*)\|^2 \,\middle|\, \mathcal{G}_t\right] \leq \|\mathbf{x}_t - \mathbf{x}^*(\mathbf{s}^*)\|^2 + \eta_t^2(\sigma_{\mathbf{x}}^2 + B_{\mathbf{x}}) \\ +2\eta_t\left(\mathcal{L}(\mathbf{x}^*(\mathbf{s}^*), \boldsymbol{\lambda}_t, \mathbf{s}^*) - \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}^*)\right).\end{aligned} \tag{5.63}$$

Likewise, we can upper bound the expectation of the square of the norm of the difference $\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}^*(\mathbf{s}^*)$ as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\left[\|\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}^*(\mathbf{s}^*)\|^2 \,\middle|\, \mathcal{G}_t\right] \leq \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*(\mathbf{s}^*)\|^2 + \eta_t^2(\sigma_{\boldsymbol{\lambda}}^2 + B_{\boldsymbol{\lambda}}) \\ +2\eta_t\left(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*(\mathbf{s}^*)\right)^\top \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t).\end{aligned} \tag{5.64}$$

Observe that by adding and subtracting $(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*(\mathbf{s}^*))^\top \mathbf{s}^*$ to $(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*(\mathbf{s}^*))^\top \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t)$ yields

$$\begin{aligned}\left(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*(\mathbf{s}^*)\right)^\top \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t) = \left(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*(\mathbf{s}^*)\right)^\top (\mathbf{s}^* - \mathbf{s}_t) \\ + \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}^*) - \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}^*(\mathbf{s}^*), \mathbf{s}^*).\end{aligned} \tag{5.65}$$

Hence, (5.64) reduces to

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\left[\|\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}^*(\mathbf{s}^*)\|^2 \,\middle|\, \mathcal{G}_t\right] \leq \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*(\mathbf{s}^*)\|^2 + \eta_t^2(\sigma_{\boldsymbol{\lambda}}^2 + B_{\boldsymbol{\lambda}}) \\ +2\eta_t\left(\mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}^*) - \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}^*(\mathbf{s}^*), \mathbf{s}^*)\right) + 2\eta_t\left(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*(\mathbf{s}^*)\right)^\top (\mathbf{s}^* - \mathbf{s}_t).\end{aligned} \tag{5.66}$$

Using similar arguments, one can upper bound the expected value of the difference $\|\mathbf{s}_{t+1} - \mathbf{s}^*\|$ by

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\|\mathbf{s}_{t+1} - \mathbf{s}^*\|_{\mathbf{K}^{-2}}^2 \,\middle|\, \mathcal{G}_t\right] \leq \|\mathbf{s}_t - \mathbf{s}^*\|_{\mathbf{K}^{-2}}^2 + \eta_t^2 B_{\mathbf{s}} + 2\eta_t\left(\mathbf{s}_t - \mathbf{s}^*\right)^\top \mathbf{K}^{-1}\left(\mathbf{K}\boldsymbol{\lambda}_t - \mathbf{s}_t\right). \tag{5.67}$$

Combining the upper bounds for the expectation of the three squares (5.63), (5.66) and (5.67), and using the definition of the constant B in (5.57) allows us to upper bound $\mathbb{E}[V_{t+1}|\mathcal{G}_t]$ by

$$\begin{aligned}\mathbb{E}\left[V_{t+1}\middle|\mathcal{G}_t\right] \leq V_t + \eta_t^2 B + \eta_t\left(\mathbf{s}_t - \mathbf{s}^*\right)^\top \left(\boldsymbol{\lambda}^*(\mathbf{s}^*) - \mathbf{K}^{-1}\mathbf{s}_t\right) \\ + \eta_t\left(\mathcal{L}(\mathbf{x}^*(\mathbf{s}^*), \boldsymbol{\lambda}_t, \mathbf{s}^*) - \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}^*(\mathbf{s}^*), \mathbf{s}^*)\right),\end{aligned} \tag{5.68}$$

Because the step size is square summable (cf., (5.52)) we can add $\sum_{s=t+1}^{\infty} B\eta_s^2$ on both sides

of previous expression. This allows us to upper bound $\mathbb{E}[S_{t+1}|\mathcal{G}_t]$ by

$$\begin{aligned}\mathbb{E}\left[S_{t+1}\middle|\mathcal{G}_t\right] \leq S_t &+ \eta_t \left(\mathbf{s}_t - \mathbf{s}^*\right)^\top \left(\boldsymbol{\lambda}^*(\mathbf{s}^*) - \mathbf{K}^{-1}\mathbf{s}_t\right) \\ &+ \eta_t \left(\mathcal{L}(\mathbf{x}^*(\mathbf{s}^*), \boldsymbol{\lambda}_t, \mathbf{s}^*) - \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}^*(\mathbf{s}^*), \mathbf{s}^*)\right).\end{aligned} \tag{5.69}$$

To show that S_t is a supermartingale, we will show that $\left(\mathbf{s}_t - \mathbf{s}^*\right)^\top \left(\boldsymbol{\lambda}^*(\mathbf{s}^*) - \mathbf{K}^{-1}\mathbf{s}_t\right) \leq 0$ and that $\left(\mathcal{L}(\mathbf{x}^*(\mathbf{s}^*), \boldsymbol{\lambda}_t, \mathbf{s}^*) - \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}^*(\mathbf{s}^*), \mathbf{s}^*)\right) \leq 0$. To see why the first term is negative write it as

$$\left(\mathbf{s}_t - \mathbf{s}^*\right)^\top \left(\boldsymbol{\lambda}^*(\mathbf{s}^*) - \mathbf{K}^{-1}\mathbf{s}_t\right) = \left(\mathbf{s}_t - \mathbf{s}^*\right)^\top \mathbf{K}^{-1} \left(\mathbf{K}\boldsymbol{\lambda}^*(\mathbf{s}^*) - \mathbf{s}_t\right), \tag{5.70}$$

and observe that $\mathbf{K}\boldsymbol{\lambda}^*(\mathbf{s}^*) = \mathbf{s}^*$ (cf., (5.15)). Hence we can write

$$\left(\mathbf{s}_t - \mathbf{s}^*\right)^\top \left(\boldsymbol{\lambda}^*(\mathbf{s}^*) - \mathbf{K}^{-1}\mathbf{s}_t\right) = -\left(\mathbf{s}_t - \mathbf{s}^*\right)^\top \mathbf{K}^{-1} \left(\mathbf{s}_t - \mathbf{s}^*\right) \leq 0, \tag{5.71}$$

because $\mathbf{K} \succ 0$. The proof is completed by noticing that the definition of a saddle point (cf., (5.8)) implies that the difference of Lagrangians $\left(\mathcal{L}(\mathbf{x}^*(\mathbf{s}^*), \boldsymbol{\lambda}_t, \mathbf{s}^*) - \mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}^*(\mathbf{s}^*), \mathbf{s}^*)\right)$ is negative. □

The previous lemma establishes the the sequence of random variables S_t defined in (5.56) is a non-negative supermartingale. Because it is a sequence whose expected value is non-increasing it converges with probability one. Hence, to show that the algorithm (5.49)–(5.51) we need to show that the limit of S_t is zero with probability one. This is the subject of the following proposition.

Proposition 4. *Let assumptions 11–15 hold. Then, the sequence* $(\mathbf{x}_t, \boldsymbol{\lambda}_t, \mathbf{s}_t)$ *that arises from the update* (5.9),(5.10) *and* (5.11) *with step-size* η_t *satisfying* (5.52) *converges to* $(\mathbf{x}^*(\mathbf{s}^*), \boldsymbol{\lambda}^*(\mathbf{s}^*), \mathbf{s}^*)$ *with probability one.*

Proof. Using the fact that S_t is a non-negative supermartingale (cf., Lemma 14) we have that with probability one $\lim_{t\to\infty} S_t = S$, where S is a random variable satisfying $\mathbb{E}[S] \leq \mathbb{E}[S_0] < \infty$ (see e.g., [29, Theorem 5.2.8]). Likewise, observe that

$$S = \lim_{t\to\infty} S_t = \lim_{t\to\infty} V_t, \tag{5.72}$$

because the limit of the tail of the series $\sum_{s=t}^{\infty} \eta_s^2 B$ is zero. Let us define the following sequence for simplicity

$$\alpha_t = -\mathcal{L}(\mathbf{x}_t, \boldsymbol{\lambda}^*(\mathbf{s}^*), \mathbf{s}^*) - \mathcal{L}(\mathbf{x}^*(\mathbf{s}^*), \boldsymbol{\lambda}_t, \mathbf{s}^*) - \|\mathbf{s}_t - \mathbf{s}^*\|^2_{\mathbf{K}^{-1/2}}. \tag{5.73}$$

Notice that for all $t \geq 0$ we have that $\alpha_t \geq 0$ (cf., Proof of Lemma 14). We will show at the

end of this proof that α_t satisfies

$$\sum_{t=0}^{\infty} \eta_t \alpha_t < \infty \quad a.e. \tag{5.74}$$

This being the case, because the sequence η_t is non sumable it follows that

$$\liminf_{t\to\infty} \alpha_t = 0 \quad a.e. \tag{5.75}$$

Because of the saddle point property of the solution $(\mathbf{x}^*(\mathbf{s}^*), \boldsymbol{\lambda}^*(\mathbf{s}^*))$ (cf., (5.8)), (5.75) implies that there exists a subsequence $\{t_s\}$ such that

$$\lim_{s\to\infty} \|\mathbf{x}_{t_s} - \mathbf{x}^*(\mathbf{s}^*)\| = 0, \tag{5.76}$$

$$\lim_{s\to\infty} \|\boldsymbol{\lambda}_{t_s} - \boldsymbol{\lambda}^*(\mathbf{s}^*)\| = 0, \tag{5.77}$$

$$\lim_{s\to\infty} \|\mathbf{s}_{t_s} - \mathbf{s}^*\| = 0. \tag{5.78}$$

Notice that the three previous conditions imply that $\lim_{s\to\infty} V_{t_s} = 0$ with probability one. Since, the limit, $\lim_{t\to\infty} V_t$ exists it has to be the case that $\lim_{t\to\infty} V_t = 0$ almost everywhere. To complete the proof we need to show that (5.74) holds. To do so, observe that the sequence

$$\sum_{t=0}^{T} \eta_t \alpha_t \tag{5.79}$$

is monotonically increasing with T because both η_t and α_t are positive for all $t \geq 0$. Hence, the Monotone Convergence Theorem (see e.g., [29, Theorem 1.6.6]) allows us to write that

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \eta_t \alpha_t\right] = \sum_{t=0}^{\infty} \mathbb{E}\left[\eta_t \alpha_t\right]. \tag{5.80}$$

Use recursively the result from Lemma 14

$$\mathbb{E}[V_{t+1}\big|\mathcal{G}_t] \leq V_t - \eta_t \alpha_t, \tag{5.81}$$

and the towering property of the conditional expectation to write

$$\mathbb{E}[V_{t+1}] \leq \mathbb{E}[V_0] - \sum_{s=0}^{t} \mathbb{E}[\eta_s \alpha_s]. \tag{5.82}$$

Re arranging the terms of the previous expression and taking limit with t going to infinity yields

$$\lim_{t\to\infty}\sum_{s=0}^{t}\mathbb{E}[\eta_s\alpha_s] \leq \lim_{t\to\infty}\mathbb{E}[V_0]-\mathbb{E}[V_{t+1}] \leq \mathbb{E}[V_0], \tag{5.83}$$

where the last inequality follows from the fact that $V_t \geq 0$ for all t. Combining the previous upper bound with the result in (5.80), we can write

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\eta_t\alpha_t\right] = \sum_{t=0}^{\infty}\mathbb{E}\left[\eta_t\alpha_t\right] \leq \mathbb{E}[V_0] < \infty \tag{5.84}$$

Because the random variable $\sum_{t=0}^{\infty}\eta_t\alpha_t$ is nonnegative, to have bounded expectation it is required that the set where the sum diverges has measure zero. Which completes the proof of the proposition. □

5.6 Numerical Experiments

Here we consider a simple example with three constraints of the form $f_i(\mathbf{x}) = \|\mathbf{x}-\mathbf{x}_i\|^2$. Where $\mathbf{x}_1 = [-3,-1]$, $\mathbf{x}_2 = [-3,1]$ and $\mathbf{x}_3 = [3,0]$. From the definition of the constraints there is no point in space that can satisfy the three at the same time. However, the intuition is that the first two should be easier to satisfy because the minimum of $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are closer than togheter than that of $f_3(\mathbf{x})$. In that sense, if we give all the constraints the same importance, i.e., $\mathbf{K} = \mathbf{I}$ we would expect the slack corresponding to the third constraint to be larger. This is confirmed by the slack plot in Figure 5.1 where in yellow we observe the evolution of the third slack $\boldsymbol{\mu}_3$, whose final value is larger than that of the other two.

5.7 Conclusion

In this chapter we considered situations in which there is no information about the problem of interest being feasible or not. We proposed a modified saddle point algorithm in which we introduce a slack variable to solve the problem in cases where the constraints and the objective function can be measured exactly and a stochastic approximation of the previous algorithm in cases where the model of the functions is probabilistic. We showed in both cases convergence to the primal dual optimal solution for a specific slack. The slack obtained is proportional to the gradient of the optimal cost with respect to the slack. The latter provides a relative measure of the hardness in satisfying the constraints, because reducing a slack that is large translates into a large cost. The information obtained through the modified Arrow-Hurwicz algorithm can be used by a high level reasoning to decide modifications of the optimization problem.

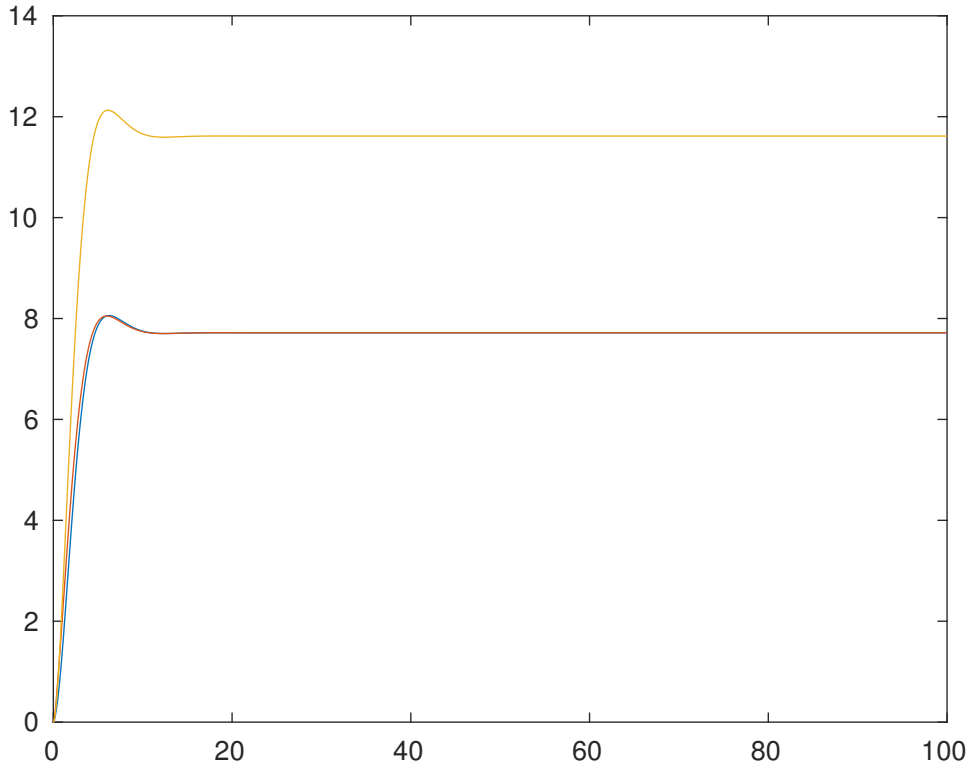


Figure 5.1: We observe the evolution of the slacks for the solutions of the dynamical system (5.9)–(5.11). In blue and red we observe the evolution of the slacks $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ corresponding to the constraints $f_i(\mathbf{x}) = \|\mathbf{x} - \mathbf{x}_i\|^2$, with $\mathbf{x}_1 = [-3, -1]$ and $\mathbf{x}_2 = [-3, 1]$. In yellow we observe the slack $\boldsymbol{\mu}_3$ for the constraint with center $\mathbf{x}_3 = [3, 0]$. Because the centers of the first two solutions are closer between them as compared to the third center. It is not surprising that the slack required to satisfy those constraints is smaller.

Chapter 6

Stochastic Policy Gradient Ascent in Reproducing Kernel Hilbert Spaces

In this chapter we consider the problem of policy optimization in the context of reinforcement learning, with the goal of maximizing an expected cumulative reward (ECR). In order to avoid discretization, we select the optimal policy to be a continuous function belonging to a reproducing Kernel Hilbert Space (RKHS). We design a policy gradient algorithm (PGA) in this context, deriving the gradients of the functional ECR and learning the unknown state transition probabilities on the way. In particular, we propose an unbiased stochastic approximation for the gradient that requires a finite number of steps. This unbiased estimator is the key enabler for a novel stochastic PGA, which provably converges to a critical point of the ECR. However, the RKHS approach increases the model order per iteration by adding extra kernels, which may render the numerical complexity prohibitive. To overcome this limitation, we prune the kernel dictionary using an orthogonal matching pursuit procedure, and prove that the modified method keeps the model order bounded for all iterations, while ensuring convergence to a neighborhood of the critical point.

6.1 Introduction

Markov decision Processes (MDPs) [44] provide a mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of a decision maker. This general framework has been used to study divers systems such as robotics [54], control [114], and finance [104]. More precisely, an MDP is a discrete time stochastic control process, where the state of the system at the next time is a random

variable, whose probability distribution depends on the current state and the action selected by decision maker. Because these transition probabilities do not depend on the history of the system they are also called memoryless systems. The actions selected by the agent determine instantaneous rewards that can be aggregated over a trajectory to determine cumulative rewards. The instantaneous rewards depend on both the state and the actions and thus, the reward along a trajectory depends on the policy under which the actions are selected based on the current state. In that sense, cumulative rewards are a measure of the quality of the decision making policy, and the objective of the agent is to find a policy that maximizes the expectation of the cumulative rewards, also known as the Q-function of the MDP [118].

In this chapter we consider reinforcement learning problems, in which the transition probabilities and the rewards are unknown and can only be accessed trough experiments that permit observation of realized transitions and rewards [118]. Solutions to these problems can be roughly divided among those that learn the Q-function to then chose for any given state the action that maximizes the function [132] and those that attempt to directly learn the optimal policy [27, 120]. Among the former, Q-learning is the most celebrated algorithm [132], its drawback, is that in general is difficult to maximize to determine the optimal policy. Algorithms that attempt to learn the optimal policy directly are based on computing (stochastic) gradients of the Q-function with respect to the policy and run gradient ascent [27, 120].

A major drawback of the previous algorithms for reinforcement learning is that they suffer from the curse of dimensionality, this is, the complexity of learning scales exponentially with the number of actions and states [37]. This is in particular the case of continuous spaces, where any reasonable discretization leads to a very large number of states and possible actions. Efforts to sidestep this issue assume that either the Q-function or the policy admits finite linear parametrization [119] or nonlinear basis expansion [13], is defined by a neural network [86] or that it belongs to a Reproducing Kernel Hilbert Space (RKHS) [61, 71, 126]. The latter provide the ability to approximate functions using nonparameteric functional representations. Although the structure of the space is determined by the choice of the kernel, the set of functions that can be represent is sufficiently rich to permit a good approximation of a large class of functions.

Here, we consider policy learning in RKHS as in [71] and we show, that it is possible to learn a policy that is a stationary point of the Q-function (Theorem 11). To do so, we construct an estimate of the gradient of the expected cumulative reward (Section 6.3) and we run stochastic gradient ascent. In the estimation of the gradient there are two main challenges that are addressed. The first one is related to the fact that the expression of the policy gradient depends on the Q-function itself and thus, it needs to be estimated. This

can be solved using a stochastic estimator of said function (Algorithm 1) that is unbiased (Proposition 5). The second difficulty when computing the gradient of the Q-function is that it depends on a state-action distribution that is not that of sample trajectories. Meaning that if one were to consider a trajectory of the system as a sample to compute the stochastic gradient, this estimate would be biased. This issue is typically reinforced by other policy gradient algorithms which consider a fixed horizon as estimate of the infinite sequence of state and action pairs. The biases introduced by the mentioned algorithms prevent to show convergence of stochastic gradient ascent to a stationary point of the Q-function. To overcome these issues, we propose to use as stopping time a random variable drawn from a geometric distribution. Such stopping time defines a horizon that is representative of the infinite time horizon problem and hence yields an unbiased estimate (Proposition 6). Whereas the setting considered in this chapter is the same as in [71], showing that the estimate of the policy gradient proposed is unbiased and the convergence of the algorithm are some of the contributions here presented.

Desipite the theoretical relevance of the previous algorithm, it has two issues of practical importance that we also address: (i) Reducing the variance of policy gradient stochastic approximations. (ii) Controlling the memory explosion of RKHS representations. To reduce the variance of stochastic policy gradient estimates we show that multiple samples from a Gaussian random policy can be related to numerical differentiation of the Q-function (Section 6.3.3). This idea has been used in the zero-th order optimization literature [39, 88]. This is, when the gradient of the function one is trying to minimize cannot be directly computed, one can estimate it by considering random samples in a neighborhood of the iterate and evaluating the objective function at those points. The problem of memory explosion has its origin in the fact that each sample used in the estimation of the stochastic gradient results in adding a kernel element. Hence, we require as many kernel elements as stochastic gradient iterations we perform. Since the convergence of stochastic gradient ascent is asymptotic we would need an infinite number of elements to represent the optimal policy. To control memory explosion [129] of RKHS representations we follow the ideas in [61] to propose the use of orthogonal matching pursuit to construct sparse kernel representations (Section 6.5). By doing so, we ensure that the model order of the representation remains bounded for all iterates at the cost of achieving convergence only to a neighborhood of a critical point of the Q-function (Theorem 12). The size of the neighborhood depends both on the learning rate – step size– selected and the error that one allows in the construction of sparse representations. Other than concluding remarks the chapter ends with numerical experiments where we consider the mountain car problem (Section 6.7).

6.2 Problem Formulation

Here, we are interested in the problem of finding a policy that maximizes the expected reward of an agent that chooses actions sequentially. Formally, let us denote the time by $t \in \{\{0\}, \mathbb{N}\}$ and let $\mathcal{S}$ be a compact set denoting the state space of the agent and $\mathcal{A} = \mathbb{R}^p$ be its action space. The transition dynamics are governed by a conditional probability $P^{a_t}_{s_t \to s_{t+1}}(s) := p(s_{t+1} = s | (s_t, a_t) \in \mathcal{S} \times \mathcal{A})$ satisfying the Markov property, i.e., $p(s_{t+1} = s \big| (s_u, a_u) \in \mathcal{S} \times \mathcal{A}, \forall u \leq t) = p(s_{t+1} = s | (s_t, a_t) \in \mathcal{S} \times \mathcal{A})$. The policy of the agent is a map $h : \mathcal{S} \to \mathcal{A}$ and we assume it to be a vector-valued function in a vector-valued RKHS $\mathcal{H}$. We formally define this notion next, with comments ensuing.

Definition 7. *A vector valued RKHS $\mathcal{H}$ is a Hilbert space of functions $h : \mathcal{S} \to \mathbb{R}^p$ such that for all $\mathbf{c} \in \mathbb{R}^p$ and $\mathbf{x} \in \mathcal{S}$, $(\kappa_x \mathbf{c})(\mathbf{y}) = \kappa(\mathbf{x}, \mathbf{y})\mathbf{c} \in \mathcal{H}$ for all $\mathbf{y} \in \mathcal{S}$, where $\kappa_x(\mathbf{y})$ is a symmetric function that is a positive definite matrix for any $\mathbf{x}, \mathbf{y} \in \mathcal{S}$ and it has the reproducing property*

$$< h, \kappa_x \mathbf{c} >_{\mathcal{H}} = h(\mathbf{x})^\top \mathbf{c}. \tag{6.1}$$

Without loss of generality we will assume that the Hilbert norm of $\kappa(\mathbf{x}, \cdot)$ is equal to one.

If $\kappa(\mathbf{x}, \mathbf{y})$ is a diagonal matrix-valued function with diagonal elements $\kappa(\mathbf{x}, \mathbf{y})_{ii}$, and $\mathbf{c}$ is the i-th canonical vector in $\mathbb{R}^p$, then (6.1) reduces to the standard one-dimensional reproducing property per coordinate $h_i(\mathbf{x}) =< h_i, \kappa(\mathbf{x}, \cdot)_{ii} >$.

Instead of choosing the action deterministically as $a = h(s)$, we randomly draw it from a multivariable Gaussian distribution with mean $h(s)$. A random policy helps the exploration of the state space and it is a good approximation of the deterministic policy as we show in Proposition 7. The conditional probability of the action is defined as $\pi_h(a|s) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$, with

$$\pi_h(a|s) = \frac{1}{\det(2\pi\Sigma)} e^{-(a-h(s))^\top \Sigma^{-1} (a-h(s))}. \tag{6.2}$$

The latter means that given a policy $h \in \mathcal{H}$ and the current state $s \in \mathcal{S}$, the agent selects an action $a \in \mathcal{A}$ from a multivariate normal distribution $\mathcal{N}(h(s), \Sigma)$. The actions selected by the agent result in a reward defined by a function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. We assume these rewards to be uniformly bounded as we formally state next.

AS16. *There exists $B_r > 0$ such that $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$, the reward function $r(s, a)$ satisfies $|r(s, a)| \leq B_r$.*

The objective is then to find a policy $h^* \in \mathcal{H}$ such that it maximizes the expected discounted reward

$$h^* := \underset{h \in \mathcal{H}}{\operatorname{argmax}}\, U(h) = \underset{h \in \mathcal{H}}{\operatorname{argmax}}\, \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \Big| h \right], \tag{6.3}$$

where the expectation is taken with respect to all states $s_0, s_1, \ldots$ and all actions $a_0, a_1, \ldots$, and $\gamma \in (0,1)$ is a discount factor that gives relative weights to the reward at different times. Values of γ close to one imply that rewards in the present are as important as future rewards, whereas smaller values of γ give origin to myopic policies that prioritize maximizing immediate rewards. It is also noticeable that $U(h)$ is indeed a function of the policy h, since policies affect the joint probabilities of the trajectories $\{s_t, a_t\}_{t=0}^{\infty}$.

Conceivably, problem (6.3) could be solved iteratively by running a gradient ascent iteration on the space of functions. In parametric problems where variables lie in a finite space, gradient ascent converges to a critical point of $U(h)$ – if $U(h)$ is upper bounded – under constant and diminishing step size [12, pp 43-45]. The same will be proved here in the case of maximizing a functional where the decision variable is a function in $\mathcal{H}$. When the functional is a convex function these results have been established in [59, 60].

The importance of this theoretical result notwithstanding, is limited by the computation of the gradient of $U(h)$ with respect to h being intractable. To see why this is the case, define the discounted long-run probability distribution $\rho(s,a)$

$$\rho(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t p(s_t = s, a_t = a) \tag{6.4}$$

where $p(s_t = s, a_t = a)$ defines the probability of reaching state s and action a at time t, and is given by

$$p(s_t, a_t) = \int \pi_h(a_t|s_t)\prod_{u=0}^{t-1} p(s_{u+1}|s_u, a_u)\pi_h(a_u|s_u)p(s_0)d\mathbf{s}d\mathbf{a} \tag{6.5}$$

and where $d\mathbf{s} = ds_0 \ldots ds_{t-1}$ and $d\mathbf{a} = da_0 \ldots da_{t-1}$ imply integration over the previous states and actions.

Let $Q(s,a;h)$ be the expected discounted reward for a policy h that at state s selects action a, formally defined as

$$Q(s,a;h) := \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\Big| h, s_0 = s, a_0 = a\right]. \tag{6.6}$$

With these functions defined, the gradient of the discounted rewards with respect to h yields [71, 120]

$$\begin{aligned}&\nabla_h U(h,\cdot) = \\ &\frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\rho(s,a)}\left[Q(s,a;h)\kappa(s,\cdot)\Sigma^{-1}\left(a - h(s)\right)\Big| h\right],\end{aligned} \tag{6.7}$$

where the dot in $(h, \cdot)$ substitutes the second variable of the kernel, belonging to $\mathcal{S}$, which is ommited to simplify notation. Observe that the expectation with respect to the distribution $\rho(s, a)$ is an integral of an infinite sum over a continuous space. In addition, the system transition density $p(s_{t+1}|s_t, a_t)$ is not known. Therefore, computing (6.7) in closed form is intractable and a large number of samples might be needed to obtain an accurate Monte Carlo approximation even if $(p_{t+1}|s_t, a_t)$ was known. An alternative to overcome this drawback is the use of stochastic approximation methods (see [52,102,107,136]), where the main idea is to compute an unbiased estimate of the policy gradient by evaluating the expression inside the expectation for one sample of a pair $(s, a) \sim \rho(s, a)$, thus reducing the cost of each iteration. Observe however, that in this particular case the evaluation of the stochastic gradient requires the Q-function defined in (6.6), which presents the same challenges that computing the gradient of the expected discounted reward, i.e., an intractable closed-form expression and a computationally prohibitive approximation. In Section 6.3.1 we present an efficient subroutine to find an unbiased estimate of the Q function which is used in Section 6.3.2 to define the stochastic gradient of the expected discounted reward. If one were to work with a deterministic policy, rather than needing an estimate of the Q-function, one needs an estimate of its derivative as we explain in Section 6.3.3. In Section 6.4, we show that by updating the policy with the stochastic estimate of $\nabla_h U(h, \cdot)$, convergence to a stationary point of $U(h)$ is achieved with probability one.

6.3 Stochastic Policy Gradient

In order to compute a stochastic approximation of $\nabla_h U(h)$ we need to sample from $\rho(s, a)$ given in (6.4). The intuition behind $\rho(s, a)$ is that it weights the probability of the system being at a specific state-action pair (s, a) at time t by a factor of $(1 - \gamma)\gamma^t$. Notice that this factor is equal to the probability of a geometric distribution of parameter γ to take the value t. Thus, for the k-th policy update, one can interpret the distribution $\rho(s, a)$ as the probability of running the system for T steps, with T randomly drawn from a geometric distribution of parameter γ. This supports steps 2-7 in Algorithm 2 which describes how to obtain a sample $(s_k, a_k) \sim \rho(s, a)$. Latter, in Proposition 6 it is claimed that an unbiased estimate of $\nabla_h U(h)$ is obtained by substituting the sample (s_k, a_k) in the stochastic gradient

$$\hat{\nabla}_h U(h, \cdot) = \frac{1}{1-\gamma} \hat{Q}(s_k, a_k; h) \kappa(s_k, \cdot) \Sigma^{-1}(a_k - h(s_k)), \tag{6.8}$$

with $\hat{Q}(s_k, a_k; h)$ being an unbiased estimate of $Q(s_k, a_k; h)$. The previous expression reveals a second challenge in computing of the stochastic gradient, namely the need of computing the function Q – or an estimate– at the state-action pair (s_k, a_k). We deal with this in Section 6.3.1, providing an unbiased estimate of $Q(s_k, a_k; h)$ that yields an unbiased

estimate of $\nabla_h U(h,\cdot)$ when substituted in (6.8).

Thus, we construct an unbiased estimate $\hat{\nabla}_h U(h,\cdot)$ in a finite number of steps. Using this estimate we propose to update the policy iteratively following the rule

$$h_{k+1} = h_k + \eta_k \hat{\nabla}_h U(h_k,\cdot), \tag{6.9}$$

where $\eta_k > 0$ is the step size of the algorithm. Under proper conditions stochastic gradient ascent methods can be shown to converge with probability one to the local maxima [101]. This approach has been widely used to solve parametric optimization problems where the decision variables are vectors in R^n. In this chapter we extend these results to non-parametric problems in RKHSs. First, we describe the algorithm to obtain the unbiased estimate $\hat{Q}(s_k,a_k;h)$ in a finite number of steps, which is instrumental for our overall non-parametric stochastic approximation strategy.

6.3.1 Unbiased Estimate of Q

A theoretically conceivable but unrealizable form of estimating the value of $Q(s,a;h)$ is to run a trajectory for infinite steps stating from $(s_0,a_0) = (s,a)$ and then compute the following infinite sum $\hat{q}_h = \sum_{t=0}^{\infty} \gamma^t r(s_t,a_t)$. Despite being an unbiased estimate, a major drawback of this approach is the need to consider an infinite number of steps. In contrast, we present the subroutine Algorithm 1 that allows to compute an unbiased estimate of $Q(s,a;h)$ by considering a representative future reward obtained after a finite number of steps. As with $U(h)$, a parameter γ closer to zero assigns similar weights to present and future rewards, and γ close to zero prioritizes the present. In that sense, when γ is very small, we do not need to let the system evolve for long time to get a representative reward. Likewise, for γ close to one we need to look far away into the future. Again, the geometric distribution allows us to represent this idea. Specifically, let T_Q be a geometric random variable with parameter γ, i.e., $P(T_Q = t) = (1-\gamma)\gamma^t$, which is finite with probability one. Then define the estimate of $Q(s,a;h)$ as the sum of rewards collected from step $t = 0$ until $t = T_Q$

$$\hat{Q}(s,a;h) := (1-\gamma) \sum_{t=0}^{T_Q} r(s_t,a_t) \tag{6.10}$$

Algorithm 1 summarizes how to obtain $\hat{Q}(s,a;h)$ as in (6.10), and Proposition 1 states that it is unbiased.

Proposition 5. *The output $\hat{Q}(s,a;h)$ of Algorithm 1 is an unbiased estimate of $Q(s,a;h)$.*

Proof. To show that the estimate is unbiased we start by computing the expectation of the

Algorithm 1 estimateQ

Input: s, a, h

Initialize: $\hat{Q} = 0$, $s_0 = s$, $a_0 = a$
Draw an integer T_Q form a geometric distribution with parameter γ, $P(T_Q = t) = (1-\gamma)\gamma^t$
for $t = 0, 1, \dots T_Q - 1$ **do**
    Collect reward and add to estimate $\hat{Q} = \hat{Q} + r(s_t, a_t)$
    Let system advance $s_{t+1} \sim P^{a_t}_{s_t \to s_{t+1}}$
    Select action $a_{t+1} \sim \pi_h(a|_{s_{t+1}})$
end for
Collect last reward $\hat{Q} = \hat{Q} + r(s_{m'}, a_{m'})$
Scale $\hat{Q} = (1-\gamma)\hat{Q}$
return $\hat{Q}$, s_{T_Q}

estimate conditioning on h and the initial state–action pair

$$\mathbb{E}\left[\hat{Q}(s,a;h)\middle| h, s_0 = s, a_0 = a\right] = \mathbb{E}\left[(1-\gamma)\sum_{t=0}^{\infty} \mathbb{1}(T_Q \geq t) r(s_t, a_t)\middle| h, s_0 = s, a_0 = a\right], \tag{6.11}$$

where we substituted ∞ for the T_Q as the last index of the sum, but added null summands for $t > T_Q$ by using the indicator function $\mathbb{1}$.

With the estimate written as in (6.11) we argue that $\hat{Q}(s,a;h)$ can be obtained equivalently by letting the system evolve towards infinity, and then keeping in the sum only those rewards for $t \leq T_Q$. Notice that according to Algorithm 1 T_Q is drawn independently of the system evolution. Furthermore, it will be argued below that the sum and expectation can be exchanged. With all this in mind we rewrite (6.11) as in

$$\begin{aligned}\mathbb{E}\left[\hat{Q}(s,a;h)\middle| h, s_0 = s, a_0 = a\right] &= (1-\gamma)\sum_{t=0}^{\infty} \mathbb{E}\left[\mathbb{1}(T_Q \geq t)\right] \mathbb{E}\left[r(s_t, a_t)\middle| h, s_0 = s, a_0 = a\right] \\ &= (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{E}\left[r(s_t, a_t)\middle| h, s_0 = s, a_0 = a\right] = Q(s,a;h)\end{aligned} \tag{6.12}$$

where we used that $T_Q \sim$Geom(γ) so that $\mathbb{E}\left[\mathbb{1}(T_Q \geq t)\right] = \gamma^t$.

It remains to proof that it is possible to exchange the sum and the expectation in the previous expression. To do so, using Assumption 16 and the triangle inequality observe that for all $N > 0$ we have that

$$\left|\sum_{t=0}^{N} \mathbb{1}(T_Q \geq t) r(s_t, a_t)\right| \leq \sum_{t=0}^{N} \mathbb{1}(T_Q \geq t) B_r. \tag{6.13}$$

Which by virtue of the monotonicity and the linearity of the expectation implies that

$$\mathbb{E}\left[\left|\sum_{t=0}^{N}\mathbb{1}(T_Q \geq t)r(s_t,a_t)\right|\right] \leq B_r\mathbb{E}\left[\sum_{t=0}^{N}\mathbb{1}(T_Q \geq t)\right]. \tag{6.14}$$

Observe that the random variable on the right is a monotonic increasing random variable and thus, by virtue of the monotone convergence theorem we have that

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\mathbb{1}(T_Q \geq t)\right] = \sum_{t=0}^{\infty}\mathbb{E}\left[\mathbb{1}(T_Q \geq t)\right] = \sum_{t=0}^{\infty}P(T_Q \geq t) = \sum_{t=0}^{\infty}\gamma^t = \frac{1}{1-\gamma}. \tag{6.15}$$

Notice that the sequence $\left|\sum_{t=0}^{N}\mathbb{1}(T_Q \geq t)r(s_t,a_t)\right|$ is dominated by $\sum_{t=0}^{\infty}\mathbb{1}\left(T_Q \geq t\right)B_r$ for all $N \geq 0$ and that the latter has bounded expectation. Then, the Dominated Convergence Theorem applies (see e.g., [29, Theorem 1.6.7]), and guarantees that indeed the expectation and sum can be exchanged in (6.11), concluding the proof. □

6.3.2 Unbiased Estimate of the Stochastic Gradient

In this section we present a subroutine that uses the estimate $\hat{Q}(s,a;h)$ produced by Algorithm 1 to obtain an unbiased estimate of $\nabla_h U(h)$. As discussed before, a sample from $\rho(s,a)$ can be obtained by sampling a time T from a geometric distribution of parameter γ and running the system T times. Although the resulting estimate in (6.8) can be shown to be unbiased, which would be enough for the purpose of stochastic approximation, we chose to introduce symmetry with respect to $h(s)$ as it is justified in Section 6.3.3. Instead of computing the approximation only at the state-action pair (s_T, a_T) we average said value with $\hat{Q}(s_T, \bar{a}_T)$, where $\bar{a}_T = h(s_T) - (a_T - h(s_T)$ is the action that is symmetric to a_T with respect to $h(s_T)$ (steps 8–11 in Algorithm 2). Hence, the proposed estimate is

$$\hat{\nabla}_h U(h,\cdot) = \frac{1}{2(1-\gamma)}\left(\hat{Q}(s_T,a_T;h) - \hat{Q}(s_T,\bar{a}_T;h)\right)\kappa(s_T,\cdot)\Sigma^{-1}(a_T - h(s_T)). \tag{6.16}$$

The subroutine presented in Algorithm 2 summarizes the algorithm to compute our stochastic approximation in (6.16). We claim that it is unbiased in the following proposition.

Proposition 6. *The output $\hat{\nabla}_h U(h,\cdot)$ of Algorithm 2 is an unbiased estimate of $\nabla_h U(h,\cdot)$ in* (6.7).

Proof. To show that the estimate is unbiased we compute the expectation of $\hat{\nabla}_h U(h,\cdot)$ conditioned to h, which in turn can be written as

$$\mathbb{E}\left[\hat{\nabla}_h U(h,\cdot)\middle|h\right] = \mathbb{E}\left[\left[\hat{\nabla}_h U(h,\cdot)\middle|s_T,a_T\right]\middle|h\right] \tag{6.17}$$

Algorithm 2 StochasticGradient

Input: h, s

Initialize: $s_0 = s$

Draw an integer T form a geometric distribution with parameter γ, $P(T = t) = (1-\gamma)\gamma^t$

Select action $a_0 \sim \pi_h(a|_s)$

for $t = 0, 1, \dots T-1$ **do**

    Advance system $s_{t+1} \sim P^{a_t}_{s_t \to s_{t+1}}$

    Select action $a_{t+1} \sim \pi_h(a|_{s_{t+1}})$

end for

Get estimate of $Q(s_T, a_T; h)$ as in Algorithm 1:

$$\hat{Q}(s_T, a_T; h) = \text{estimateQ}(s_T, a_T; h)$$

Given a_T, find symmetric $\bar{a}_T = h(s_T) - (a_T - h(s_T))$

Get estimate of $Q(s_T, \bar{a}_T; h)$ as in Algorithm 1:

$$\hat{Q}(s_T, \bar{a}_T; h) = \text{estimateQ}(s_T, \bar{a}_T; h)$$

Compute the stochastic gradient $\hat{\nabla}_h U(h, \cdot)$ as in (6.16) **return** $\hat{\nabla}_h U(h, \cdot)$

Using the linearity of the expectation and the fact that $\kappa(s_T, \cdot)\Sigma^{-1}(a_T - h(s_T))$ is measurable with respect of the sigma algebra generated by $s_0 \dots s_T$ and $a_0 \dots a_T$ we have that

$$\mathbb{E}\left[\hat{\nabla}_h U(h, \cdot)\Big|h\right] = \mathbb{E}\left[\mathbb{E}\left[\hat{Q}(s_T, a_T; h) - \hat{Q}(s_T, \bar{a}_T; h)\Big|s_T, a_T\right] \frac{\kappa(s_T, \cdot)}{2(1-\gamma)}\Sigma^{-1}(a_T - h(s_T))\Big|h\right]. \tag{6.18}$$

Using the result of Proposition 5 the previous expression reduces to

$$\mathbb{E}\left[\hat{\nabla}_h U(h, \cdot)\Big|h\right] = \mathbb{E}\left[(Q(s_T, a_T; h) - Q(s_T, \bar{a}_T; h)) \frac{\kappa(s_T, \cdot)}{2(1-\gamma)}\Sigma^{-1}(a_T - h(s_T))\Big|h\right]. \tag{6.19}$$

Since a_T is normally distributed with mean $h(s_T)$ we have that $a_T - h(s_T)$ and $h(s_T) - a_T$ are both normally distributed with zero mean. Moreover, $\bar{a}_T$ has the same distribution as a_T. Hence the two expectations on the right hand side of the previous equality are the same. Adding them yields

$$\mathbb{E}\left[\hat{\nabla}_h U(h, \cdot)\Big|h\right] = \frac{1}{1-\gamma}\mathbb{E}\left[Q(s_T, a_T; h)\kappa(s_T, \cdot)\Sigma^{-1}(a_T - h(s_T))\Big|h\right]. \tag{6.20}$$

The previous expression is equivalent to

$$\mathbb{E}\left[\hat{\nabla}_h U(h, \cdot)\Big|h\right] = \frac{1}{1-\gamma}\mathbb{E}\left[\sum_{t=0}^{\infty} \mathbb{1}(T = t)Q(s_t, a_t; h)\kappa(s_t, \cdot)\Sigma^{-1}(a_t - h(s_t))|h\right]. \tag{6.21}$$

Next, we argue that it is possible to exchange the infinity sum and the expectation in the previous expression. Observe that only one of terms inside the sum can be different than zero. Denote by t^* the index corresponding to that term and upper bound the norm of $\hat{\nabla}_h U(h)$ by

$$(1-\gamma)\left\|\hat{\nabla}_h U(h)\right\| \leq |Q(s_{t^*}, a_{t^*}; h)| \, \|\kappa(s_{t^*}, \cdot)\| \left\|\Sigma^{-1}(a_{t^*} - h(s_{t^*}))\right\|. \tag{6.22}$$

Using that $\|\kappa(s_t, \cdot)\| = 1$ (cf., Definition 7) and $|Q(s, a; h)| \leq B_r/(1-\gamma)$ (cf., Lemma 19), we can upper bound the previous expression by

$$\begin{aligned}\left\|\hat{\nabla}_h U(h)\right\| &\leq \frac{B_r}{(1-\gamma)^2} \left\|\Sigma^{-1}(a_{t^*} - h(s_{t^*}))\right\| \\ &\leq \frac{B_r}{(1-\gamma)^2 \, \lambda_{\min}(\Sigma^{-1/2})} \left\|\Sigma^{-1/2}(a_{t^*} - h(s_{t^*}))\right\|,\end{aligned} \tag{6.23}$$

Notice that $\Sigma^{-1/2}(a_t - h(s_t))$ are identically distributed mutlivariate normal distributions, and thus the expectation of its norm is bounded. The Dominated Convergence Theorem can be hence used to exchange the sum and the expectation in (6.21). In addition, the draw of the random variable T is independent of the evolution of the system until infinity. Hence (6.21) yields

$$\begin{aligned}\mathbb{E}\left[\hat{\nabla}_h U(h, \cdot)\middle| h\right] &= \sum_{t=0}^{\infty} \frac{P(t=T)}{1-\gamma} \mathbb{E}\left[Q(s_t, a_t; h)\kappa(s_t, \cdot)\Sigma^{-1}(a_t - h(s_t))|h\right] \\ &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}\left[Q(s_t, a_t; h)\kappa(s_t, \cdot)\Sigma^{-1}(a_t - h(s_t))|h\right] = \nabla_h U(h, \cdot).\end{aligned} \tag{6.24}$$

where the last equality coincides with that in (6.7). To be able to write the last equality we need to justify that it is possible to exchange the sum with the expectation. We do so next in order to complete proof that the stochastic gradient estimated by Algorithm 2 is unbiased. Let us define the following sequence of random variables

$$S_k = \sum_{t=0}^{k} \gamma^t Q(s_t, a_t; h)\kappa(s_t, \cdot)\Sigma^{-1}(a_t - h(s_t)). \tag{6.25}$$

Use the triangle inequality along with the bounds on $Q(s_t, a_t; h)$ and $\kappa(s_t, \cdot)$ from (6.23) to bound the norm of S_k by

$$\|S_k\| \leq \frac{B_r}{1-\gamma} \sum_{t=0}^{k} \gamma^t \left\|\Sigma^{-1}(a_t - h(s_t))\right\|. \tag{6.26}$$

Observe that the sum in the right is an increasing random variable because all terms in the sumands are positive. Hence, by virtue of the Monotone Convergence Theorem (see e.g., [29, Theorem 1.6.6]) we have that

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t \left\|\Sigma^{-1}(a_t - h(s_t))\right\|\right] = \sum_{t=0}^{\infty}\gamma^t\mathbb{E}\left[\left\|\Sigma^{-1}(a_t - h(s_t))\right\|\right]. \tag{6.27}$$

Because $\Sigma^{-1/2}((a_t - h(s_t))$ is normally dsitributed, its norm has bounded expectation. Use in addition the fact that the geometric series converges to ensure that the right hand side of the previous expression is bounded. S_k is therefore dominated by a random variable with finite expectation. Thus, the Dominated Convergence Theorem allows us to write that

$$\lim_{k\to\infty}\mathbb{E}[S_k] = \mathbb{E}[\lim_{k\to\infty} S_k]. \tag{6.28}$$

The latter corresponds to exchanging the sum and the expectation in (6.24). □

Now we are in conditions of presenting the complete algorithm for policy gradient in RKHSs. Each iteration consists of the estimation of $\hat{\nabla}_h U(h_k, \cdot)$ as described in Algorithm 2 – which uses Algorithm 1 as a subroutine to get unbiased estimates of $Q(s, a; h)$ – and of the updated

$$h_{k+1} = h_k + \eta_k \hat{\nabla}_h U(h_k, \cdot), \tag{6.29}$$

where η_k is non-summable and square summable, i.e.

$$\sum_{k=0}^{\infty}\eta_k = \infty \quad \text{and} \quad \sum_{k=0}^{\infty}\eta_k^2 < \infty. \tag{6.30}$$

Algorithm 3 Stochastic Policy Gradient Ascent

Input: step size η_0

1: *Initialize*: $h_0 = 0$
2: **for** $k = 0 \ldots$ **do**
3: Draw an initial state s_0 for Algorithm 2
4: Compute the stochastic gradient:

$$\hat{\nabla}_h U(h_k, \cdot) = \text{StochasticGradient}(h_k, s_0)$$

5: Gradient ascent step $h_{k+1} = h_k + \eta_k \hat{\nabla}_h U(h_k, \cdot)$
6: **end for**

Theorem 11. *Let* $\{h_k, k \geq 0\}$ *be the sequence of functions given by* (6.29)*, where* η_k *is as step size satisfying* (6.30) *and* $\hat{\nabla}_h U(h_k, \cdot)$ *is an unbiased estimator of the gradient of the*

functional. With probability one we have that $\lim_{k\to\infty} h_k = H^*$*, where* H^* *is a random variable taking values in the set of critical points of the functional* $U(h)$ *defined in* (6.3)*.*

Proof. The proof of this result is the matter of Section 6.4. □

The previous result establishes that h_k converges with probability one to a critical point of the functional $U(h)$. A major drawback of Algorithm 3 is that at each iteration the stochastic gradient ascent iteration will add a new element to the kernel dictionary. Indeed, for each iteration $\hat{\nabla}_h U(h_k, \cdot)$ introduces an extra kernel centered at a new s_T (cf., (6.16)). Hence for any $k > 0$ in order to represent h_k we require k dictionary elements. This translates into memory explosion and thus Algorithm 3, while theoretically interesting, is not practical. To overcome this limitation, we introduce in the next section a projection on a smaller Hilbert space so that we can control the model order. Before that, we introduce a discussion regarding the use of random policies. .

6.3.3 Gaussian policy as an approximation

Our reason to use a randomized Gaussian policy is two-fold: it yields a good approximation of the gradient of the q-function that would result from a deterministic policy as we show in Proposition 7, and it effects numerical derivatives when the gradients are handled via stochastic approximation (see also [89]). Building on these hints, we will propose alternative estimates for faster convergence. In this direction, we consider the Gaussian bell $\pi_h(a|_s)$ with covariance Σ as an approximation to the Dirac's impulse [112], and its gradient $\nabla_a \pi_h(a|_s) = \Sigma^{-1}(a - h(s))\pi_h(a|_s)$ as an approximation of the impulse's gradient. Then, the next proposition follows

Proposition 7. *Consider a family of Gaussian policies with* Σ *and let* $U_\Sigma(s; h)$ *and* $Q_\Sigma(s, a; h)$ *be the cumulative rewards and q-functions that results from such policies, respectively. Correspondingly, let* $Q_0(s, a; h) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| h, s_0 = s, a_0 = a\right]$ *be the q-function that results from a deterministic policy* $a_t = h(s_t)$*. If* $\nabla_a Q_\Sigma(s, \Sigma^{1/2}\eta + h(s), h)$ *is bounded for all* s, a, h *and* Σ*, then*

$$\lim_{\Sigma\to 0} \int Q_\Sigma(s, a; h)\Sigma^{-1}(a - h(s))\pi_h(a|_s)da = \nabla_a Q_0(s, a; h) \tag{6.31}$$

and

$$\lim_{\Sigma\to 0} \nabla_h U_\Sigma(h, \cdot) \qquad\qquad = \frac{1}{1-\gamma} \int \nabla_a Q_0(s, a; h)\rho(s)\kappa(s, \cdot)ds$$

where $\rho(s)$ *is defined such that* $\rho(s, a) = \rho(s)\pi_h(a|_s)$*.*

Proof. Integrating by parts the expression (6.31) yields

$$\begin{aligned}\int Q_\Sigma(s,a;h)\Sigma^{-1}(a-h(s))\pi_h(a|_s)da = -Q_\Sigma(s,a;h)\pi_h(a|_s)\Big|_{-\infty}^{\infty} \\ + \int \nabla_a Q_\Sigma(s,a;h)\pi_h(a|s)\,da.\end{aligned} \tag{6.32}$$

The first term is zero because $Q_\Sigma(s,a,h)$ is bounded for all s,a,h and Σ (cf., Lemma 19) and the Gaussian goes to zero at infinity. To work with the second integral, consider the following variable $\eta = \Sigma^{-1/2}(a-h(s))$. By introducing this change of variable $\pi_h(a|s)da = \phi(\eta)d\eta$, where $\phi(\eta)$ is the multivariate normal distribution. Hence, it follows that

$$\begin{aligned}\int Q_\Sigma(s,a;h)\Sigma^{-1}(a-h(s))\pi_h(a|_s)da &= \int \nabla_a Q_\Sigma(s,a;h)\pi_h(a|s)\,da \\ &= \int \nabla_a Q_\Sigma(s,\Sigma^{1/2}\eta + h(s),h)\phi(\eta)\,d\eta.\end{aligned} \tag{6.33}$$

Because $\nabla_a Q_\Sigma(s,\Sigma^{1/2}\eta+h(s),h)$ is bounded for all s,a,h and Σ we can use the Dominated Convergence Theorem to exchange the limit and the integral in (6.31) . Hence, it follows that

$$\lim_{\Sigma\to 0}\int Q_\Sigma(s,a;h)\Sigma^{-1}(a-h(s))\pi_h(a|_s)da = \int \lim_{\Sigma\to 0}\nabla_a Q_\Sigma(s,\Sigma^{1/2}\eta + h(s);h)\phi(\eta)\,d\eta. \tag{6.34}$$

We will show afterwards that indeed $\lim_{\Sigma\to 0} Q_\Sigma(s,a;h) = Q_0(s,a;)$ the q-function that results from a deterministic policy $a_t = h(s_t)$. This being the case the previous integral reduces to

$$\begin{aligned}\lim_{\Sigma\to 0}\int Q_\Sigma(s,a;h)\Sigma^{-1}(a-h(s))\pi_h(a|_s)da &= \int \nabla_a Q_0(s,h(s);h)\phi(\eta)\,d\eta \\ &= \nabla_a Q_0(s,h(s);h),\end{aligned} \tag{6.35}$$

where in the previous expression we had swaped the derivative with respect to a and the limit. The proof of this is analogous to the proof that $\lim_{\Sigma\to 0} Q_\Sigma(s,a;h) = Q_0(s,a;)$ the q-function that results from a deterministic policy $a_t = h(s_t)$. We do this next to complete the proof. Observe that for any Σ the q-function can be written as

$$Q_\Sigma(a_0,s_0;h) = \sum_{t=0}^{\infty}\gamma^t \int r(s_t,a_t)\prod_{u=0}^{t-1} p(s_{u+1}|s_u,a_u)\pi_h(a_{u+1},s_{u+1})\,d\mathbf{s}d\mathbf{a}. \tag{6.36}$$

Taking the limit with $\Sigma\to 0$, we have that $\pi_h(a|s) = \delta(a-h(s))$. Hence, the previous

expression yields

$$\begin{aligned}\lim_{\Sigma\to 0} Q_\Sigma(a_0, s_0; h) &= r(s_0, a_0) + \sum_{t=1}^{\infty} \gamma^t \int r(s_t, a_t) \prod_{u=0}^{t-1} p(s_{u+1}|s_u, a_u)\delta(a_{u+1} - h(s_{u+1})\, d\mathbf{s}d\mathbf{a}. \\ &= r(s_0, a_0) + \sum_{t=1}^{\infty} \gamma^t \int r(s_t, h(s_t))p(s_1|s_0, a_0) \prod_{u=1}^{t-1} p(s_{u+1}|s_u, h(s_u))\, d\mathbf{s}.\end{aligned} \tag{6.37}$$

Which shows that that $\lim_{\Sigma\to 0} Q_\Sigma(s, a; h)$ is indeed the q-function that results from a deterministic policy $a_t = h(s_t)$. The proof of the second part of the proposition follows analogously. □

The assumption of $\nabla_a Q_\Sigma(s, a; h)$ being bounded is satisfied if for instance the derivatives of $r(s, a)$ and $p(s_{t+1}|s, a)$ with respect to a are bounded. This interpretation of the integral in (6.31) as the gradient of $Q(s, a; h)$ can be seen from the perspective of stochastic approximation. For notational brevity we define $I_\pi := \int Q(s, a; h)\nabla_a \pi_h(a|_s)da$, and express it in terms of expectations

$$I_\pi = \mathbb{E}_{a\sim\pi_h}[Q(s, a; h)\Sigma^{-1}(a - h(s))] \tag{6.38}$$

Then, an unbiased stochastic approximation can be obtained by sampling two (or more) instances a and a' from $\pi_h(a|_s)$ and averaging as in $\hat{I}_\pi = \frac{1}{2}Q(s, a; h)\Sigma^{-1}(a - h(s)) + \frac{1}{2}Q(s, a'; h)\Sigma^{-1}(a' - h(s))$. Furthermore, if a' is the symmetric action of a with respect to the mean $h(s)$, then the estimator is still unbiased. Define the zero-mean Gaussian variable $\eta = a - h(s)$ to be the deviation of a from $h(s)$. Thus by symmetry, $a' - h(s) = -\eta$, and we can rewrite the symmetric estimate as the finite difference

$$\hat{I}_\pi = \frac{\Sigma^{-1}\eta}{2}(Q(s, h(s) + \eta; h) - Q(s, h(s) - \eta; h)), \tag{6.39}$$

revealing the gradient structure hidden in (6.38). The interpretation of (6.39) as a derivative is relevant to our policy method because it reveals the underlying reinforcement mechanisms, in the sense that the policy update favors directions that improve the reward. Fig.6.1 (left) represents the field $Q(s, a; h)$ as a function of $a \in \mathbb{R}^2$, and the gradient estimate $\hat{I}_\pi$ in (6.39) that is obtained by weighting two opposite directions with the corresponding rewards. Since the reward in the direction η is relatively higher $\hat{I}_\pi(Q)$ points in this direction.

Fig. 6.1 (right) shows that the direction of $\nabla_a Q(s, a; h)$ can be approximated more accurately at the expense of sampling the reward at $2d$ points in quadrature.

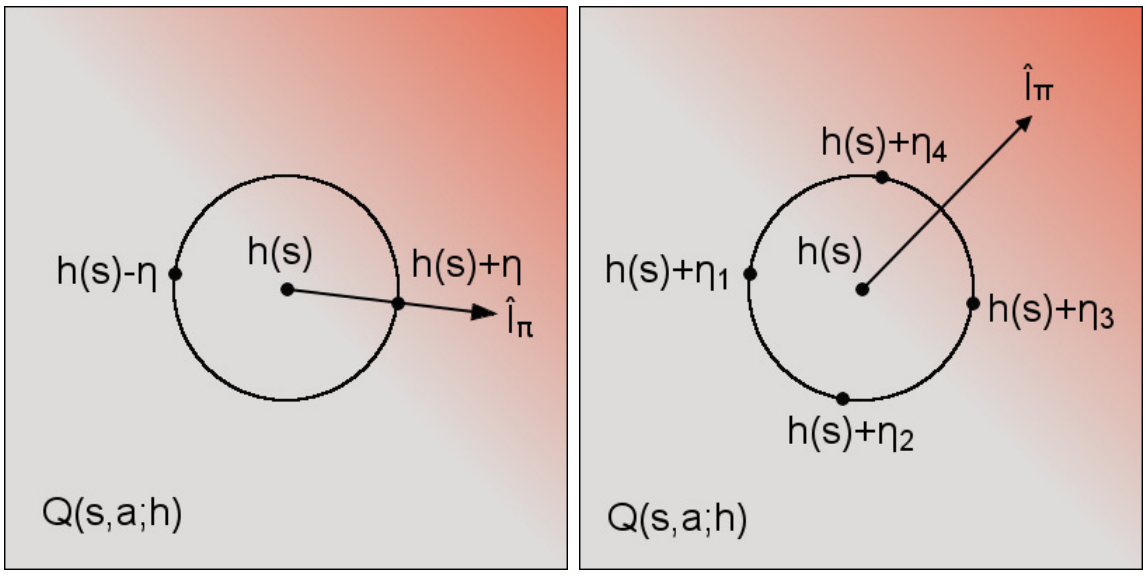


Figure 6.1: Numerical gradient via stochastic approximation; (left) two-sample approximation, (right) full-dimension. Red levels represents higher values of $Q(s, a; h)$.

6.4 Convergence Analysis for Unbiased Stochastic Gradient Ascent

This section contains the proof of Theorem 11. For this purpose let us introduce a probability space $(\Omega, \mathcal{F}, P)$ and define the following filtration defined as a sequence of increasing sigma-algebras $\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_k \subset \ldots \subset \mathcal{F}_\infty \subset \mathcal{F}$, where for each k we have that $\mathcal{F}_k$ is the sigma algebra generated by the random variables $h_0, \ldots, h_k$. Then, define a sequence the following sequence of random variables

$$V_k = U(h_k) - B \sum_{j=k}^{\infty} \eta_j^2 \tag{6.40}$$

where $B = \left(L_1\sigma^2 + L_2\eta_0\sigma^3\right)$, σ is the constant defined in Lemma 22 and L_1 and L_2 are those defined in Lemma 21. Since the sequence η_k is square summable and the expected discounted reward $U(h)$ is bounded (cf., Lemma 19), the random variable V_k is bounded for all $k \geq 0$. We next show that the sequence (6.40) is a bounded submartingale.

Lemma 14. *The sequence V_k defined in* (6.40) *is a bounded submartingale and it verifies that*

$$\mathbb{E}\left[V_{k+1}|\mathcal{F}_k\right] \geq V_k + \eta_k \left\|\nabla_h U(h_k)\right\|_{\mathcal{H}}^2. \tag{6.41}$$

Proof. According to Lemma 19 the value function $U(h_k)$ in (6.40) is upper-bounded. Thus

V_k is also upper bounded since the stepsizes are square-summable according to (6.30). Observe as well that by definition $h_k \in \mathcal{F}_k$ for all k and therefore V_k is adapted to the sequence of sigma-algebras. To show that V_k is a submartingale it suffices to show (6.41) which we do next. Writing the Taylor expansion of $U(h_{k+1})$ around h_k, yields

$$V_{k+1} = U(h_{k+1}) - B\sum_{j=k+1}^{\infty}\eta_j^2 = U(h_k) + \langle \nabla_h U(f_k), h_{k+1} - h_k\rangle_{\mathcal{H}} - B\sum_{j=k+1}^{\infty}\eta_j^2, \tag{6.42}$$

where $f_k = \lambda h_k + (1-\lambda)h_{k+1}$ with $\lambda \in [0,1]$. Adding and subtracting $\langle \nabla_h U(h_k), h_{k+1} - h_k\rangle_{\mathcal{H}}$ to the previous expression, using the Cauchy-Schwartz inequality and the result of Lemma 21 we can rewrite the previous expression as

$$\begin{aligned} V_{k+1} &= U(h_k) + \langle \nabla_h U(h_k), h_{k+1} - h_k\rangle_{\mathcal{H}} + \langle \nabla_h U(f_k) - \nabla_h U(h_k), h_{k+1} - h_k\rangle_{\mathcal{H}} - B\sum_{j=k+1}^{\infty}\eta_j^2 \\ &\geq U(h_k) + \langle \nabla_h U(h_k), h_{k+1} - h_k\rangle_{\mathcal{H}} - L_1\left\|h_{k+1} - h_k\right\|_{\mathcal{H}}^2 - L_2\left\|h_{k+1} - h_k\right\|_{\mathcal{H}}^3 - B\sum_{j=k+1}^{\infty}\eta_j^2. \end{aligned} \tag{6.43}$$

Let us consider the conditional expectation of the random variable V_{k+1} with respect to the sigma-field $\mathcal{F}_k$. Combine the monotonicity and the linearity of the expectation with the fact that h_k is measurable with respect to $\mathcal{F}_k$ to write

$$\begin{aligned} \mathbb{E}\left[V_{k+1}|\mathcal{F}_k\right] \geq U(h_k) + \langle \nabla_h U(h_k), \mathbb{E}\left[h_{k+1} - h_k|\mathcal{F}_k\right]\rangle_{\mathcal{H}} - L_1\mathbb{E}\left[\left\|h_{k+1} - h_k\right\|_{\mathcal{H}}^2|\mathcal{F}_k\right] \\ - L_2\mathbb{E}\left[\left\|h_{k+1} - h_k\right\|_{\mathcal{H}}^3|\mathcal{F}_k\right] - B\sum_{j=k+1}^{\infty}\eta_j^2. \end{aligned} \tag{6.44}$$

Substitute h_{k+1} by its expression in (6.9) to write the expectation of the quadratic term as

$$L_1\mathbb{E}\left[\left\|h_{k+1} - h_k\right\|_{\mathcal{H}}^2|\mathcal{F}_k\right] = \eta_k^2 L_1\mathbb{E}\left[\left\|\nabla_h U(h_k,\cdot)\right\|_{\mathcal{H}}^2|\mathcal{F}_k\right] \leq \eta_k^2 L_1\sigma^2, \tag{6.45}$$

where the inequality follows from the bound on the second moment of the stochastic gradient derived in Lemma 22. Likewise, using the bound for the third moment of the stochastic gradient, also in Lemma 22 and the fact that η_k is a non increasing sequence, we can write

$$L_2\mathbb{E}\left[\left\|h_{k+1} - h_k\right\|_{\mathcal{H}}^3|\mathcal{F}_k\right] \leq \eta_k^3 L_2\mathbb{E}\left[\left\|\nabla_h U(h_k,\cdot)\right\|_{\mathcal{H}}^3|\mathcal{F}_k\right] \leq \eta_k^2\eta_0 L_2\sigma^3. \tag{6.46}$$

Substituting and in (6.44) with $\eta_k^2 L_1 \sigma^2 + \eta_k^2 \eta_0 L_2 \sigma^3 = B$ denoting their sum, it results

$$\mathbb{E}\left[V_{k+1}|\mathcal{F}_k\right] \geq U(h_k) - \sum_{j=k}^{\infty} \eta_k^2 (L_1\sigma^2 + \eta_0 L_2 \sigma^3) + \langle \nabla_h U(h_k), \mathbb{E}\left[h_{k+1} - h_k|\mathcal{F}_k\right]\rangle_{\mathcal{H}} \tag{6.47}$$

$$= V_k + \langle \nabla_h U(h_k), \mathbb{E}\left[h_{k+1} - h_k|\mathcal{F}_k\right]\rangle_{\mathcal{H}}\,. \tag{6.48}$$

To complete the proof observe that according to (6.9) $h_{k+1} - h_k = \eta_k \hat{\nabla}_h U(h_k)$ and that the stochastic gradient is an unbiased estimate of the gradient (cf. Proposition 6). □

The previous Lemma establishes that V_k is a submartingale. A submartingale is in a sense a generalization of an increasing function and because it is bounded above it is expected that it converges. In fact this can be formalized (cf., [29, Theorem 5.2.8]). Moreover, the expression in (6.41) and the convergence of V_k suggest that the norm of the gradient $\|\nabla_h U(h_k, \cdot)\|$ goes to zero as k goes to infinity. We show that this is the case in what follows. By virtue of Lemma 14 it follows that the sequence V_k defined in (6.40) is a bounded submartingale and therefore it converges almost everywhere to a limiting random variable $V := \lim_{k\to\infty} V_k$ such that $\mathbb{E}|V| < \infty$ (cf., [29, Theorem 5.2.8]). Continuing the proof of Theorem 11, consider the conditional expectation of V_{k+1} with respect to the sigma algebra $\mathcal{F}_{k-1}$. Since $\mathcal{F}_{k-1} \subset \mathcal{F}_k$ it holds that

$$\mathbb{E}\left[V_{k+1}\middle|\mathcal{F}_{k-1}\right] = \mathbb{E}\left[\mathbb{E}\left[V_{k+1}\middle|\mathcal{F}_k\right]\middle|\mathcal{F}_{k-1}\right]. \tag{6.49}$$

Then, substitute (6.41) (6.49) to obtain

$$\begin{aligned}\mathbb{E}\left[V_{k+1}\middle|\mathcal{F}_{k-1}\right] &\geq \mathbb{E}\left[V_k + \eta_k \left\|\nabla_h U(h_k, \cdot)\right\|^2 \middle|\mathcal{F}_{k-1}\right] \\ &= \mathbb{E}\left[V_k\middle|\mathcal{F}_{k-1}\right] + \eta_k \mathbb{E}\left[\left\|\nabla_h U(h_k, \cdot)\right\|^2 \middle|\mathcal{F}_{k-1}\right],\end{aligned} \tag{6.50}$$

Next, use again (6.41) to lower bound the first term on the right hand side of the previous equation

$$\mathbb{E}\left[V_{k+1}\middle|\mathcal{F}_{k-1}\right] \geq V_{k-1} + \eta_{k-1}\left\|\nabla_h U(h_{k-1}, \cdot)\right\|^2 + \eta_k \mathbb{E}\left[\left\|\nabla_h U(h_k, \cdot)\right\|^2 \middle|\mathcal{F}_{k-1}\right]. \tag{6.51}$$

Repeating this procedure of conditioning on the previous sigma algebras recursively one obtains

$$\mathbb{E}\left[V_{k+1}\right] \geq V_0 + \eta_0 \left\|\nabla_h U(h_0, \cdot)\right\|^2 + \sum_{j=1}^{k} \eta_j \mathbb{E}\left[\left\|\nabla_h U(h_j, \cdot)\right\|^2\right]. \tag{6.52}$$

Since V_k is a sequence of bounded random variables, then by virtue of the Dominated Convergence Theorem we have that $\mathbb{E}\left[V\right] = \lim_{k\to\infty} \mathbb{E}\left[V_k\right]$. This result applied to the

previous inequality results in

$$\mathbb{E}\left[V\right] \geq V_0 + \eta_0 \left\|\nabla_h U(h_0,\cdot)\right\|^2 + \sum_{j=1}^{\infty} \eta_j \mathbb{E}\left[\left\|\nabla_h U(h_j,\cdot)\right\|^2\right], \tag{6.53}$$

with $\mathbb{E}|V| < \infty$, hence

$$\sum_{j=1}^{\infty} \eta_j \mathbb{E}\left[\left\|\nabla_h U(h_j,\cdot)\right\|^2\right] < \infty. \tag{6.54}$$

The monotone convergence theorem applied to the sum $\sum_{j=1}^{k} \eta_j \left\|\nabla_h U(h_j,\cdot)\right\|^2$ implies that

$$\lim_{k\to\infty} \mathbb{E}\left[\sum_{j=1}^{k} \eta_k \left\|\nabla_h U(h_j,\cdot)\right\|^2\right] = \mathbb{E}\left[\sum_{j=1}^{\infty} \eta_k \left\|\nabla_h U(h_j,\cdot)\right\|^2\right]. \tag{6.55}$$

Since the left hand side of the previous expression is bounded by virtue of (6.54) the latter implies that

$$\lim_{k\to\infty} \sum_{j=0}^{k} \eta_j \left\|\nabla_h U(h_j,\cdot)\right\|^2 < \infty \quad \text{a.e.} \tag{6.56}$$

Because the sequence of step sizes η_j is non-summable (cf., (6.30)) the previous expression implies that

$$\liminf_{k\to\infty} \left\|\nabla_h U(h_k,\cdot)\right\|^2 = 0. \tag{6.57}$$

We are left to show that $\limsup_{k\to\infty} \left\|\nabla_h U(h_k,\cdot)\right\| = 0$ almost everywhere, which we do by contradiction. Assume that $\limsup_{k\to\infty} \left\|\nabla_h U(h_k(\omega),\cdot)\right\| = \epsilon > 0$ for some $\omega \in \Omega$. Then, there exist subsequences $\{m_j\}$ and $\{n_j\}$ such that $m_j < n_j < m_{j+1}$ and

$$\left\|\nabla_h U(h_k,\cdot)\right\| > \frac{\epsilon}{3} \tag{6.58}$$

for $m_j \leq k < n_j$ and

$$\left\|\nabla_h U(h_k,\cdot)\right\| \leq \frac{\epsilon}{3} \tag{6.59}$$

for $n_j \leq k < m_{j+1}$, where we have dropped the ω to simplify the notation, but hereafter we argue for a specific sample point in the probability space. It is proved in Lemma 23 in the appendix, that the sequence

$$S_k = \sum_{j=0}^{k} \eta_j \left(\hat{\nabla}_h U(h_j) - \nabla_h U(h_j)\right) = \sum_{j=0}^{k} \eta_j e_j \tag{6.60}$$

converges to a finite limit with probability one. By virtue of this result and (6.56) there

exists $\bar{j}$ such that

$$\sum_{k=m_{\bar{j}}}^{\infty} \eta_k \left\|\nabla_h U(h_k,\cdot)\right\|^2 < \min\left\{\frac{\epsilon^2}{36L_1}, \frac{\epsilon^{3/2}}{6\sqrt{6L_2}}\right\} \tag{6.61}$$

and

$$\left\|\sum_{k=m_{\bar{j}}}^{\infty} \eta_k e_k\right\| < \min\left\{\frac{\epsilon}{12L_1}, \frac{\epsilon^{1/2}}{2\sqrt{6L_2}}\right\} \tag{6.62}$$

For any $j \geq \bar{j}$ and any m with $m_j \leq m < n_j$, by virtue of Lemma 21, we have

$$\left\|\nabla_h U(h_{n_j},\cdot) - \nabla_h U(h_m,\cdot)\right\| \leq L_1 \left\|h_{n_j} - h_m\right\| + L_2 \left\|h_{n_j} - h_m\right\|^2, \tag{6.63}$$

Recall that the difference $h_{n_j} - h_m$ can be written as

$$h_{n_j} - h_m = \sum_{k=m}^{n_j-1} \eta_k \hat{\nabla}_h U(h_k,\cdot) = \sum_{k=m}^{n_j-1} \eta_k \nabla_h U(h_k,\cdot) + \sum_{k=m}^{n_j-1} \eta_k \left(\hat{\nabla}_h U(h_k,\cdot) - \nabla_h U(h_k,\cdot)\right). \tag{6.64}$$

Thus, defining the error $e_k = \hat{\nabla}_h U(h_k,\cdot) - \nabla_h U(h_k,\cdot)$, the following upper bound holds

$$\left\|h_{n_j} - h_m\right\| \leq \sum_{k=m}^{n_{j-1}} \eta_k \left\|\nabla_h U(h_k,\cdot)\right\| + \left\|\sum_{k=m}^{n_{j-1}} \eta_k e_k\right\| \leq \frac{3}{\epsilon} \sum_{k=m}^{n_{j-1}} \eta_k \left\|\nabla_h U(h_k,\cdot)\right\|^2 + \left\|\sum_{k=m}^{n_{j-1}} \eta_k e_k\right\|, \tag{6.65}$$

where in the last inequality we used that that according to (6.58) for all k such $m \leq k < n_j$ we have that $(3/\epsilon)\left\|\nabla_h U(h_k,\cdot)\right\| \geq 1$. Using the bounds on the tails (6.61) and (6.4) it holds that

$$\left\|h_{n_j} - h_m\right\| \leq \frac{3}{\epsilon}\frac{\epsilon^2}{36L_1} + \frac{\epsilon}{12L_1} = \frac{\epsilon}{6L_1} \tag{6.66}$$

and that

$$\left\|h_{n_j} - h_m\right\| \leq \frac{3}{\epsilon}\frac{\epsilon^{3/2}}{6\sqrt{6L_2}} + \frac{\epsilon^{1/2}}{2\sqrt{6L_2}} = \sqrt{\frac{\epsilon}{6L_2}}. \tag{6.67}$$

Replacing the previous bounds in (6.63) yields $\left\|\nabla_h U(h_{n_j},\cdot) - \nabla_h U(h_m)\right\| \leq \epsilon/3$. The latter together with (6.59) implies that $\left\|\nabla_h U(h_m,\cdot)\right\| < 2\epsilon/3$ for all m such $m_j \leq m < n_j$, which contradicts (6.59) and therefore the assumption that $\limsup_{k\to\infty} \left\|\nabla_h U(h_k,\cdot)\right\| > 0$. Hence, it must hold that $\lim_{k\to\infty} \left\|\nabla_h U(h_k,\cdot)\right\| = 0$.

6.5 Sparse Projections in the Function Space

As observed before, the update (6.9) requires the introduction of a new element $\kappa(s_{T_k}, \cdot)$ of the kernel dictionary at each iteration, thus resulting in memory explosion. To overcome this limitation we modify the stochastic gradient ascent by introducing a projection over a RKHS of lower dimension as long as the induced error remains below a given compression budget. This algorithm is known as Orthogonal Match and Pursuit [129] and we summarize and adapt it to policy gradient ascent it in Algorithm 4. Starting with the policy $h_0 \equiv 0$, each stochastic gradient ascent iteration defines a new policy

$$\tilde{h}_{k+1} = h_k + \eta \hat{\nabla}_h U(h_k, \cdot), \tag{6.68}$$

where $\hat{\nabla}_h U(h_k, \cdot)$ is that in (6.16). The difference between the updates (6.68) and (6.29) is that in (6.68) $h_k = \sum_{j=1}^{M_k} w_j^{(k)} \kappa(s_j^{(k)}, \cdot)$ is represented by a reduced $M_k \leq k$ number of states $s_j^{(k)}$ and weights $w_j^{(k)}$, as it results from the pruning procedure below, (cf., $M_k = k$ for h_{k+1} in (6.29)).

With state s_{T_k} being s_T in step 8 of Algorithm 2, and

$$\tilde{w}_k := \eta \frac{\hat{Q}(s_{T_k}, a_{T_k}; h_k) - \hat{Q}(s_{T_k}, \bar{a}_{T_k}; h_k)}{2(1-\gamma)} \Sigma^{-1}(a_{T_k} - h_k(s_{T_k})), \tag{6.69}$$

$$\tilde{h}_{k+1} = \sum_{j=1}^{M_k} w_j^{(k)} \kappa(s_j^{(k)}, \cdot) + \tilde{w}_k \kappa(S_{T_k}, \cdot). \tag{6.70}$$

Hence, h_k is represented by dictionary $D_k = [s_1^{(k)}, \ldots, s_{M_k}^{(k)}]$ and associated weights $\mathbf{w}_k = \left[\left(w_1^{(k)} \right)^\top, \ldots, \left(w_{M_k}^{(k)} \right)^\top \right]^\top$, and $\tilde{h}_{k+1}$ is represented by the updated $\tilde{D}_{k+1} = [D_k, s_{T_k}]$ and $\tilde{\mathbf{w}}_{k+1} = [\mathbf{w}_k^\top, \tilde{w}_k^\top]^\top$, which has model order $\tilde{M}_{k+1} = M_k + 1$. Then, to avoid memory explosion, we prune the dictionary as long as the induced error stays below a prescribed bound $\epsilon > 0$. We start by storing copies of the previous dictionary, i.e., define $D_{k+1} = \tilde{D}_{k+1}$ and $\mathbf{w}_{k+1} = \tilde{\mathbf{w}}_{k+1}$. Let $\mathcal{H}_{D_{k+1}^j}$ be the space spanned by all the elements of D_{k+1} except for the j-th one. For each $j = 1 \ldots M_{k+1}$ we identify the less informative dictionary element by solving

$$e_j = \min_{h \in \mathcal{H}_{D_{k+1}^j}} \left\| h - \tilde{h}_{k+1} \right\|_{\mathcal{H}}^2 = \mathbf{c}_j + \min_{\mathbf{w} \in \mathbb{R}^{pM_{k+1}-1}} \mathbf{w}^\top \mathbf{K}_{D_{k+1}^j, D_{k+1}^j} \mathbf{w} - 2\mathbf{w}^\top \mathbf{K}_{D_{k+1}^j, \tilde{D}_{k+1}} \tilde{\mathbf{w}}_{k+1}, \tag{6.71}$$

which results from expanding the square after substituting h and $\tilde{h}_{k+1}$ by their representa-

tions as weighted sums of kernel elements, and upon defining the block matrices $\mathbf{K}_{D_{k+1}^j, D_{k+1}^j}$ and $\mathbf{K}_{D_{k+1}^j, \tilde{D}_{k+1}}$ whose (l,m)-th blocks of size $p \times n$ are $\kappa(s_l^{(k)}, s_m^{(k)})$ and $\kappa(s_l^{(k)}, \tilde{s}_m^{(k)})$, respectively, with $s_l^{(k)}$ and $s_m^{(k)}$ being the l-th and m-th elements of D_{k+1}^j, and with $\tilde{s}_l^{(k)}$ correspondingly in $\tilde{D}_{k+1}$. Problem (6.71) is a least-squares problem with the following

Algorithm 4 Kernel Orthogonal Matching Pursuit (KOMP)

Input: function $\tilde{h}_k$ defined by Dictionary $\tilde{D}_k \in \mathbb{R}^{n \times \tilde{M}_k}$ weights $\tilde{\mathbf{w}}_k \in \mathbb{R}^{p \times \tilde{M}_k}$ and compression budget $\epsilon > 0$

Initialize: $D_k = \tilde{D}_k$, $W_k = \tilde{W}_k$, $M_k = \tilde{M}_k$, $e^* = 0$
while $e^* < \epsilon$ and $M_k > 0$ **do**
  for $j = 1 \dots M_k$ **do**
    Find minimal error e_j by solving (6.71)
  end for
  Less informative element $j^* = \text{argmin}_j \, e_j$
  Save error $e^* = e_{j^*}$
  if Error smaller than compression budget $e^* < \epsilon$ **then**
    Prune Dict., $D_k \leftarrow D_k^{j^*}$, $M_k \leftarrow M_k - 1$
    Update Weights as in (6.72)

$$\mathbf{w}_k = \mathbf{K}^{\dagger}_{D_k^j, D_k^j} \mathbf{K}_{D_k^j, \tilde{D}_k} \tilde{\mathbf{w}}_k$$

  end if
end while
return $D_k, \mathbf{w}_k$

closed-form solution

$$\mathbf{w}_j^* = \mathbf{K}^{\dagger}_{D_{k+1}^j, D_{k+1}^j} \mathbf{K}_{D_{k+1}^j, \tilde{D}_{k+1}} \tilde{\mathbf{w}}_{k+1}, \tag{6.72}$$

where, $(\cdot)^{\dagger}$ denotes the Moore-Penrose pseudo-inverse. After computing all compression errors e_j we chose the dictionary element that yields the smallest error $j^* = \text{argmin}_{j=1 \dots M_{k+1}} e_j$, we remove the j^*-th column from the dictionary D_{k+1}, i.e., we redefine $D_{k+1} = D_{k+1}^{j^*}$ and the model order $M_{k+1} = M_{k+1} - 1$ and update the corresponding weights as $\mathbf{w}_{k+1} = \mathbf{w}_{j^*}^*$. We repeat the process as long as the minimum compression error remains below the compression budget, i.e., $\min_{j=1 \dots M_{k+1}} e_j < \epsilon$. The output of the pruning process is a function h_{k+1} that is represent by at most the same number of elements than $\tilde{h}_{k+1}$ and such that the error introduced in this approximation is, by construction, smaller than the compression budget ϵ. This output can be interpreted as a projection over a RKHS of smaller dimension. Let D_{k+1} be the dictionary that Algorithm 4 outputs. Then, the resulting policy can be expressed as

$$h_{k+1} = \mathcal{P}_{\mathcal{H}_{D_{k+1}}} \left[\tilde{h}_{k+1} \right] = \mathcal{P}_{\mathcal{H}_{D_{k+1}}} \left[h_k + \eta \hat{\nabla}_h U(h_k, \cdot) \right], \tag{6.73}$$

where the operation $\mathcal{P}_{\mathcal{H}_{D_{k+1}}}[\cdot]$ refers to the projection onto the RKHS spanned by the dictionary D_{k+1}. The algorithm described by (6.68) and (6.73) is summarized in Algorithm 5. By projecting over a smaller subspace we control the model order of the policy h_k. However, the induced error translates into an estimation bias on the estimate of $\nabla_h U(h,\cdot)$ as we detail in the next proposition

Algorithm 5 Projected Stochastic Policy Gradient Ascent

Input: step size η_0, compression budget ϵ

1: *Initialize*: $h_0 = 0$
2: **for** $k = 0 \dots$ **do**
3: Compute $\hat{\nabla}_h U(h_k,\cdot) = \text{StochasticGradient}(h_k)$
4: Update policy via stochastic gradient ascent

$$\tilde{h}_{k+1} = h_k + \eta\hat{\nabla}_h U(h_k,\cdot)$$

5: Reduce model order $h_{k+1} = \text{KOMP}(\tilde{h}_{k+1}, \epsilon)$
6: **end for**

Proposition 8. *The update of Algorithm 5 is equivalent to running biased stochastic gradient ascent, with bias*

$$b_k = \mathcal{P}_{\mathcal{H}_{D_{k+1}}}\left[h_k + \eta\hat{\nabla}_h U(h_k,\cdot)\right] - \left(h_k + \eta\hat{\nabla}_h U(h_k,\cdot)\right). \tag{6.74}$$

bounded by the compression budget ε for all k.

This proposition allow us to rewrite (6.73) as

$$h_{k+1} = h_k + \eta\hat{\nabla}_h U(h,\cdot) + b_k, \tag{6.75}$$

with $\|b_k\| \leq \epsilon$.

Proof. From (6.73) and adding and subtracting $\eta\hat{\nabla}_h U(h_k,\cdot)$, it is possible to write the difference $h_{k+1} - h_k$ as

$$h_{k+1} - h_k = \mathcal{P}_{\mathcal{H}_{D_{k+1}}}\left[h_k + \eta\hat{\nabla}_h U(h_k,\cdot)\right] - \left(h_k + \eta\hat{\nabla}_h U(h_k,\cdot)\right) + \eta\hat{\nabla}_h U(h_k,\cdot). \tag{6.76}$$

Using the definition of the bias (6.74) the previous expression can be written as

$$h_{k+1} = h_k + \eta\hat{\nabla}_h U(h,\cdot) + b_k. \tag{6.77}$$

To complete the proof, notice that by definition b_k is the error of the compression and thus its norm is bounded by the compression budget ε. □

As stated by the previous proposition the effect of introducing the KOMP algorithm is that of updating the policy by running gradient ascent, where now the estimate is biased. Hence, we claim in the following result that Stochastic Policy Gradient Ascent (Algorithm 5) converges to a neighborhood of a critical point of the expected discounted reward, whose size depends on the step-size of the algorithm as well as on compression error allowed. However, whereas the model order of the function obtained via stochastic gradient ascent without projection (Algorithm 3) could grow without bound, for the projected version we can ensure that the model order obtained is always bounded. We formalize these results next.

Theorem 12. *Let $\eta > 0$ and $\epsilon > 0$ for all $k \geq 0$. Then there exists a constant $C := C(\gamma, \eta, \epsilon, \Sigma, B_r,)$ such that*

$$\liminf_{k\to\infty} \|\nabla_h U(h_k, \cdot)\|_{\mathcal{H}} \leq \frac{\epsilon}{2\eta} + \frac{\sqrt{\epsilon^2 + 4\eta^3 C}}{2\eta}, \tag{6.78}$$

with probability one. Moreover, there exists a constant $M := M(\epsilon) > 0$ such that for every $k \geq 0$ the model order M_k needed to represent the function h_k is such that $M_k \leq M$.

Proof. The proof of this result is the matter of Section 6.6. □

Observe that the optimal selection is $\epsilon = O(\eta^{3/2})$ in the sense that selecting a smaller compression factor, the total error bound is of $O(\eta^{3/2})$. In that sense, such selection is not optimal, because we force a small compression error – which entails larger model order – and there is no benefit in terms of the convergence error. Then the parameter η is to be chosen trading-off accuracy for speed of convergence.

6.6 Convergence Analysis of Sparse Policy Gradient

This section contains the proof of Theorem 12. It starts by providing a lower bound on the expectation of random variables $U(h_{k+1})$ conditioned to the sigma field $\mathcal{F}_k$

Lemma 15. *The sequence of random variables $U(h_k)$ satisfies the following inequality*

$$\mathbb{E}\left[U(h_{k+1})|\mathcal{F}_k\right] \geq U(h_k) - \eta^2 C + \eta \|\nabla_h U(h_k)\|_{\mathcal{H}} \left(\|\nabla_h U(h_k)\|_{\mathcal{H}} - \frac{\epsilon}{\eta} \right), \tag{6.79}$$

where C is the following positive constant

$$C = L_1 \left(\sigma^2 + 2\frac{\epsilon}{\eta}\sigma + \frac{\epsilon^2}{\eta^2} \right) + \eta L_2 \left(\sigma^2 + 2\frac{\epsilon}{\eta}\sigma + \frac{\epsilon^2}{\eta^2} \right)^{3/2}, \tag{6.80}$$

where L_1 and L_2 are the constants defined in Lemma 21 and σ is the constant defined in Lemma 22.

Proof. Consider the Taylor expansion of $U(h_{k+1})$ around h_k,

$$U(h_{k+1}) = U(h_k) + \langle \nabla_h U(f_k, \cdot), h_{k+1} - h_k \rangle_{\mathcal{H}} . \tag{6.81}$$

where $f_k = \lambda h_k + (1-\lambda) h_{k+1}$ with $\lambda \in [0,1]$. Adding and subtracting

$$\langle \nabla_h U(h_k, \cdot), h_{k+1} - h_k \rangle_{\mathcal{H}} \tag{6.82}$$

to the previous expression, using the Cauchy-Schwartz inequality and the result of Lemma 21 we can rewrite (6.81) as

$$\begin{aligned} U(h_{k+1}) &= U(h_k) + \langle \nabla_h U(h_k, \cdot), h_{k+1} - h_k \rangle_{\mathcal{H}} + \langle \nabla_h U(f_k, \cdot) - \nabla_h U(h_k, \cdot), h_{k+1} - h_k \rangle_{\mathcal{H}} \\ &\geq U(h_k) + \langle \nabla_h U(h_k, \cdot), h_{k+1} - h_k \rangle_{\mathcal{H}} - L_1 \left\| h_{k+1} - h_k \right\|_{\mathcal{H}}^2 - L_2 \left\| h_{k+1} - h_k \right\|_{\mathcal{H}}^3 . \end{aligned} \tag{6.83}$$

Let us consider the conditional expectation of the random variable $U(h_{k+1})$ with respect to the sigma-field $\mathcal{F}_k$. Combine the monotonicity and the linearity of the expectation with the fact that h_k is measurable with respect to $\mathcal{F}_k$ to write

$$\begin{aligned} \mathbb{E}\left[U(h_{k+1}) | \mathcal{F}_k\right] \geq U(h_k) &+ \langle \nabla_h U(h_k, \cdot), \mathbb{E}\left[h_{k+1} - h_k | \mathcal{F}_k\right] \rangle_{\mathcal{H}} \\ &- L_1 \mathbb{E}\Big[\left\| h_{k+1} - h_k \right\|_{\mathcal{H}}^2 | \mathcal{F}_k \Big] - L_2 \mathbb{E}\Big[\left\| h_{k+1} - h_k \right\|_{\mathcal{H}}^3 | \mathcal{F}_k \Big]. \end{aligned} \tag{6.84}$$

Substitute (6.75) for h_{k+1} to write the expectation of the quadratic term in the right hand side of (6.84) as

$$L_1 \mathbb{E}\left[\left\| h_{k+1} - h_k \right\|_{\mathcal{H}}^2 | \mathcal{F}_k \right] = L_1 \mathbb{E}\left[\left\| \eta \hat{\nabla}_h U(h, \cdot) + b_k \right\|_{\mathcal{H}}^2 | \mathcal{F}_k \right] \tag{6.85}$$

$$\leq L_1 \left(\eta^2 \mathbb{E}\left[\left\| \hat{\nabla}_h U(h_k, \cdot) \right\|_{\mathcal{H}}^2 | \mathcal{F}_k \right] + 2\eta\epsilon \mathbb{E}\left[\left\| \hat{\nabla}_h U(h_k, \cdot) \right\|_{\mathcal{H}} | \mathcal{F}_k \right] + \epsilon^2 \right), \tag{6.86}$$

where we have used that $\|b_k\| \leq \epsilon$ as stated in Proposition 8. Using the bounds provided in Lemma 22, the previous expression can be upper bounded by

$$L_1 \mathbb{E}\left[\left\| h_{k+1} - h_k \right\|_{\mathcal{H}}^2 | \mathcal{F}_k \right] \leq \eta^2 L_1 \left(\sigma^2 + 2\frac{\epsilon}{\eta}\sigma + \frac{\epsilon^2}{\eta^2} \right). \tag{6.87}$$

With a similar procedure we obtain

$$L_2\mathbb{E}\left[\|h_{k+1}-h_k\|_{\mathcal{H}}^3\,|\mathcal{F}_k\right] \leq \eta^2\eta_0 L_2\left(\sigma^2+2\frac{\epsilon}{\eta}\sigma+\frac{\epsilon^2}{\eta^2}\right)^{3/2} \tag{6.88}$$

Observe that the sum of (6.87) and (6.88) is equal to $\eta^2 C$ in (6.80). Then, substitute (6.87) and (6.88) in (6.84) to obtain

$$\mathbb{E}\left[U(h_{k+1})|\mathcal{F}_k\right] \geq U(h_k) - C\eta^2 + \langle \nabla_h U(h_k), \mathbb{E}\left[h_{k+1}-h_k|\mathcal{F}_k\right]\rangle_{\mathcal{H}}\,. \tag{6.89}$$

Finally, (6.79) results from applying the Cauchy-Schwartz inequality to the inner product in (6.89) and then substituting (6.75) for h_{k+1}, with $\|b_k\| \leq \epsilon$.

□

The previous Lemma establishes a lower bound on the expectation of $U(h_{k+1})$ conditioned to the sigma algebra $\mathcal{F}_k$. This lower bound however, is not enough for $U(h_k)$ to be a submartingale, since the sign of the term added to $U(h_k)$ in the right hand side of (6.79) depends on the norm of $\nabla_h U(h_k)$. This is in contrast with the situation in Lemma 14, where the term was always positive. The origin of this issue lies on the bias introduced by the sparsification. However, when the norm of the gradient is large the term is negative and we have a submartingale while a neighborhood of the critical point is not reached. To formalize this idea let us define the neighborhood as

$$\|\nabla_h U(h_k,\cdot)\|_{\mathcal{H}} \leq \frac{\epsilon}{2\eta}+\frac{\sqrt{\epsilon^2+4\eta^3 C}}{2\eta}, \tag{6.90}$$

and the corresponding stopping time

$$N = \min_{k\geq 0}\left\{\|\nabla_h U(h_k,\cdot)\|_{\mathcal{H}} \leq \frac{\epsilon}{2\eta}+\frac{\sqrt{\epsilon^2+4\eta^3 C}}{2\eta}\right\}. \tag{6.91}$$

In order to prove (6.78) we will argue that either the limit exists and satisfies the bound in (6.78), or $P(N<\infty)=1$, in which case (6.90) must be recursively satisfied after a finite number of iterations so that (6.78) holds. In this direction we define $V_k = (U(h^*)-U(h_k))\,\mathbb{1}(k\leq N)$, with $\mathbb{1}(\cdot)$ being the indicator function, and prove that V_k is a non-negative submartingale. Indeed, since $U(h^*)$ maximizes $U(h)$, V_k is always non-negative. In addition $V_k \in \mathcal{F}_k$ since $U(h_k)\in\mathcal{F}_k$ and $\mathbb{1}(k-1\leq N)\in\mathcal{F}_k$. To show that $\mathbb{E}[V_{k+1}|\mathcal{F}_k]\leq V_k$ start by using that $\mathbb{1}(k\leq N)\in\mathcal{F}_k$ and write

$$\mathbb{E}\left[V_{k+1}|\mathcal{F}_k\right] = \mathbb{1}(k+1\leq N)\mathbb{E}\left[U(h^*)-U(h_{k+1})|\mathcal{F}_k\right] \tag{6.92}$$

Using (6.79) we can upper bound $\mathbb{E}\left[V_{k+1}|\mathcal{F}_k\right]$ as

$$\mathbb{E}\left[V_{k+1}|\mathcal{F}_k\right] \leq \mathbb{1}(k+1 \leq N)\left(U(h^*) - U(h_k)\right) - \mathbb{1}(k+1 \leq N)W_k. \tag{6.93}$$

with

$$W_k := \eta \left\|\nabla_h U(h_k, \cdot)\right\|_{\mathcal{H}}^2 - \epsilon \left\|\nabla_h U(h_k, \cdot)\right\|_{\mathcal{H}} - \eta^2 C \tag{6.94}$$

Notice that the bound in (6.90) is root of (6.94) as a polynomial in the variable $\|\nabla_h U(h_k, \cdot)\|$. It follows that $W_k > 0$ as long as $k < N$, so that $\mathbb{1}(k+1 \leq N)W_k \geq 0$ for all k. Also notice that the indicator function $\mathbb{1}(k \leq N)$ is non-increasing with k, so that $\mathbb{1}(k+1 \leq N) \leq \mathbb{1}(k+1 \leq N)$. Using these two facts, it follows from (6.93) that $\mathbb{E}[V_{k+1}|\mathcal{F}_k] \leq V_k$. Thus, V_k is a nonnegative submartingale and therefore it converges to random variable V such that $\mathbb{E}[V] \leq \mathbb{E}[V_0]$ (see e.g., [29, Theorem 5.29]). Rearranging the terms in (6.93) and considering the total expectation in both sides of the inequality we have that

$$\mathbb{E}\left[\sum_{j=0}^{k} \mathbb{1}(j < N)W_k\right] \leq \mathbb{E}[V_0] - \mathbb{E}\left[V_{k+1}\right]. \tag{6.95}$$

Again, by definition of the stopping time N, $\mathbb{1}(k < N)W_k$ is nonnegative, and thus the sequence of random variables

$$S_k = \sum_{j=0}^{k} \mathbb{1}(j < N)W_j, \tag{6.96}$$

is monotonically increasing. Hence, use the Monotone Convergence Theorem (see e.g., [29, Theorem 1.6.6]) to write

$$\lim_{k\to\infty} \mathbb{E}\left[\sum_{j=0}^{k} \mathbb{1}(j < N)W_j\right] = \mathbb{E}\left[\sum_{j=0}^{\infty} \mathbb{1}(j < N)W_j\right]. \tag{6.97}$$

On the other hand, $U(h_k)$ is bounded according to Lemma 19, thus V_k is a bounded sequence and then we use the Dominated Convergence Theorem (see e.g. [29, Theorem 1.6.7]) to obtain

$$\mathbb{E}[V] = \mathbb{E}[\lim_{k\to\infty} V_k] = \lim_{k\to\infty} \mathbb{E}[V_k]. \tag{6.98}$$

Taking the limit of k going to infinity in both sides of (6.95) and using (6.97) and (6.98)

we have that

$$\mathbb{E}\left[\sum_{j=0}^{\infty}\mathbb{1}(j<N)W_j\right]\leq\mathbb{E}[V_0]-\mathbb{E}[V]<\infty. \tag{6.99}$$

Observe that the expectation on the left hand side of the previous expression can be computed as

$$P(N<\infty)\mathbb{E}\left[\sum_{j=0}^{N-1}W_j\Big|_{N<\infty}\right]+P(N=\infty)\mathbb{E}\left[\sum_{j=0}^{\infty}W_j\Big|_{N=\infty}\right]. \tag{6.100}$$

By virtue of Lemma 20, $\|\nabla_h U(h,\cdot)\|$ is uniformly bounded for all $h\in\mathcal{H}$. Thus, the first sum in the previous expression is finite. Hence,

$$P(N=\infty)\mathbb{E}\left[\sum_{j=0}^{\infty}W_j\Big|N=\infty\right]<\infty. \tag{6.101}$$

The latter can only hold if $P(N=\infty)=0$ or if the expectation of the sum is bounded. If the former happens it means that infinitely often $\|\nabla_h U(h,\cdot)\|$ visits the neighborhood (6.90), and thus (6.78) holds. It remains to analyze the case where the expectation of the sum is finite. Using the Monotone Convergence Theorem one can exchange the expectation with the sum and therefore we have that

$$\sum_{j=0}^{\infty}\mathbb{E}\left[W_j\Big|N=\infty\right]<\infty, \tag{6.102}$$

which implies that $\lim_{k\to\infty}\mathbb{E}[W_k|N=\infty]=0$. Thus

$$\lim_{k\to\infty}\mathbb{E}\left[\left(\eta\left\|\nabla_h U(h_k,\cdot)\right\|^2-\left\|\nabla_h U(h_k,\cdot)\right\|\epsilon-\eta^2C\right)\right]=0. \tag{6.103}$$

Moreover, because the norm of the gradient is bounded, the Dominated Convergence Theorem allows us to write

$$\mathbb{E}\left[\lim_{k\to\infty}\left(\eta\left\|\nabla_h U(h_k,\cdot)\right\|^2-\left\|\nabla_h U(h_k,\cdot)\right\|\epsilon-\eta^2C\right)\right]=0. \tag{6.104}$$

Because the random variable is nonnegative it must hold that

$$\lim_{k\to\infty}\|\nabla_h U(h_k,\cdot)\|_{\mathcal{H}}=\frac{\epsilon}{2\eta}+\frac{\sqrt{\epsilon^2+4\eta^3C}}{2\eta}. \tag{6.105}$$

Thus, (6.78) holds as well if $P(N=\infty)>0$. It remains to be shown that the model

order of the representation is bounded for all k. The proof of this result is identical to that in [60, Theorem 3].

6.7 Numerical Experiments

We benchmarked Stochastic Projected Stochastic Policy Gradient Ascent on a classic control problem, the Continuous Mountain Car [2], which is featured in OpenAI Gym [1]. In this problem, the state space is $n = 2$ dimensional, consisting of position and velocity, bounded within $[-1.2, 0.6]$ and $[-0.07, 0.07]$, respectively. The action space is a scalar representing the real valued force on the car. The reward function is 100 when the car reaches the goal at position 0.6, and in every episode it substract $0.1 \sum_{t=t_0}^{t_f} a_t^2$, where a_t are the actions selected. Because of the penalization of the actions, in the space of policies there are local maxima around policies that keep the car stationary in order to realize roughly zero reward. In order to avoid converging to such policy, we set h_0 to have kernels at $(0.65, -0.02)$ and $(-0.35, 0.02)$ with respective weights 0.5 and -0.5. In particular, we work with Gaussian kernels, that are nonsymmetric due to the difference in the scales of position and velocities attained by the mountain cart. Their covariance matrix is given by $diag([0.15, 0.015])$. The results obtained with Algorithm 5 for the following paramters: $\gamma = 0.001$, $\Sigma = 1.0$, $\eta = 0.0005$ and $\epsilon = 0.005$ are given in figures 6.2 and 6.3. In the former, we plot the average reward during training (top figure), and the model order (bottom figure). The policy learned from this experiment is given in Figure 6.3, where we plot the policy learned after after 50,000 iterations. From 6.2 we can observe that the policy converges to a solution that allows to solve the problem in about 15000 training examples with the exception of the two dips that can be observed around 10,000 and 30,000 iterations. These are probably the result of using a step size that is too large since due to the sensitivity of the reward function – which incurs a positive reward only when the objective is reached – small changes in the policy might entail large changes in the reward. The challenge in the mountain car is that by just accelerating to the right it is not possible to escape the valey. Hence, the optimal policy needs to be such that it increases its velocity. In particular, in Figure 6.3 we can observe that for positive velocities the acceleration is mostly positive, while when the velocity is negative the force is also negative.

In contrast to other Kernel based RL algorithms, such as [126], ours manages to significantly reduce the computational complexity by only updating the dictionary after a sequence of actions. In practice, our algorithm performs cheap actions (as measured by time and computational complexity) in order to perform relatively few computationally intensive learning steps. In particular, the most costly subroutine is KOMP (Algorithm 4) and we resource to it only once per episode.

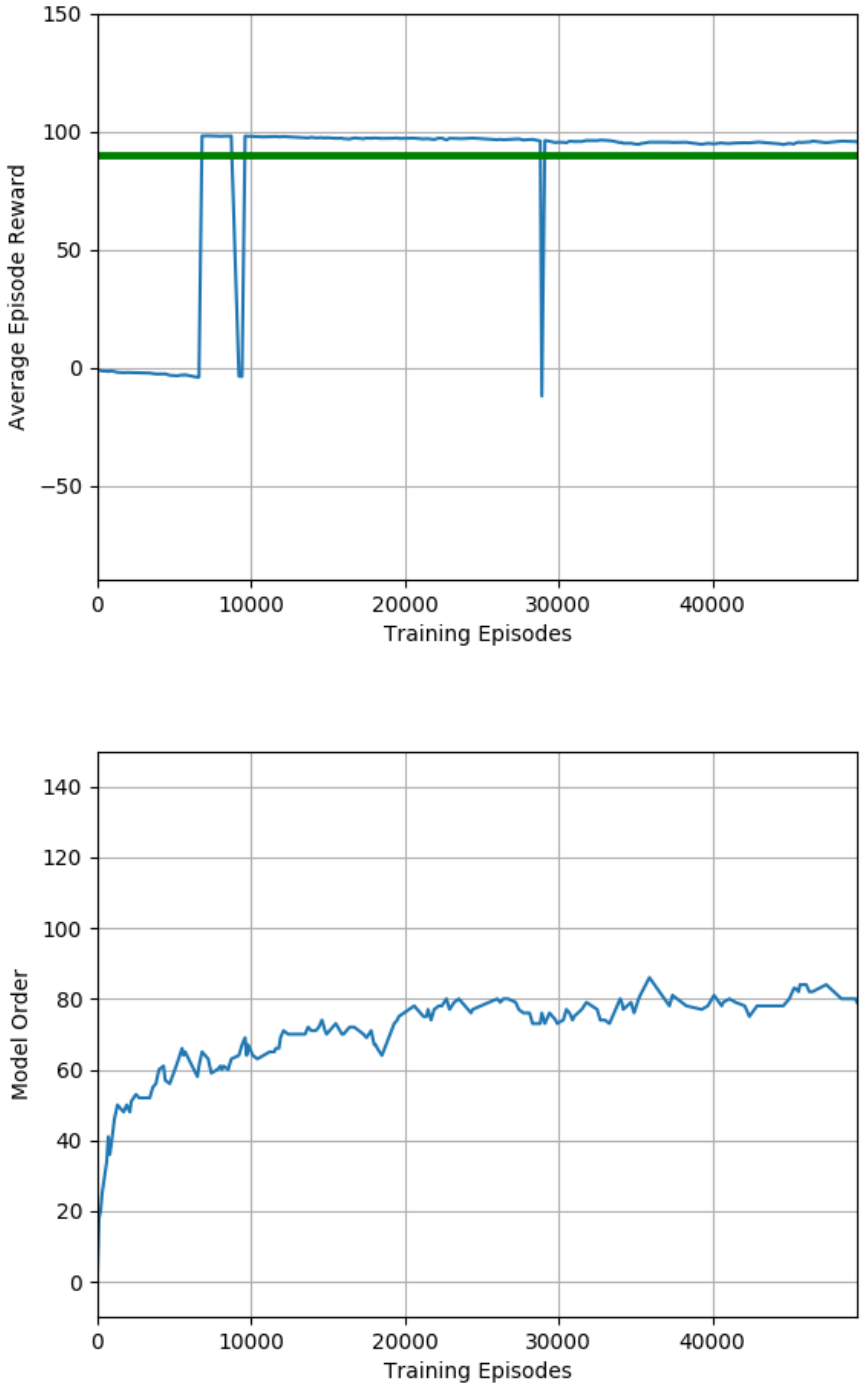


Figure 6.2: Result of representative run of Algorithm 5 over 50,000 Continuous Mountain Car episodes. The top figure shows the average reward obtained by the policy –showed in Figure 6.3– after each training step (episode). An average reward over 90 (green) indicates that we have solved the problem, reaching the goal location. The bottom figure shows the model complexity (number of Dictionary elements) during training remains bounded.

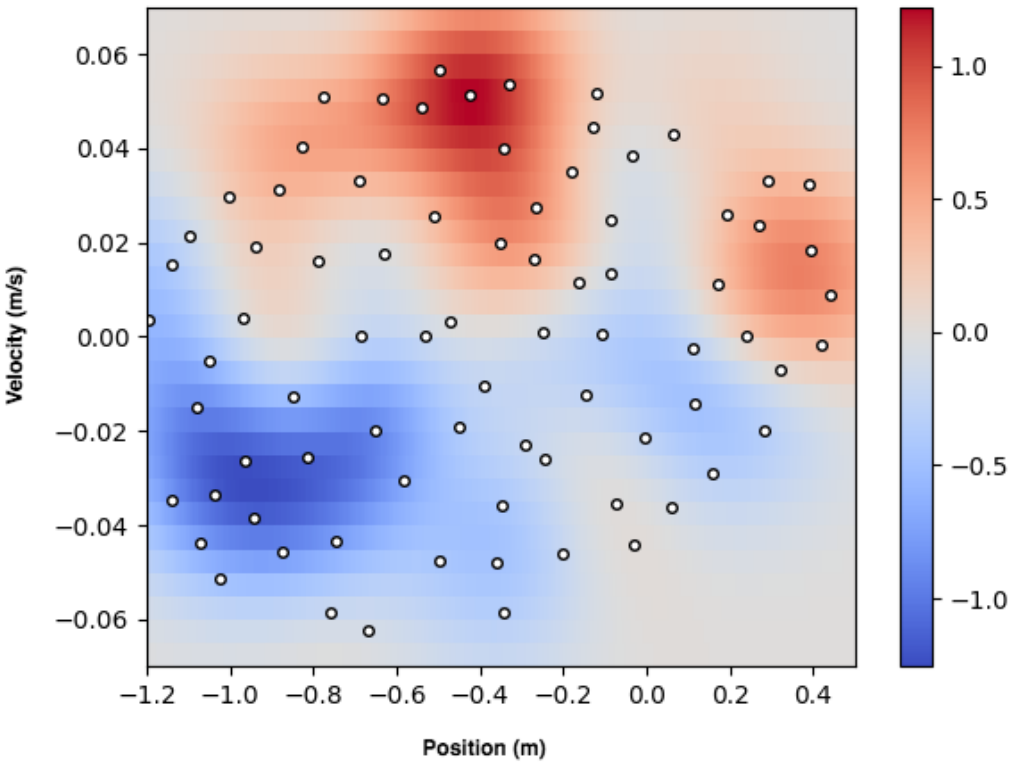


Figure 6.3: Learned policy for Continuous Mountain Car after 50,000 episodes.

6.8 Conclusion

We have considered the problem of learning a policy that belongs to a RKHS in order to maximize the functional defined by the expected discounted cumulative reward that an agent receives. In particular, we presented an algorithm that allows to obtain an unbiased estimate of the gradient of the functional with respect to the policy. By running stochastic gradient ascent in the RKHS we were able to show convergence of the algorithm to a critical point of the functional. This algorithm, of theoretical interest, is not practical since the number of kernel elements that requires grows unbounded. To overcome this limitation, we combined the previous algorithm with destructive Kernel Orthogonal Matching Pursuit to ensure that the model order remains bounded. The later comes at the price of loosing accuracy in the solution and thus, the convergence is to a neighborhood of the critical points. We tested this algorithm in the mountain car problem and its online version in a navigation problem in an environment with obstacles.

Chapter 7

Future Work

The taxonomy presented in Chapter 1 takes into account three mayor characteristics of a minimum definition of autonomy. These are related to the complexity in which the agent operates, the type of information available about the environment and whether the agent is myopic or farsighted. In Chapters 2 through 6 we provide solutions for different situations in which different levels of complexity in each of these characteristics where present. The philosophy of all these solutions is to always use greedy controllers, which due to their simplicity do not require the involvement of logic. The reason for doing so, is to reduce at the minimum the logic required in a complex system to perform relatively simple tasks, so it can be fully devoted to the high level thinking and reasoning. While we provided solutions for some of those problems, there are scenarios that still need to be addressed. We briefly describe these and its possible solutions in what follows.

7.1 Saddle Point algorithms in punctured spaces

The solutions of unconstrained optimization problems in punctured spaces developed for complete and deterministic information in Chapter 2 and for local and stochastic information in Chapter 3 suggest that such approaches could be combined with Saddle Points algorithms as the one described in Chapter 4. Such approach would allow the agent to achieve sublinear regret and fit in a space with obstacles. In the case where the information about the constraints is either deterministic or stochastic, if we were given the optimal Lagrange multiplier for the problem λ^*, we could find the solution of the system by minimizing the Lagrangian evaluated at λ^*. This is

$$x^* = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \mathcal{L}(x, \lambda^*). \tag{7.1}$$

Because, the Lagrangian is a convex function in x, the previous problem is not different than the problems studied in Chapters 2 and 3. And it can be solved by descending along the negative gradient of the Lagrangian with respect to x. It is clear on the other hand, that having the value of λ^* is not a realistic scenario in most applications, where the constraints are measured in operation time. In that sense a posibility – similar to the classic Saddle Point algorithm– would be to run gradient descent along a potential of the Koditscheck-Rimon form, where the attractive potential is now $\mathcal{L}(x, \lambda)$. The multipliers, in turn could be updated by runing gradient ascent with respect to the dual variables, yielding an update that is proportional to the constraint violation.

7.2 Reinforcement Learning with constraints

Similarly to the problem with punctured spaces, in the current formulation of non-myopic agents discussed in Chapter 6 we do not take into account constraints that need to be satisfied but just one functional that needs to be maximized. The hope in this case is that under some problem restrictions, we could be able to generalize the saddle point algorithm to find a policy that is able to satisfy a set of constraints, in the same sense that the stochastic policy gradient ascent discussed in Chapter 6 generalizes stochastic gradient ascent. In this setting the agent would be faced with a set of $m+1$ rewards $r_i(s, a)$ with $i = 0 \dots m$ that represent each one of the constraints to be satsified in the long run and the objective function. Defining

$$U_i(h) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t)\right], \tag{7.2}$$

where the expectations are with respect to the probability distribution of the trajectories of the system. The optimization problem would be therefore, to find a policy $h \in \mathcal{H}$ such that

$$\begin{aligned} h^* := \text{argmax } & U_0(h) \\ \text{s.t.} \quad & U_i(h) \geq 0 \quad \text{for all} \quad i = 1 \dots m \end{aligned} \tag{7.3}$$

As in the case of the parametric optimization, we could think about constructing a Lagrangian for the previous optimization problem, by defininining a set of multipliers $\lambda \in \mathbb{R}_+^m$ and weighting each constraint by its multiplier

$$\mathcal{L}(h, \lambda) = U_0(h) + \sum_{i=1}^{m} \lambda_i U_i(h). \tag{7.4}$$

Then, a possible solution to the problem (7.3) would be to update the policy by running gradient ascent

$$h_{k+1} = h_k + \eta_k \nabla_h \mathcal{L}(h_k, \lambda_k, \cdot), \tag{7.5}$$

and the weights of the multipliers by running gradient descent

$$\lambda^i_{k+1} = \left[\lambda^i_k - \eta_k U_i(h_k, \lambda_k)\right]^+. \tag{7.6}$$

In the previous setting the problems of of the form (7.3) are not necessarily convex, so the aim would be to get local convergence results.

Appendix A

Appendix

A.1 Proofs of the results in Chapter 2

A.1.1 Proof of Lemma 3

Since φ_k is twice continuously differentiable and its maximum is attained in the boundary of the compact set $\mathcal{F}$ (cf., Lemma 1) it must be the case that there exists $x_c \in \text{int}(\mathcal{F})$ such that $\nabla\varphi_k(x_c) = 0$. In Lemma 1 it was argued that for all $x \in \mathcal{F}$ it holds that $f_0^k(x) + \beta(x) > 0$. Hence $\nabla\varphi_k(x_c) = 0$ (cf., (A.44)) if and only if

$$k\beta(x_c)\nabla f_0(x_c) = f_0(x_c)\nabla\beta(x_c) \tag{A.1}$$

In cases where $\nabla\beta(x^*) = 0$ or $f_0(x^*) = 0$ then the previous equation is satisfied for $x_c = x^*$ and x^* is a critical point. By virtue of Lemma 2 there are not critical points in the boundary of the free space, hence the left hand size of the above equation is not zero for any $x_c \neq x^*$. Since $x^* \in \text{int}(\mathcal{F})$ (see Assumption 2) there exists $\delta_0 > 0$ such that for any $\delta \in (0, \delta_0]$ we have

$$\big\{x \in \mathcal{F}\big|\beta(x) < \delta\big\} \cap \big\{x \in \mathcal{F}\big|\|\nabla f_0(x)\| < \delta\big\} = \emptyset \tag{A.2}$$

Since f_0 is non negative and both f_0, β are twice continuously differentiable (see Assumption 2) and $\mathcal{F}$ is a compact set, there exists $C > 0$ such that $f_0(x)\|\nabla\beta(x)\| < C$ for all $x \in \mathcal{F}$. Hence, from (A.1) we have that for any $\delta_1 \in (0, \delta_0]$ there exists $K_1 > 0$ such that if $k > K_1$ then

$$\beta(x_c)\|\nabla f_0(x_c)\| < \delta_1^2. \tag{A.3}$$

By construction both $\beta(x_c)$ and $\|\nabla f_0(x_c)\|$ cannot be smaller than δ_1 and if they are both larger than δ_1 then the above inequality is violated. Hence, either $\beta(x_c) < \delta_1$ or $\|\nabla f_0(x_c)\| < \delta_1$. Moreover, using the same argument for the individual functions $\beta_i(x)$, since the obstacles do not intersect (cf., Assumption 1) there exists $\varepsilon_0' > 0$ such that for any $\varepsilon \in (0, \varepsilon_0']$ there

exists $K_0\prime(\varepsilon) > 0$ such that if $k > K_0\prime(\varepsilon)$ then x_c is such that either $\|\nabla f_0(x_c)\| < \varepsilon$ or for exactly one i we have that $\beta_i(x_c) < \varepsilon$. We next show that the critical points cannot be pushed towards the external boundary of the free space. Assume that for all $\varepsilon \in (0, \varepsilon_0']$ there exists $K_0\prime(\varepsilon)$ such that for all $k > K_0\prime(\varepsilon)$ there is a critical point x_c satisfying $\beta_0(x_c) < \varepsilon$. Let us write the gradient of $\nabla\beta(x_c)$ as in (2.26)

$$\nabla\beta(x_c) = \bar{\beta}_0(x_c)\nabla\beta_0(x_c) + \beta_0(x_c)\nabla\bar{\beta}_0(x_c) \tag{A.4}$$

Since the workspace is a convex set, it is the super level set of a concave function (cf., (2.6)). Thus it holds that $\nabla\beta_0(x_s)^\top(x_s - x^*) < 0$. Since $\nabla\bar{\beta}_0$ is continuous (cf. Assumption 1), over the compact set $\mathcal{F}$ it is bounded. Then, choose $\varepsilon_0 < \varepsilon_0\prime$ such that $\nabla\beta(x_s)^\top(x_s - x^*) < 0$. It follows from (A.1) that at a critical point $\nabla\beta(x_s)$ and $\nabla f_0(x_s)$ point in the same direction and therefore there exists $K_0(\varepsilon_0) > 0$ such that if $k > K_0(\varepsilon_0)$ then $\nabla f_0(x_s)^\top(x_s - x^*) < 0$. The latter however contradicts the first order condition of convexity (see e.g. [16]). Hence, for any $\varepsilon < \varepsilon_0$ there exists $K_0(\varepsilon) > 0$ such that if $k > K_0(\varepsilon)$ for any critical point we have that $\beta_0(x_c) > \varepsilon_0$.

A.1.2 Proof of Lemma 4

Let x_s be a critical point such that $\beta_i(x_s) < \varepsilon_0$ for some $i = 1 \dots m$ where ε_0 is that of Lemma 3 and let v be a unit vector normal to $\nabla\beta(x_s)$. If we prove that $v^\top\nabla^2\varphi_k(x_s)v < 0$ for some direction v, then x_s is not a local minimum. Differentiating (A.44) and using the fact that for a critical point (A.1) holds, we can write

$$\begin{aligned}\nabla^2\varphi_k(x_s) = \Big(f_0^k(x_s) + \beta(x_s)\Big)^{-1-\frac{1}{k}} \Bigg(&\beta(x_s)\nabla^2 f_0(x_s) + (1 - \frac{1}{k})\nabla f_0(x_s)\nabla\beta(x_s)^\top \\ &- \frac{f_0(x_s)\nabla^2\beta(x_s)}{k}\Bigg).\end{aligned} \tag{A.5}$$

In Lemma 1 we argued that $\forall x \in \mathcal{F}$ it holds that $f_0^k(x) + \beta(x) > 0$. Thus, along a direction v satisfying $v^\top\nabla\beta(x_s) = 0$, we have that $v^\top\nabla^2\varphi_k(x_s)v < 0$ if and only if

$$k\beta(x_s)v^\top\nabla^2 f_0(x_s)v - f_0(x_s)v^\top\nabla^2\beta(x_s)v < 0. \tag{A.6}$$

Since $x^* := \operatorname{argmin} f_0(x)$, then $\nabla f_0(x^*) = 0$ and we can use (2.12) to lower bound $\nabla f_0(x_s)^\top(x_s - x^*)$ as

$$\lambda_{\min}\|x_s - x^*\|^2 \leq \nabla f_0(x_s)^\top(x_s - x^*). \tag{A.7}$$

Since x_s is a critical point (A.1) holds. Multiply both sides of the equation by $(x_s - x^*)$ to write

$$k\beta(x_s)\nabla f_0(x_s)^\top(x_s - x^*) = f_0(x_s)\nabla\beta(x_s)^\top(x_s - x^*). \tag{A.8}$$

From Lemma 3 we have that $\|\nabla f_0(x_s)\| > \varepsilon_0$ which is independent of k, hence $\|x_s - x^*\|$ is bounded away from zero by a constant independent of k. Therefore we can upper bound $k\beta(x_s)$ by

$$k\beta(x_s) \leq f_0(x_s)\frac{\nabla\beta(x_s)^\top(x_s - x^*)}{\lambda_{\min}\|x_s - x^*\|^2}. \tag{A.9}$$

Substituing $\nabla\beta(x_s)$ in (A.9) by its expression in (2.26) yields

$$\begin{aligned} k\beta(x_s) \leq & \frac{f_0(x_s)}{\lambda_{\min}\|x_s - x^*\|^2}\bar{\beta}_i(x_s)\nabla\beta_i(x_s)^\top(x_s - x^*) \\ & + \frac{f_0(x_s)}{\lambda_{\min}\|x_s - x^*\|^2}\beta_i(x_s)\nabla\bar{\beta}_i(x_s)^\top(x_s - x^*). \end{aligned} \tag{A.10}$$

We argue next that the second term of (A.10) is bounded by a constant. As argued in the previous paragraph $\|x_s - x^*\|$ is bounded away from zero by a constant independent of k. In addition the remaining factors are the product of continuous functions in a bounded set, thus they are uniformly bounded as well. Let $B > 0$ be a constant bounding the terms multiplying $\beta_i(x_s)$ in the second term of (A.10), i.e,

$$\frac{f_0(x_s)}{\lambda_{\min}\|x_s - x^*\|^2}\nabla\bar{\beta}_i(x_s)^\top(x_s - x^*) \leq B. \tag{A.11}$$

Now, let us focus on the second term of (A.6), in particular the Hessian of $\beta(x_s)$ can be computed by differentiating (2.26)

$$\nabla^2\beta(x_s) = \beta_i(x_s)\nabla^2\bar{\beta}_i(x_s) + \bar{\beta}_i(x_s)\nabla^2\beta_i(x_s) + 2\nabla\beta_i(x_s)\nabla^\top\bar{\beta}_i(x_s).$$

It follows from the result of Lemma 3 and the non negativity of the objective function (cf., Assumption 2) that both $f_0(x_s)$ and $\bar{\beta}_i(x_s)$ are bounded away form zero. Then, combine (2.26) and (A.1) to express the gradient of $\nabla\beta_i(x_s)$ as

$$\nabla\beta_i(x_s) = k\beta_i(x_s)\frac{\nabla f_0(x_s)}{f_0(x_s)} - \beta_i(x_s)\frac{\nabla\bar{\beta}_i(x_s)}{\bar{\beta}_i(x_s)}. \tag{A.12}$$

Recall from (A.1) that at a critical point $\nabla\beta(x_s)$ and $\nabla f_0(x_s)$ are collinear, thus $v^\top\nabla f_0(x_s) = 0$ since v is perpendicular to $\nabla\beta(x_s)$. Hence

$$v^\top\nabla\beta_i(x_s) = -\beta_i(x_s)v^\top\frac{\nabla\bar{\beta}_i(x_s)}{\bar{\beta}_i(x_s)}. \tag{A.13}$$

Combine (A.12) and (A.13) to evaluate the quadratic form associated with the Hessian of $\beta(x_s)$ along the direction v

$$v^\top \nabla^2 \beta(x_s) v = v^\top \nabla^2 \beta_i(x_s) v \bar{\beta}_i(x_s) + \beta_i(x_s) \left(v^\top \nabla^2 \bar{\beta}_i(x_s) v - 2 \frac{\|v^\top \nabla \bar{\beta}_i(x_s)\|^2}{\bar{\beta}_i(x_s)} \right). \quad \text{(A.14)}$$

In the above equation the absolute value of the function multiplying $\beta_i(x_s)$ is upper bounded by a constant independent of k. Let $B' > 0$ be this constant. Then, the second term of (A.6) is upper bounded by

$$-f_0(x_s) v^\top \nabla^2 \beta(x_s) v \leq -v^\top \nabla^2 \beta_i(x_s) v \bar{\beta}_i(x_s) f_0(x_s) + \beta_i(x_s) B'. \quad \text{(A.15)}$$

Use the bounds (A.10), (A.11) and (A.15) and the fact tht $v^\top \nabla f_0(x_s) v \leq \lambda_{\max}$ to bound the left hand side of (A.6) by

$$\begin{aligned} v^\top \left(k\beta(x_s) \nabla^2 f_0(x_s) - f_0(x_s) \nabla^2 \beta(x_s) \right) v \leq v^\top \nabla^2 f_0(x_s) v \frac{f_0(x_s) \bar{\beta}_i(x_s)}{\lambda_{\min} \|x_s - x^*\|^2} \nabla \beta_i(x_s)^\top (x_s - x^*) \\ - v^\top \nabla^2 \beta_i(x_s) v f_0(x_s) \bar{\beta}_i(x_s) + \beta_i(x_s) \left(B \lambda_{\max} + B' \right). \end{aligned} \quad \text{(A.16)}$$

As argued previously $\beta_j(x_s)$ is bounded away from zero by a constant independent of k for all $j \neq i$. The same holds for $f_0(x_s)$. Then, we have that $v^\top \nabla^2 \varphi_k(x_s) v < 0$ if

$$v^\top \nabla^2 f_0(x_s) v \frac{\nabla \beta_i(x_s)^\top (x_s - x^*)}{\lambda_{\min} \|x_s - x^*\|^2} - v^\top \nabla^2 \beta_i(x_s) v \leq -\beta_i(x_s) B'', \quad \text{(A.17)}$$

where $B'' > 0$ is a bound for $(B\lambda_{\max} + B')/(\bar{\beta}_i(x_s) f_0(x_s))$. From Assumption 2 we have that $v^\top \nabla^2 f_0(x_s) v \leq \lambda_{\max}$ and $v^\top \nabla^2 \beta_i(x_s) v \geq \mu^i_{\min}$, then $v^\top \nabla^2 \varphi(x_s) v < 0$ if

$$\frac{\lambda_{\max}}{\lambda_{\min}} \frac{\nabla \beta_i(x_s)^\top (x_s - x^*)}{\|x_s - x^*\|^2} - \mu^i_{\min} \leq -\beta_i(x_s) B''. \quad \text{(A.18)}$$

By hypothesis the left hand side of the above equation is strictly negative in the boundary of the obstacle, and the right hand side takes the value zero. Therefore there exists $\varepsilon_1 > 0$ such that for any $\varepsilon \in (0, \varepsilon_1]$ if $\beta_i(x_s) < \varepsilon$ the above inequality is satisfied. Thus, from the result in Lemma 2 there exists some $K_1(\varepsilon) > K_0(\varepsilon)$ such that for any $k > K_1(\varepsilon)$ the critical point is not a minimum.

A.1.3 Proof of Lemma 5

Since $\varphi_k(x)$ is a twice continuously differentiable function and it attains its maximum at the boundary of a compact set (see Lemma 1) it must have a minimum in the interior of

$\mathcal{F}$. In virtue of Lemma 4 for any $\varepsilon < \varepsilon_1$ there exists $K_1(\varepsilon) > 0$ such that if $k > K_1(\varepsilon)$ the critical points x_c such that $\beta_i(x_c) < \varepsilon$ are not local minima. Hence the minimum for $\varphi_k(x)$ is such that $\|\nabla f_0(x_c)\| < \varepsilon$. We next show that any critical point satisfying $\|\nabla f_0(x_c)\| < \varepsilon$ is a nondegenerate minimum. Using the same arguments as in Lemma 4 we have that $\nabla^2 \varphi_k(x_c) \succ 0$ if and only if

$$\beta(x_c)\nabla^2 f_0(x_c) + (1 - \frac{1}{k})\nabla\beta(x_c)\nabla f_0^\top(x_c) - \frac{f_0(x_c)}{k}\nabla^2\beta(x_c) \succ 0. \tag{A.19}$$

Since $\|\nabla f_0(x_c)\| < \varepsilon < \varepsilon_0$ it follows from Lemma 3 that each $\beta_i(x_c) > \varepsilon_0$ and therefore $\beta(x_c) > \varepsilon_0^{m+1}$. Hence the first term in the previous equation satisfies

$$\beta(x_c)\nabla^2 f_0(x_c) \succ \lambda_{\min}\varepsilon_0^{m+1} I \succ 0. \tag{A.20}$$

From (A.1) it follows that $\nabla f_0(x_c)$ and $\nabla\beta(x_c)$ point in the same direction, thus the second term in (A.19) is a positive semi definite matrix for any $k > 1$. Therefore for $\nabla^2\varphi_k(x_c)$ to be positive definite it suffices that

$$\frac{f_0(x_c)}{k}\nabla^2\beta(x_c) \prec \lambda_{\min}\varepsilon_0^{m+1} I. \tag{A.21}$$

Since f_0 and β are twice continuously differentiable (see Assumption 2) $f_0(x_c)\nabla^2\beta(x_c)$ is bounded by a constant independent of k because the free space is compact. Therefore there exists $K_2\prime(\varepsilon_0) > 1$ such that if $k > K_2\prime(\varepsilon_0)$, the above equation holds and therefore any critical point satisfying $\|\nabla f_0(x_c)\| < \varepsilon$ is a minimum. We are left to show that the minimum is unique. Let $c > f_0(x^*)$ be such that for any $x \in \mathcal{F}$ if $f_0(x) < c$, then $\|\nabla f_0(x)\| < \varepsilon_0$ and define the set $\Omega_c = \{x \in \mathcal{F} \big| f_0(x) < c\}$. By definition of the previous set and because the previous discussion all critical points in Ω_c are minima. We show next that for large enough k, Ω_c is positively invariant for the flow $\dot{x} = -\nabla\varphi_k(x)$. Compute the derivative of $f_0(x)$ along the trajectories of the flow and evaluate at the boundary of Ω_c

$$\dot{f}_0(x) = -\nabla f_0(x)^\top \nabla\varphi_k(x). \tag{A.22}$$

The previous inner product is negative if and only if

$$\beta(x)\|\nabla f_0(x)\|^2 - \nabla f_0(x)^\top \nabla\beta(x)\frac{f_0(x)}{k} > 0. \tag{A.23}$$

Observe that first term in the above equation is lower bounded by a constant independent of k in $\partial\Omega_c$ since $c > f_0(x^*)$ and $\beta_i(x) > \varepsilon_0$. Moreover since β and f_0 are twice continuously differentiable the second term in the previous equation is lower bounded by $-C/k$, where C is independent of k. Therefore there exists $K_2\prime\prime(\varepsilon_0) > 1$ such that if $k > K_2\prime\prime(\varepsilon_0)$,

then Ω_c is positively invariant, hence the limit set of the flow $\dot{x} = -\nabla\varphi_k(x)$ restricted to Ω_c converges to a local minimum. If there were more than one degenerate minimum in Ω_c, since the stable manifold of minimums are open sets, then it would be possible to write $\partial\Omega_c$ as a disjoint union of open sets – in the topology relative to the boundary of Ω_c. This contradicts the connexity of the boundary. Hence, for any $\varepsilon > 0$ there exists $K_2(\varepsilon) = \max\{K_1(\varepsilon), K_2\prime(\varepsilon), K_2\prime\prime(\varepsilon)\}$ such that if $k > K_2(\varepsilon)$, φ_k is polar with minimum at $\bar{x}$, where $\|\bar{x} - x^*\| < \varepsilon$. Finally from the discussion in Lemma 2 we have that $\bar{x} = x^*$ if $f_0(x^*) = 0$ or $\nabla\beta(x^*) = 0$.

A.1.4 Proof of Theorem 3

In the particular case where the functions β_i take the form (2.21), condition (2.18) of Theorem 2 yields

$$\frac{\lambda_{\max}}{\lambda_{\min}} \frac{(x_s - x_i)^\top A_i (x_s - x^*)}{\|x_s - x^*\|^2} - \mu^i_{\min} < 0. \tag{A.24}$$

Since A_i is positive definite, there exists $A_i^{1/2}$ such that

$$A_i = \left(A_i^{1/2}\right)^\top A_i^{1/2}. \tag{A.25}$$

Consider the change of variables $z = A_i^{1/2} x$, and write

$$\frac{(x_s - x_i)^\top A_i (x_s - x^*)}{\|x_s - x^*\|^2} = \frac{(z_s - z_i)^\top (z_s - z^*)}{\|A_i^{-1/2}(z_s - z^*)\|^2}. \tag{A.26}$$

Denote by $\mu^i_{\max}$ the maximum eigenvalue of the matrix A_i

$$\frac{1}{\mu^i_{\max}} \|(z_s - z^*)\|^2 \leq \|A_i^{-1/2}(z_s - z^*)\|^2. \tag{A.27}$$

Use the above inequality to bound the left hand side of (A.24)

$$\frac{\lambda_{\max}}{\lambda_{\min}} \frac{(x_s - x_i)^\top A_i (x_s - x^*)}{\|x_s - x^*\|^2} - \mu^i_{\min} \leq \frac{\lambda_{\max}}{\lambda_{\min}} \frac{(z_s - z_i)^\top (z_s - z^*)}{\|z_s - z^*\|^2} \mu^i_{\max} - \mu^i_{\min}. \tag{A.28}$$

The change of coordinates transforms the elliptical obstacle into a sphere of radius $r_i(\mu^i_{\min})^{1/2}$ since the function β_i takes the following form for the variable z

$$\beta_i(z) = \|z - z_i\|^2 - r_i^2 \mu^i_{\min}. \tag{A.29}$$

Since the obstacle is, after considering the change of coordinate, a sphere we define for convenience the radial direction $\hat{e}_r$, with $\|\hat{e}_r\| = 1$. Let θ be the angle between $\hat{e}_r$ and the

direction $z_i - z^*$. Further define $\tilde{r}$ to be the distance between the critical point z_s and z_i. Notice that if $|\theta| \leq \pi/2$ then

$$\frac{(x_s - x_i)^\top (x_s - x^*)}{\|x_s - x^*\|^2} \leq 0, \tag{A.30}$$

and in that case the right hand side of (A.28) is negative which completes the proof of the lemma. However if $|\theta| > \pi/2$ then the term under consideration is positive. In particular the larger the norm of $\tilde{r}$ the larger the value. Hence define $\tilde{r}_{\max} = r_i(\mu^i_{\min})^{1/2} + \varepsilon$, and the following bound holds

$$\frac{(z_s - z_i)^\top (z_s - z^*)}{\|z_s - z^*\|^2} \leq \frac{\tilde{r}_{\max}(\tilde{r}_{\max} - d_i \cos\theta)}{\tilde{d_i}^2 + \tilde{r}^2_{\max} - 2\tilde{d}_i \tilde{r}_{\max} \cos\theta}, \tag{A.31}$$

where $\tilde{d}_i$ is the distance between z_s and z^*. Differentiating the right hand side of the above equation with respect to θ we conclude that its critical points are multiples of π. Notice that for multiples of π of the form $2k\pi$, with $k \in \mathbb{Z}$ will correspond to negative values and and for multiples of π of the form $(2k+1)\pi$ with $k \in \mathbb{Z}$, we have that

$$RHS(2k\pi + \pi) = \frac{\tilde{r}_{\max}(\tilde{r}_{\max} + \tilde{d}_i)}{\left(\tilde{d}_i + \tilde{r}_{\max}\right)^2} = \frac{\tilde{r}_{\max}}{\tilde{d}_i + \tilde{r}_{\max}} \tag{A.32}$$

Combine the previous bound with (A.28) to upper bound (A.24)

$$\frac{\lambda_{\max}}{\lambda_{\min}} \frac{(x_s - x_i)^\top A_i (x_s - x^*)}{\|x_s - x^*\|^2} \mu^i_{\max} - \mu^i_{\min} \leq \frac{\lambda_{\max}}{\lambda_{\min}} \frac{\tilde{r}_{\max}}{\tilde{d}_i + \tilde{r}_{\max}} \mu^i_{\max} - \mu^i_{\min}. \tag{A.33}$$

Notice that a lower bound for that distance is given by $\tilde{d}_i \geq \mu^i_{\min} d_i$. Aince z_s can be placed arbitrarily close to the boundary of the obstacle $\mathcal{O}_i$ we have that $\tilde{r} \leq r_i(\mu^i_{\min})^{1/2} + \varepsilon$. To complete the proof observe that

$$\frac{\tilde{r}_{\max}}{\tilde{d}_i + \tilde{r}_{\max}} = \frac{r_i + \frac{\varepsilon}{\mu^i_{\min}}}{d_i + r_i + \frac{\varepsilon}{\mu^i_{\min}}}, \tag{A.34}$$

hence since ε can be made arbitrarily small by increasing k we have that (A.24) holds if

$$\frac{\lambda_{\max}}{\lambda_{\min}} \frac{\mu^i_{\max}}{\mu^i_{\min}} < 1 + \frac{d_i}{r_i}. \tag{A.35}$$

Thus condition (2.18) takes the form stated in the theorem.

A.1.5 Proof of Theorem 4

Let us consider the evolution of the dynamical system (2.34) from some time $t_0 > 0$. Notice that if (2.18) holds, then in virtue of Theorem 2 for large enough k the function $\varphi_{k,\mathcal{A}_c(t_0)}(x)$ defined in (2.33) is a navigation function for the set $\mathcal{F}_{\mathcal{A}_c(t_0)} = \mathcal{X} \setminus \bigcup_i \mathcal{O}_{i\in\mathcal{A}_c(t_0)}$. On one hand, this ensures the avoidance of the obstacles $\mathcal{O}_i$ with $i \in \mathcal{A}_c(t_0)$, furthermore it ensures convergence to x^* – or to a point arbitrarily close to x^*– unless a new obstacle is visited. If the first happens the proof is completed. In the second case, we need to show that the time lapsed until the agent reaches the neighborhood of a new obstacle is finite. This being the case it would take a finite time $T \geq 0$ to visit all obstacles before having $\varphi_{k,\mathcal{A}_c(t)}(x) = \varphi_k(x)$ for all $x \in \mathcal{F}$. Then for any $t \geq T$ we are in the situation where the obstacles are known and Theorem 2 holds, which completes the proof. Let t_f be the first instant in which the agent reaches the c-neighborhood of an obstacle of which he is not aware. Formally, this is

$$t_f = \min\left\{t > t_0 \middle| \beta_j(x(t)) \leq c \quad \text{for some} \quad j \notin \mathcal{A}_c(t_0)\right\}. \tag{A.36}$$

Notice that by the definition of the time t_f we have that $\mathcal{A}_c(t) = \mathcal{A}_c(t_0)$ for all $t \in [t_0, t_f)$. And therefore $\varphi_{k,\mathcal{A}_c(t)}(x) = \varphi_{k,\mathcal{A}_c(t_0)}(x)$ is a navigation function for the free space $\mathcal{F}_{\mathcal{A}_c(t_0)} = \mathcal{X} \setminus \bigcup_{i\in\mathcal{A}_c(t)}$ for all $t \in [t_0, t_f)$. Therefore the critical points of the function (2.33) are arbitrarily close to x^* or arbitrarily close to the obstacles $\mathcal{O}_i$ with $i \in \mathcal{A}_c(t_0)$ (cf., lemma 3). Thus the norm of the gradient of the partial navigation function is bounded below for any $x(t)$ with $t \in [t_0, t_f)$ for a set of initial conditions of measure one. Hence, there exists a constant $L > 0$ such that

$$\left\|\nabla\varphi_{k,\mathcal{A}_c(t_0)}(x(t))\right\| \geq L, \forall t \in [t_0, t_f). \tag{A.37}$$

From the fundamental theorem of calculus we can write

$$\varphi_{k,\mathcal{A}_c(t_0)}(x(t_f)) - \varphi_{k,\mathcal{A}_c(t_0)}(x(t_0)) = \int_{t_0}^{t_f} \dot{\varphi}_{k,\mathcal{A}_c(t_0)}(x(s))ds. \tag{A.38}$$

Write the right hand side of the above equation as

$$\int_{t_0}^{t_f} \dot{\varphi}_{k,\mathcal{A}_c(s)}(x(s))ds = \int_{t_0}^{t_f} \nabla\varphi^\top_{k,\mathcal{A}_c(t_0)}(x(s))\dot{x}ds \tag{A.39}$$

and substitute $\dot{x}$ by the expression in (2.34)

$$\int_{t_0}^{t_f} \dot{\varphi}_{k,\mathcal{A}_c(s)}(x(s))ds = -\int_{t_0}^{t_f} \left\|\nabla\varphi_{k,\mathcal{A}_c(t_0)}(x(s))\right\|^2 ds. \tag{A.40}$$

Finally combine the above expression with (A.38) and the bound in (A.37) to write

$$\varphi_{k,\mathcal{A}_c(t_f)}(x(t_f)) - \varphi_{k,\mathcal{A}_c(t_0)}(x(t_0)) \leq \int_{t_0}^{t_f} L^2 ds. \tag{A.41}$$

By integrating the right hand side of the above expression we get the following upper bound for t_f

$$t_f \leq t_0 + \frac{\varphi_{k,\mathcal{A}_c(t_0)}(x(t_0)) - \varphi_{k,\mathcal{A}_c(t_0)}(x(t_f))}{L^2}. \tag{A.42}$$

Since the navigation function is always bounded (cf., Definition 1) the time until the agent visits a new obstacle if finite, which completes the proof of the theorem.

A.2 Proofs of the results in Chapter 3

A.2.1 An estimator of the navigation function

In this section we analize a possible estimate of the gradient of a Rimon-Koditschek navigation function based on local and stochastic observations of the objective function and the obstacles that motivates Assumptions 3 and 4. The estimate proposed is based on the fact that the gradient of the potential defined in (2.17) is collinear to

$$\beta(x)\nabla f_0(x) - \frac{f_0(x)\nabla\beta(x)}{k}. \tag{A.43}$$

Indeed, by differentiating (2.17) one has that (cf., (A.44))

$$\nabla\varphi_k(x) = \left(f_0^k(x) + \beta(x)\right)^{-1-\frac{1}{k}} \left(\beta(x)\nabla f_0(x) - \frac{f_0(x)\nabla\beta(x)}{k}\right). \tag{A.44}$$

By virtue of assumptions 1 and 2 and the definition of the function $\beta(x)$ in (2.8) one has that the factor that distinguishes the expressions in (A.43) and (A.44) is strictly positive.

Since the objective function is typically a physical magnitude that must be minimized or maximized one can assume that the robot has estimates of the value of the function $f_0(x)$ and its gradient at the current location. For instance, in the problem of climbing a forested hill, the function $f_0(x)$ represents the height profile of the hill. The value of such function can be estimated with a GPS whereas its gradient – the slope of the hill –can be inferred with an inertial measurement unit (IMU). Denote these estimates at time t by $\hat{f}_0(x_t, \theta_t)$ and $\hat{\nabla} f_0(x_t, \theta_t)$ respectively, where θ_t is a random vector, representing the measurement noise, measurable with respect to the sigma algebra $\mathcal{G}_t$. We assume the estimates to be unbiased, i.e.,

$$\mathbb{E}\left[\hat{f}_0(x_t, \theta_t)\middle|\mathcal{G}_t\right] = f_0(x_t), \quad \mathbb{E}\left[\hat{\nabla} f_0(x_t, \theta_t)\middle|\mathcal{G}_t\right] = \nabla f_0(x_t). \tag{A.45}$$

In order to estimate the obstacles – the trees – the agent may have information available gathered by a rangefinder. Due to physical limitations like the range of the sensor or the fact that obstacles can be "hidden" behind others the agent is not able to sense all the obstacles at a given position x. The set obstacles that can be estimated is composed by those that are at a distance smaller than a given limit $c > 0$

$$\mathcal{A}_c(x) = \left\{ i = 1 \dots m \middle| d_i(x) \leq c \right\}, \tag{A.46}$$

where $d_i(x)$ is the euclidean distance to the i–th obstacle defined as in Assumption 3. Depending on the belief that the agent has about the world, the "obstacle function" will be different. We discuss the case where the obstacle model is spherical [26]. To describe such

obstacles three estimations are needed: distance to the obstacle, direction from the obstacle to the agent and curvature of the obstacle. Denote these quantities for the i–th obstacle by $d_i(x)$, $\mathbf{n}_i(x)$ and $R_i(x)$, and describe the obstacle with the function

$$\tilde{\beta}_i(x) = d_i^2(x) + 2R_i(x)d_i(x), \tag{A.47}$$

and corresponding gradient

$$\tilde{\nabla}\beta_i(x) = 2\left(d_i(x) + R_i(x)\right)\mathbf{n}_i(x). \tag{A.48}$$

Observe that the previous expression is a representation of what the gradient would be if the obstacle were indeed a sphere and it is not the derivative of (A.47). Indeed, notice that if an obstacle is a sphere of center x_i and radius R_i one has that

$$\beta_i(x) = \|x - x_i\|^2 - R_i^2 = (d_i(x) + R_i)^2 - R_i^2 = d_i(x)^2 + 2R_i d_i(x). \tag{A.49}$$

and by differentiating the previous expression we get

$$\nabla\beta_i(x) = 2(x - x_i) = 2(d_i(x) + R_i)\mathbf{n}_i(x). \tag{A.50}$$

Hence, the model of obstacles (A.47)–(A.48) corresponds to spherical obstacles. Denoting the noisy estimates of distance, direction and curvature of the i-th obstacle by $\hat{d}_i(x_t, \theta_t)$, $\hat{\mathbf{n}}_i(x_t, \theta_t)$ and $\hat{R}_i(x_t, \theta_t)$ respectively, a natural estimation for it, is

$$\hat{\beta}_i(x_t, \theta_t) = \hat{d}_i^2(x_t, \theta_t) + 2\hat{R}_i(x_t, \theta_t)\hat{d}_i(x_t, \theta_t), \tag{A.51a}$$

$$\hat{\nabla}\beta_i(x_t, \theta_t) = 2\left(\hat{d}_i(x_t, \theta_t) + \hat{R}_i(x_t, \theta_t)\right)\hat{\mathbf{n}}_i(x_t, \theta_t). \tag{A.51b}$$

Observe that if the estimates of distance, direction and curvature are independent and unbiased we have that

$$\mathbb{E}\left[\hat{\beta}_i(x_t, \theta_t)\middle|\mathcal{G}_t\right] = d_i^2(x_t) + \sigma_{d_i}^2 + 2R_i(x_t)d_i(x_t) = \tilde{\beta}_i(x_t) + \sigma_{d_i}^2(x_t), \tag{A.52}$$

where $\sigma_{d_i}^2(x_t)$ the variance of the estimate of the distance. This variance needs not be constant, but a function of the position since for instance, it could become smaller the closer the robot is to the obstacle. Likewise

$$\mathbb{E}\left[\hat{\nabla\beta}_i(x_t, \theta_t)\middle|\mathcal{G}_t\right] = 2R_i(x_t)d_i(x_t)\mathbf{n}_i(x_t) = \tilde{\nabla}\beta_i(x_t). \tag{A.53}$$

With these estimates and inspired by (A.43) a possible estimate of the direction of the gradient of the navigation function is

$$\hat{g}(x_t,\theta_t) := \hat{\nabla} f_0(x_t,\theta_t) \prod_{i\mathcal{A}_c(x_t)} \hat{\beta}_i(x_t,\theta_t) - \frac{\hat{f}_0(x_t,\theta_t)}{k} \sum_{i\in\mathcal{A}_c(x_t)} \hat{\nabla\beta}_i(x_t,\theta_t) \prod_{j\in\mathcal{A}_c(x_t),j\neq i} \hat{\beta}_j(x_t,\theta_t). \tag{A.54}$$

By taking the expectation of the estimate with respect to $\mathcal{G}_t$ and assuming independence across estimates it is possible to show that the estimate (A.54) satisfies (3.1). Indeed, write

$$\mathbb{E}\left[\hat{g}\middle|\mathcal{G}_t\right] = \mathbb{E}\left[\hat{\nabla} f_0\middle|\mathcal{G}_t\right] \prod_{i\mathcal{A}_c} \mathbb{E}\left[\hat{\beta}_i\middle|\mathcal{G}_t\right] - \frac{\mathbb{E}\left[\hat{f}_0\middle|\mathcal{G}_t\right]}{k} \sum_{i\in\mathcal{A}_c} \mathbb{E}\left[\hat{\nabla\beta}_i\middle|\mathcal{G}_t\right] \prod_{j\in\mathcal{A}_c,j\neq i} \mathbb{E}\left[\hat{\beta}_j\middle|\mathcal{G}_t\right], \tag{A.55}$$

where we dropped the variables x_t and θ_t to simplify the notation. Substituting (A.45), (A.52) and (A.53) in the previous expression yields

$$\begin{aligned}\mathbb{E}\left[\hat{g}(x_t,\theta_t)\middle|\mathcal{G}_t\right] = \nabla f_0(x_t) &\prod_{i\mathcal{A}_c(x_t)} (\tilde{\beta}_i(x_t) + \sigma^2_{d_i}(x_t)) \\ &- \frac{f_0(x_t)}{k} \sum_{i\in\mathcal{A}_c(x_t)} \tilde{\nabla\beta}_i(x_t) \prod_{j\in\mathcal{A}_c(x_t),j\neq i} (\tilde{\beta}_j(x_t) + \sigma^2_{d_j}(x_t)).\end{aligned} \tag{A.56}$$

Let $\alpha : \mathbb{R}^n \to \mathbb{R}_{++}$ be defined as

$$\alpha(x) = \frac{\prod_{i\mathcal{A}_c(x)}(\tilde{\beta}_i(x) + \sigma^2_{d_i}(x))}{\beta(x)} \left(f_0(x)^k + \beta(x)\right)^{1+1/k}, \tag{A.57}$$

Observe that the previous function is continuous at the boundary of the free space if the variance of the distance vanishes fast when approaching it. Then, the discontinuities in $\alpha(x)$ are due to the inclusion or removal of an obstacle from the set $\mathcal{A}_c(x)$. Moreover, $\alpha(x)$ is strictly positive. With this definition, one can write (A.56) as

$$\begin{aligned}\mathbb{E}\left[\hat{g}(x_t,\theta_t)\middle|\mathcal{G}_t\right] = \frac{\alpha(x_t)}{\left(f_0(x_t)^k + \beta(x_t)\right)^{1+1/k}} \Bigg(&\nabla f_0(x_t)\beta(x_t) \\ &- \frac{f_0(x_t)\beta(x_t)}{k} \sum_{i\in\mathcal{A}_c(x_t)} \frac{\tilde{\nabla}\beta_i(x_t)}{\tilde{\beta}_i(x_t) + \sigma^2_{d_i}(x_t)} \Bigg).\end{aligned} \tag{A.58}$$

Adding and substracting $(f_0(x_t)\beta(x_t)/k)\sum_{i=0}^m \nabla\beta_i(x_t)/\beta_i(x_t)$ inside the parenthesis of the

previous expression, substituting (A.44) and defining

$$b_k(x) = \frac{f_0(x)\beta(x)}{k\left(f_0(x)^k + \beta(x)\right)^{1+1/k}} \times \left(\sum_{i=0}^{m} \frac{\nabla \beta_i(x)}{\beta_i(x)} - \sum_{i \in \mathcal{A}_c(x)} \frac{\tilde{\nabla} \beta_i(x)}{\tilde{\beta}_i(x) + \sigma_{d_i}^2(x)} \right), \tag{A.59}$$

yields

$$\mathbb{E}\left[\hat{g}(x_t, \theta_t) \middle| \mathcal{G}_t \right] = \alpha(x_t) \left(\nabla \varphi_k(x_t) + b_k(x_t) \right). \tag{A.60}$$

Which shows that the proposed estimate is of the form (3.1). We next analyze some properties of the estimate $\hat{g}(x_t, \theta_t)$ proposed. These properties inspire the assumptions of navigable estimates in Section 3.2.1. The first one if this properties is that the estimate is bounded. Observe that (A.54) has bounded norm as long as the individual estimates are since the computation only involves products and sums. Further notice, that when an agent is close to the obstacle $\mathcal{O}_i$ we have that $\beta_i(x_t) \approx 0$. Therefore, the direction $\hat{g}(x_t, \theta_t)$ is approximately given by

$$\hat{g}(x_t, \theta_t) \approx -\frac{\hat{f}_0(x_t, \theta_t)}{k} \prod_{j \in \mathcal{A}_c(x_t), j \neq i} \hat{\beta}_j(x_t, \theta) \hat{\nabla} \beta_i(x_t, \theta_t). \tag{A.61}$$

Since the update of the position is in the direction of $-\hat{g}(x_t, \theta_t)$ (cf., (3.10)), the previous expression shows that this update pushes the agent outwards the obstacle nearby. These observations made for this particular estimator correspond to Assumption 3 in Section 3.2.1 for the general case. We next devote our attention to the properties of the bias $b_k(x)$ defined in (A.59). The bias depends on three main factors as we detail next. These do not have an origin in the stochastic nature of the measurements but on the fact that we are making systematic errors in the estimation of the obstacles. The limitation in the number of obstacles that can be measured is one of the factors and it translates in the fact that the two sums in (A.59) are not over the same indices. The second one is the difference between the free space and the belief of the agent, this translates into the fact that in one of the sums in (A.59) we have terms corresponding to the real obstacles, while in the other one we have terms corresponding to the hallucinated obstacles. The closer the belief the agent to the reality the smaller the bias is. The third element is due to non-linearity in the estimation of the obstacles which translates in the presence of the standard deviation in the estimation of the distance to the obstacle. We show in what follows that the difference of the sums in (A.59) is bounded in the free space. Observe that it could be unbounded only at the boundary of the free space, where $\beta(x) = 0$. Let us consider the limit of the difference of the sum when $x \to \partial\mathcal{O}_i$. If the agent approaches the i–th obstacle it means that $\beta_j(x)$ is bounded away from zero for any $j \neq i$ (cf., Assumption 1). Hence, it suffices

to show that the following limit is bounded

$$\lim_{x\to\partial\mathcal{O}_i} \frac{\nabla\beta_i(x)}{\beta_i(x)} - \frac{\tilde{\nabla}\beta_i(x)}{\tilde{\beta}_i(x)+\sigma_{d_i}^2(x)}. \tag{A.62}$$

The previous expression can be re-written as

$$\lim_{x\to\partial\mathcal{O}_i} \frac{(\tilde{\beta}_i(x)+\sigma_{d_i}^2(x))\nabla\beta_i(x) - \beta_i(x)\tilde{\nabla}\beta_i(x)}{\beta_i(x)\left(\tilde{\beta}_i(x)+\sigma_{d_i}^2(x)\right)}. \tag{A.63}$$

Let us write the Taylor's expansion of the function $\beta_i(x)$ and its gradient at the projection $\tilde{x}$ of a point x in the boundary of the obstacle.

$$\beta_i(x) = \nabla\beta_i(z)^\top(x-\tilde{x}) = \nabla\beta_i(z)^\top\mathbf{n}(x)d_i(x), \tag{A.64}$$

where $z \in \{y \in \mathbb{R}^n : y = \lambda x + (1-\lambda)\tilde{x}, \lambda \in [0,1]\}$

$$\nabla\beta_i(x) = \nabla\beta_i(\tilde{x}) + \nabla^2\beta_i(z')(x-\tilde{x}) = \nabla\beta_i(\tilde{x}) + \nabla^2\beta_i(z')\mathbf{n}(x)d_i(x), \tag{A.65}$$

where $z' \in \{y \in \mathbb{R}^n : y = \lambda x + (1-\lambda)\tilde{x}, \lambda \in [0,1]\}$. From (A.64) and (A.47) one can observe that both $\tilde{\beta}_i(x)$ and $\beta_i(x)$ are functions that depend linearly on the distance nearby the obstacles. Therefore, as long as the variance of the estimation vanishes as we approach the obstacle faster than the function $\tilde{\beta}_i$, the denominator in (A.63) is of the order of $O(d_i(x)^2)$. Thus, the limit in (A.63) exists if the numerator is $O(d_i(x)^2)$ as well. We next work towards proving the latter. Using the definition of (A.47) and the expansion of the gradient of $\nabla\beta_i(x)$, the first term in the numerator of (A.63) when $x \to \partial\mathcal{O}_i$, yields

$$\lim_{d_i(x)\to 0} 2R_i(x)d_i(x)\nabla\beta_i(\tilde{x}) + O(d_i(x)^2). \tag{A.66}$$

Likewise the second term in the numerator yields

$$\nabla\beta_i(z)^\top\mathbf{n}(x)d_i(x)\left(2(R_i(x)+d_i(x))\mathbf{n}(x)\right), \tag{A.67}$$

and its limit is

$$\lim_{d_i(x)\to 0} 2R_i(x)d_i(x)\nabla\beta_i(z)^\top\mathbf{n}(x)\mathbf{n}(x) + O(d_i(x)^2). \tag{A.68}$$

Combining this two terms we have that the limit of the numerator can be written as

$$\lim_{d_i(x)\to 0} 2R_i(x)d_i(x)\left(\nabla\beta_i(\tilde{x}) - \nabla\beta_i(z)^\top\mathbf{n}(x)\mathbf{n}(x)\right) + O(d_i^2(x)). \tag{A.69}$$

To complete this proof observe that $\nabla\beta_i(\tilde{x})$ and $\mathbf{n}(x)$ are collinear, hence the first term in the previous expression is zero when multiplied by a perpendicular vector of $\mathbf{n}(x)$. Thus, in that direction the previous limit is a function of order $O(d_i(x)^2)$. Along the direction of $\mathbf{n}(x)$ the difference in the brackets yields $\|\nabla\beta_i(\tilde{x}) - \nabla\beta_i(z)\|$, which goes to zero at least linearly when $x \to \partial\mathcal{O}_i$. Hence we have showed that the numerator of (A.63) is of the order of $d_i(x)^2$ and thus there exists a constant $B' > 0$ such that for all $x \in \mathcal{F}$ it holds that

$$\left\| \sum_{i=0}^{m} \frac{\nabla\beta_i(x)}{\beta_i(x)} - \sum_{i\in\mathcal{A}_c(x)} \frac{\tilde{\nabla}\beta_i(x)}{\tilde{\beta}_i(x) + \sigma_{d_i}^2(x)} \right\| \leq B', \tag{A.70}$$

Since the gradient of $\varphi_k(x)$ has a factor of $1/\left(f_0(x)^k + \beta(x)\right)^{1+1/k}$ it is more convenient to work with the following scaling of the bias

$$\tilde{b}_k(x) = \Big(f_0(x)^k + \beta(x)\Big)^{1+1/k} b_k(x), \tag{A.71}$$

and the following scaling of the gradient of $\varphi_k(x)$

$$\tilde{\nabla}\varphi_k(x) = \Big(f_0(x)^k + \beta(x)\Big)^{1+1/k} \nabla\varphi_k(x), \tag{A.72}$$

A first consequence of the bias being bounded in the free space is that for any $x \in \partial\mathcal{F}$ we have $\tilde{b}_k(x) = b_k(x) = 0$ since $\beta(x) = 0$. Further observe that the norm of $\tilde{b}_k(x)$ is decreasing at the rate $1/k$ for any point in the interior of the free space and in particular $\lim_{k\to\infty} \tilde{b}_k(x) = 0$. As in the case of the function $\alpha(x)$, the function $\tilde{b}_k(x)$ is piece-wise twice differentiable and the discontinuities are due to changes in the set $\mathcal{A}_c(x)$. Therefore, the discontinuities occur away from the obstacles. Furthermore, since $\lim_{k\to\infty} \|\tilde{b}_k(x)\| = 0$ we have that for large enough k the region where $\nabla\varphi_k(x)^\top \left(\nabla\varphi_k(x) + b_k(x)\right) \leq 0$ are disjoint regions around the critical points of $\varphi_k(x)$.

In what follows we argue that near the saddle points of $\varphi_k(x)$ the bias is smaller than $\nabla\varphi_k(x)$ in the C^1 sense. Notice that the saddle points x_c of $\nabla\varphi_k(x)$ satisfy that $\beta(x_c) \leq L/k$ where L is a non-negative constant (see the proof of Lemma 3) and therefore the scaled bias (cf., (A.59) and (A.72)) satisfies $\left\|\tilde{b}_k(x_c)\right\| = O(1/k^2)$. The Jacobian of the bias however is at least of norm $O(1/k)$ and thus the C^1 norm of the bias is defined by the Jacobian. To see why this is the case let us compute the Jacobian of the bias

$$J_{\tilde{b}_k}(x) = \frac{1}{k} D(x) \left(f_0(x)\nabla\beta(x) + \beta(x)\nabla f_0(x)\right) + \frac{f_0(x)\beta(x)}{k} J_D(x), \tag{A.73}$$

where for simplicity we defined $D(x)$ to be

$$D(x) = \sum_{i=0}^{m} \frac{\nabla \beta_i(x)}{\beta_i(x)} - \sum_{i \in \mathcal{A}_c(x)} \frac{\tilde{\nabla} \beta_i(x)}{\tilde{\beta}_i(x) + \sigma_{d_i}^2(x)}, \tag{A.74}$$

and $J_D(x)$ is

$$\begin{aligned} J_D(x) = \sum_{i=1}^{m} \frac{\nabla^2 \beta_i(x)}{\beta_i(x)} - \sum_{i \in \mathcal{A}_c(x)} \frac{J_{\tilde{\nabla}\beta_i}(x)}{\tilde{\beta}_i(x) + \sigma_{d_i}^2(x)} - \sum_{i=1}^{m} \frac{\nabla \beta_i(x) \nabla \beta_i(x)^\top}{\beta_i(x)^2} \\ + \sum_{i \in \mathcal{A}_c(x)} \frac{\tilde{\nabla} \beta_i(x) \nabla \left(\tilde{\beta}_i(x) + \sigma_{d_i}^2(x) \right)^\top}{\left(\tilde{\beta}_i(x) + \sigma_{d_i}^2(x) \right)^2}. \end{aligned} \tag{A.75}$$

Let $v = \nabla \beta(x_c) / \|\nabla \beta(x_c)\|$ and $v_\perp$ a unit vector satisfying $v^\top v_\perp = 0$. Since at the critical points $\nabla \beta(x_c)$ is collinear with $\nabla f_0(x)$ (cf., A.44) we have that

$$v^\top J_{\tilde{b}_k}(x_c) v = \frac{v^\top D(x_c)}{k} \left(f_0(x_c) \|\nabla \beta(x_c)\| + \beta(x) \|\nabla f_0(x_c)\| \right) + \frac{f_0(x_c)\beta(x_c)}{k} v^\top J_D(x_c) v. \tag{A.76}$$

Notice that the norm of $D(x_c)$ is bounded (cf., (A.70)). By an analogous analysis one can show that $\beta(x_c) J_D(x_c)$ is bounded as well. Therefore the right hand side of the previous equality is of the order of $1/k$. On the other hand the quadratic form associated to the Hessian of $\varphi_k(x_c)$ can be shown to be of the order of (k^0) along the direction of v (cf., proof of Lemma 6). These facts combined imply that

$$\left\| \frac{v^\top J b_k(x_c) v}{v^\top J_{\tilde{\nabla} \varphi_k(x_c)} v} \right\| = O(1/k), \tag{A.77}$$

Observe that in the boundary of the free space $\nabla \beta(x)$ is collinear with $\nabla \beta_i(x)$, thus $D(x_c)$ is almost perpendicular to $v_\perp$ The same holds for part of the expression of $v_\perp^\top J_D(x_c) v_\perp$. And thus, the quadratic form of the Jacobian of the bias evaluated at $v_\perp$ can be approximated by

$$v_\perp^\top J_b k(x_c) v_\perp \approx \frac{f_0(x_c)\beta(x_c)}{k} v_\perp^\top \left(\sum_{i=1}^{m} \frac{\nabla^2 \beta_i(x)}{\beta_i(x)} - \sum_{i \in \mathcal{A}_c(x)} \frac{J_{\tilde{\nabla}\beta_i}(x)}{\tilde{\beta}_i(x) + \sigma_{d_i}^2(x)} \right) v_\perp. \tag{A.78}$$

The second factor of the previous expression can be show to be bounded by an analysis similar to that of the bound of $D(x)$. Since the critical points satisfy $\beta(x_c) = O(1/k)$ the previous expression is of the order of $1/k^2$. On the other hand the eigenvalues of the Hessian

of the navigation function are of the order of $1/k$ along the directions $v_\perp$. To observe the latter, evaluate the derivative of (A.43) along the direction $v_\perp$, i.e.,

$$v_\perp^\top J_{\tilde{\nabla}\varphi_k}(x_c)v_\perp = \beta(x_c)v_\perp^\top \nabla^2 f_0(x_c)v_\perp - \frac{f_0(x_c)}{k} v_\perp^\top \nabla^2 \beta(x_c) v_\perp. \tag{A.79}$$

And therefore along the direction $v_\perp$ it holds as well that

$$\left\| \frac{v_\perp^\top J b_k(x_c) v_\perp}{v_\perp^\top J_{\tilde{\nabla}\varphi_k(x_c)} v_\perp} \right\| = O(1/k). \tag{A.80}$$

The previous analysis shows that the C^1 norm of the gradient of the navigation function dominates by a factor of k that of the bias. Because $\varphi_k(x)$ is a Morse function, the gradient vector field is structurally stable [108]. Thus, this suggests that adding the bias will result in a topologically equivalent flow. These observations about the bias for the particular estimate here presented motivate Assumption 4 for a generic estimate.

A.2.2 Proof of Lemma 10

Let us start by defining a gradient-like vector field and by stating a result that is a direct consequence of Theorem B [115].

Definition 8 (Gradient like vector field [40]). *Let $x \in \mathbb{R}^n$ and let $g : \mathbb{R}^n \to \mathbb{R}^n$ be a smooth function, we say that $g(x)$ is a gradient like vector field if its non-wandering set consists of finitely many hyperbolic equilibrium states and the stable and unstable manifolds of singular points intersect transversally.*

Theorem 13 ([40]). *Let M^n be a smooth closed orientable manifold and let $g(x) : M^n \to [0, n]$ be a gradient-like vector field, then, there exists a function $V : M^n \to \mathbb{R}$ such that*

(i) *is twice differentiable and Morse*

(ii) *its critical points coincide with the set of the critical points of $g(x)$*

(iii) *$\dot{V}(x) = \nabla V(x)^\top g(x) < 0$, for any x such that $g(x) \neq 0$*

(iv) *V(x) = ind(x) for x such that $g(x) = 0$.*

Proof. See Theorem B in [115]. □

In what follows we will show that there exists a function satisfying (i)–(iv) for $g(x) = -\mathbb{E}\left[\hat{g}(x, \theta_t)\middle|\mathcal{G}_t\right]$. Equivalently we show that such function exists for a positive scaling of $g(x)$. Define then,

$$\tilde{g}(x) = -\left(\beta(x)\nabla f_0(x) - f_0(x)\nabla\beta(x)/k + \tilde{b}_k(x)\right) = -\left(\tilde{\nabla}\varphi_k(x) + \tilde{b}_k(x)\right). \tag{A.81}$$

Because the bias is not differentiable, $\tilde{g}(x)$ cannot be a gradient-like vector field and Theorem 13 cannot be applied directly. Hence, we will define a continuously differentiable approximation of the bias and show that said approximation is gradient like. To be precise, for every $\varepsilon > 0$ and for every $k > 0$ define the following neighborhood of the critical points of $\varphi_k(x)$

$$\mathcal{N}(\varepsilon, k) := \left\{ x \in \mathcal{F}, \left\| \tilde{\nabla} \varphi_k(x) \right\| < \varepsilon \right\}. \tag{A.82}$$

Since the bias is differentiable at x^* (cf., Assumption 4) and the minimum of the navigation function can be placed arbitrarily close to x^* (cf., Theorem 2) we can choose $\varepsilon_0' > 0$ and $K_0(\varepsilon_0') > 0$ such that the artificial potential is a navigation function and such that the bias is differentiable in a neighborhood of its minimum for any $k > K_0(\varepsilon_0')$. Likewise, since the discontinuities of the bias occur at distance $D > 0$ of the obstacles (cf., Assumption 4) there exists $\varepsilon_0'' > 0$ such that if $\varepsilon < \min\{\varepsilon_0', \varepsilon_0 \prime\prime\} = \varepsilon_0$ then $\tilde{b}_k(x)$ is C^1 in $\mathcal{N}(\varepsilon, k)$. Define then, for any $\varepsilon < \varepsilon_0$ the function $\tilde{b}_{\varepsilon,k}^{diff} : \mathbb{R}^n \to \mathbb{R}^n$ to be C^1 and to satisfy

$$\left\| \tilde{b}_{\varepsilon,k}^{diff}(x) \right\| < O(1/k), \quad \tilde{b}_{\varepsilon,k}^{diff}(x) = \tilde{b}_k(x) \forall x \in \mathcal{N}(\varepsilon, k). \tag{A.83}$$

In the following lemma we show that a perturbation of $\nabla \varphi_k(x)$ by an approximation of the bias satisfying (A.83) is gradient-like.

Lemma 16. *Under the Hypothesis of Lemma 10, for all $\varepsilon < \varepsilon_0$ and large enough k, the vector field $\tilde{g}^{diff}(x) = -\left(\tilde{\nabla} \varphi_k(x) + \tilde{b}_{k,\varepsilon}^{diff}(x)\right)$, with $\tilde{b}_{k,\varepsilon}^{diff}(x)$ satisfying* (A.83) *is gradient-like.*

Proof. We start by showing that the Lie derivative of $\varphi_k(x)$ along $\tilde{g}^{diff}(x)$ is negative for any $x \notin \mathcal{N}(\varepsilon, k)$ and therefore no point in $\mathcal{F} \setminus \mathcal{N}(\varepsilon, k)$ can belong to the non-wandering set of $\tilde{g}^{diff}(x)$.

$$\mathcal{L}_{\tilde{g}^{diff}(x)} \varphi_k(x) = -\nabla \varphi_k(x)^\top \left(\tilde{\nabla} \varphi_k(x) + \tilde{b}_{\varepsilon,k}^{diff}(x) \right). \tag{A.84}$$

Since $\tilde{\nabla} \varphi_k(x)$ is a scaling of $\nabla \varphi_k(x)$ we have that

$$\mathcal{L}_{\tilde{g}^{diff}(x)} \varphi_k(x) < -\left\| \nabla \varphi_k(x) \right\| \left(\left\| \tilde{\nabla} \varphi_k(x) \right\| - \left\| \tilde{b}_{\varepsilon,k}^{diff}(x) \right\| \right). \tag{A.85}$$

Because $\left\| \tilde{b}_{\varepsilon,k}^{diff}(x) \right\| < O(1/k)$ there exists $K_0(\varepsilon) > 0$ such that for any $k > K_0(\varepsilon)$ we have that $\| \tilde{b}_{\varepsilon,k}^{diff}(x) \| < \varepsilon$. Then, by definition of $\mathcal{N}(\varepsilon, k)$ we have that $\left\| \tilde{\nabla} \varphi_k(x) \right\| > \varepsilon$ which shows that $\mathcal{L}_{\tilde{g}^{diff}(x)} \varphi_k(x) < 0$ in $x \notin \mathcal{N}(\varepsilon, k)$ for $\varepsilon < \varepsilon_0$ and $k > K_0(\varepsilon)$. We are therefore left to show that in the neighborhood of the critical points the vector field is gradient-like. In particular, observe that in the neighborhood of the saddle points the field is topologically equivalent to that of $\nabla \varphi_k(x)$ because of Assumption 4. Since $\varphi_k(x)$ is Morse, the set of non-wandering points in each one of the neighborhoods is one hyperbolic equilibrium state

and the stable and unstable manifolds intersect transversally. We are thus left to show that in the same holds in the neighborhood of the minimum of $\varphi_k(x)$. Since the norm of the bias can be made arbitrarily small by increasing k, it suffices to show that the C^1 norm of $\tilde{\nabla}\varphi_k$ is constant with respect to k in the neighborhood of the minimum of $\varphi_k(x)$. We proceed to show the latter by analyzing the Jacobian of $\tilde{\nabla}\varphi_k(x)$

$$J_{\tilde{\nabla}\varphi_k}(x) = \beta(x)\nabla^2 f_0(x) + \left(1 - \frac{1}{k}\right)\nabla\beta(x)\nabla f_0(x)^\top - \frac{f_0(x)}{k}\nabla^2\beta(x). \tag{A.86}$$

Observe that the last term goes to zero as k goes to infinity and so does the second one. The reason for the latter is that the larger k the closer the local minimum of $\nabla\varphi_k(x)$ is to that of $\nabla f_0(x)$. So we are left to analyze the first term. Since the minimum of $\varphi_k(x)$ is away from the obstacles, the function $\beta(x)$ is bounded away from zero for all k. In addition $f_0(x)$ is strongly convex (cf., Assumption 2) hence its Hessian is bounded away from zero. This two facts together imply that the norm of the Jacobian of $\tilde{\nabla}\varphi_k(x)$ is $O(k^0)$. Since the original vector field $\tilde{\nabla}\varphi_k(x)$ is gradient-like the vector field $\tilde{g}^{diff}_{\varepsilon,k}(x)$ is it as well in the neighborhood of the minimum of $\nabla\varphi_k(x)$. The latter completes the proof that $\tilde{g}^{diff}(x)$ is gradient-like in $\mathcal{F}$. □

Since $\tilde{g}^{diff}(x)$ is gradient like, by virtue of Theorem 13 there exists a function $V_{\varepsilon,k}(x)$ satisfying (i)–(iv). We show next that (ii)–(iv) also hold for $\tilde{g}(x)$. Let us define the following set

$$\mathcal{N}'(k,\varepsilon,\varepsilon') = \left\{\nabla V_{\varepsilon,k}(x)^\top \tilde{g}^{diff}(x) > -\varepsilon'\right\}. \tag{A.87}$$

Since the norms of both $\tilde{b}^{diff}_{\varepsilon,k}(x)$ and $\tilde{b}_k(x)$ decrease at a rate of $1/k$, for every $\varepsilon' > 0$ there exists $K_1(\varepsilon,\varepsilon')$ such that for every $k > K_1(\varepsilon,\varepsilon')$ we have for all x that

$$\left|\nabla V_{\varepsilon,k}(x)^\top \left(\tilde{b}^{diff}_{\varepsilon,k}(x) - \tilde{b}_k(x)\right)\right| < \varepsilon'. \tag{A.88}$$

Hence, for any $x \notin \mathcal{N}'(k,\varepsilon,\varepsilon')$ we have that

$$\begin{aligned}\nabla V_{\varepsilon,k}(x)^\top \tilde{g}(x) &= \nabla V_{\varepsilon,k}(x)^\top \left(\tilde{g}^{diff}(x) - \tilde{b}^{diff}_{\varepsilon,k}(x) + \tilde{b}_k(x)\right)\\ &\leq -\varepsilon' + \left|\nabla V_{\varepsilon,k}(x)^\top \left(\tilde{b}^{diff}_{\varepsilon,k}(x) - \tilde{b}_k(x)\right)\right|,\end{aligned} \tag{A.89}$$

which shows that the Lie derivative of $V_{\varepsilon,k}(x)$ along the flow $\dot{x} = \tilde{g}(x)$ is negative for all $x \notin \mathcal{N}'(k,\varepsilon,\varepsilon')$. Thus (iii) holds outside $\mathcal{N}'(k,\varepsilon,\varepsilon')$ and since there are no critical points of $\tilde{g}(x)$ in said neighborhood (ii) and (iv) hold trivially. Next choose ε' to satisfy $\mathcal{N}'(k,\varepsilon,\varepsilon') \subset \mathcal{N}(k,\varepsilon)$. For any $x \in \mathcal{N}'(k,\varepsilon,\varepsilon') \subset \mathcal{N}(\varepsilon,k)$ we have that $\tilde{b}_{k,\varepsilon}(x) = \tilde{b}^{diff}_{k,\varepsilon}(x)$. Thus, because (ii)– (iv) hold for $\tilde{g}^{diff}(x)$ they also do for $\tilde{g}(x)$.

To complete the proof we are left to show that the critical points of $V_k(x)$ are arbitrarily

close to those of $\varphi_k(x)$ and that they have the same indices. The critical points of $V_k(x)$ satisfy

$$\tilde{g}(x) = \tilde{\nabla}\varphi_k(x) + \tilde{b}_k(x) = 0. \tag{A.90}$$

Since we have that $\|\tilde{b}_k(x)\| = O(1/k)$ the critical points of $\tilde{g}(x)$ satisfy that $\left\|\tilde{\nabla}\varphi_k(x)\right\| = O(1/k)$ which shows that the critical points can be placed arbitrarily close to those of $\varphi_k(x)$. The fact that their indices are the same is a consequence that in the neighborhood of the critical points $\tilde{\nabla}\varphi_k(x)$ and $\tilde{g}(x)$ are topologically equivalent as it was shown in the proof of Lemma 16.

A.3 Proofs of the results in Chapter 4

A.3.1 Proof of Lemma 11

In order to develop this proof we need to define the tangent cone and to state Lemma 17 relating the projection of a vector over it and the projection over a convex set

Definition 9 (Tangent cone)**.** *Let $X \subset \mathbb{R}^n$ be a closed convex set. We define the tangent cone to X at x_0 as*

$$T_X(x_0) = \overline{\bigcup_{\theta>0, x\in X} \theta(x - x_0)}. \tag{A.91}$$

The above union is over all the points of the set X and over all the positive reals θ. Notice that the $\bigcup_{\theta>0} \theta(x - x_0)$ is the ray from x_0 and intersecting the point x. Thus, the tangent cone is the closure of the cone formed by all rays emanating from x_0 and intersecting at least one point $x \in X$ with $x \neq x_0$.

Lemma 17. *Let $X \in \mathbb{R}^n$ be a closed convex set, let $x_0 \in X$ and let $v \in \mathbb{R}^n$. Then the projection of v over the set X at x_0 defined in* (4.16) *is*

$$\Pi_X(x_0, v) = P_{T_X(x_0)}(v). \tag{A.92}$$

Proof. The proof follows from Lemma 4.6 in [135]. □

Proof of Lemma 11. Consider the case in which $x_0 \in int(X)$. Then, for any v there exits a small enough $\delta > 0$ such that $x_0 + \delta v \in X$. Hence $P_X(x_0 + \delta v) = x_0 + \delta v$ and it holds that $P_X(x_0 + \delta v) - x_0 = v\delta$. Thus $\Pi_X(x, v) = v$ and (4.21) is verified. When $x_0 \in \partial X$ two cases are possible; either $x_0 + \delta v \in T_X(x_0)$ for small enough $\delta > 0$ or $x_0 + \delta v \notin T_X(x_0)$ for all $\delta > 0$. In the first case because of Lemma 17 it is verified that

$$\Pi_X(x_0, v) = P_{T_X(x_0)}(v) = v. \tag{A.93}$$

And therefore (4.21) holds. Let us now consider the last case in which $x_0 \in \partial X$ and $x_0 + \delta v \notin T_X(x_0)$. Because X is a convex set there exists a vector $a \in \mathbb{R}^n$ with $\|a\| = 1$ defining a supporting hyperplane at x_0 $\mathcal{H} = \{x \in \mathbb{R}^n : a^\top(x - x_0) = 0\}$, and for all $x \in X$ we have that

$$a^\top(x - x_0) \leq 0. \tag{A.94}$$

If the set X is smooth at x_0 then the border of the tangent cone at the point x_0 is contained in the hyperplane $\mathcal{H}$, therefore $\Pi_X(x_0, v) \subset \mathcal{H}$. Thus, $a^\top \Pi_X(x_0, v) = 0$ and we have as well that $a^\top v \geq 0$, otherwise there must exists a $\delta > 0$ such that $x_0 + \delta v \in T_X(x_0)$. On the other hand if there is a corner at x_0 there are infinite supporting hyperplanes.

One of them verifies that $a^\top v \geq 0$ and contains the boundary of the tangent cone, thus $a^\top \Pi_X(x_0, v) = 0$. Since $\Pi_X(x_0, v)$ is the projection of v over the tangent cone, we have that: $\Pi_X(x_0, v) = P_{T_X(x_0)}(v) = (a_\perp^\top v)a_\perp$, where $a_\perp \in \mathbb{R}^n$ and verifies that $a^\top a_\perp = 0$ and $\|a_\perp\| = 1$. Projecting the vectors $x_0 - x$ and v over a and $a_\perp$, we have

$$(x_0 - x)^\top v = (x_0 - x)^\top a v^\top a + (x_0 - x)^\top a_\perp v^\top a_\perp. \tag{A.95}$$

From the previous discussion the above equation reduces to

$$(x_0 - x)^\top v = (x_0 - x)^\top a v^\top a + (x_0 - x)^\top \Pi_X(x_0, v). \tag{A.96}$$

By combining the fact that $v^\top a \geq 0$ and (A.94) the left hand side of the above equality can be lower bounded by

$$(x_0 - x)^\top v \geq (x_0 - x)^\top \Pi_X(x_0, v). \tag{A.97}$$

Hence we have proved the lemma for all posible cases. □

A.4 Proofs of the results in Chapter 5

Lemma 18. *Let $f_0 : \mathbb{R}^n \to \mathbb{R}$ and $f : \mathbb{R}^n \to \mathbb{R}^m$ be convex functions. Then, for every matrix $\mathbf{K} \succ 0$, there exists $\mathbf{s}^* \in \mathbb{R}^m$ such that*

$$\boldsymbol{\lambda}^*(\mathbf{s}^*) = \mathbf{K}^{-1}\mathbf{s}^* \quad \textit{and} \quad \|\mathbf{s}^*\| < \infty, \tag{A.98}$$

where $\boldsymbol{\lambda}^*(\mathbf{s}^*)$ is the dual variable of problem (5.2) with $\mathbf{s} = \mathbf{s}^*$.

Proof. Because both $f_0(\mathbf{x})$ and $f(\mathbf{x})$ are convex functions, the primal probem $p^*(\mathbf{s})$ defined in (5.2) is a convex function on $\mathbf{s}$ [16, Section 5.6.1]. Let us next define the following regularized function

$$q^*(\mathbf{s}) = p^*(\mathbf{s}) + \frac{1}{2}\|\mathbf{s}\|^2_{\mathbf{K}^{-1}}, \tag{A.99}$$

where $\mathbf{K} \succ 0$. By introducing the regularizer, $q^*(\mathbf{s})$ is a strongly convex function. Hence its minimum $\mathbf{s}^* := \operatorname{argmin}_{\mathbf{s}\in\mathbb{R}^m} q^*(\mathbf{s})$ is such that $\|\mathbf{s}^*\| < \infty$. Likewise, $\mathbf{s}^*$ satisfies that $\nabla q^*(\mathbf{s}^*) = 0$. The latter is equivalent to

$$\nabla q^*(\mathbf{s}^*) = \nabla p^*(\mathbf{s}^*) + \mathbf{K}^{-1}\mathbf{s}^* = 0. \tag{A.100}$$

The proof is then completed by using the fact that $\boldsymbol{\lambda}^*(\mathbf{s}) = -\nabla p^*(\mathbf{s})$ for all $\mathbf{s} \in \mathbb{R}^m$ [16, Section 5.6.3]. □

A.5 Proofs of the results in Chapter 6

In this appendix we present some properties of the expected discounted reward and its gradient which are needed in the convergence analysis of functional stochastic gradient ascent.

Lemma 19. *Under Assumption 16 the expected discounted reward defined in* (6.3) *and the q-function defined in* (6.6) *satisfy*

$$|U(h)| < \frac{B_r}{1-\gamma} \quad \textit{and} \quad |Q(s,a;h)| < \frac{B_r}{1-\gamma} \quad \forall \quad h \in \mathcal{H}. \tag{A.101}$$

Proof. The triangle inequality applied to $|U(h)|$, with $U(h)$ defined in (6.3), yields

$$|U(h)| \leq \mathbb{E}\left[\left|\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right| \Big| h\right] \leq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t |r(s_t, a_t)| \Big| h\right], \tag{A.102}$$

Since the absolute value of the reward function $r(s,a)$ is bounded by B_r for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ (cf., Assumption 16) it follows that

$$|U(h)| \leq B_r \sum_{t=0}^{\infty} \gamma^t = \frac{B_r}{1-\gamma}. \tag{A.103}$$

The proof of the result for $Q(s,a;h)$ is analogous. □

Lemma 20. *Let Assumption 16 hold, then* $\nabla_h U(h, \cdot)$ *defined as in* (6.7) *is bounded for all* $h \in \mathcal{H}$.

Proof. Staring from (6.7) and considering $\|k(s,\cdot)\| = 1$ (cf., Definition 7), one can write

$$\|\nabla_h U(h,\cdot)\| \leq \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim\rho(s,a)} \left[|Q(s,a;h)| \left\| \Sigma^{-1} \left(a - h(s)\right) \right\| \right]. \tag{A.104}$$

And then use the result of Lemma 19 to further upper bound the norm of the gradient by

$$\|\nabla_h U(h,\cdot)\| \leq \frac{B_r}{(1-\gamma)^2} \mathbb{E}_{(s,a)\sim\rho(s,a)} \left[\left\| \Sigma^{-1} \left(a - h(s)\right) \right\| \right]. \tag{A.105}$$

By construction $\Sigma^{-1/2}(a - h(s)$ is a multivariate normal distribution, hence the expectation of its norm is bounded. □

Lemma 21. *Let Assumption 16 hold, with constant* B_r. *Then the gradient of the expected discounted reward satisfies*

$$\|\nabla_h U(h_1,\cdot) - \nabla_h U(h_2,\cdot)\|_{\mathcal{H}} \leq L_1 \|h_1 - h_2\|_{\mathcal{H}} + L_2 \|h_1 - h_2\|_{\mathcal{H}}^2, \tag{A.106}$$

for all $h_1, h_2 \in \mathcal{H}$ with L_1 and L_2 given by

$$L_1 = B_r \frac{(1-\gamma+p(1+\gamma))}{\lambda_{\min}(\Sigma)(1-\gamma)^3},\ L_2 = B_r \frac{(1+\gamma)\sqrt{p}}{\left(\lambda_{\min}(\Sigma)\right)^{3/2}(1-\gamma)^3}.$$

Proof. Consider the following bound to be used later

$$\|h(s)\| = |\langle h, \kappa(s,\cdot)\rangle_{\mathcal{H}}| \leq \|h\|. \tag{A.107}$$

due to the Cauchy-Scwartz inequality and with $\|\kappa(s,.)\| = 1$ (cf., Definition 7). Substituting (6.6) for $Q(s,a;h)$ in (6.24) it holds

$$\nabla_h U(h,\cdot) = \sum_{t=0}^{\infty}\sum_{u=0}^{\infty}\gamma^{t+u}\mathbb{E}_{p_h}\left[r(s_{t+u},a_{t+u})\kappa(s_t,\cdot)\zeta_t^h\right] \tag{A.108}$$

where we have defined the Gaussian variable $\zeta_t^h := \Sigma^{-1}(a_t - h(s_t))$ for notational brevity. The expectation in (A.108) is integrated with

$$p_h(\mathbf{s},\mathbf{a}) := p_{t+u}(\mathbf{s},\mathbf{a})\prod_{r=0}^{t+u}\pi_{h_1}(a_r|s_r) \tag{A.109}$$

with $\mathbf{s}$ and $\mathbf{a}$ collecting states and actions up to time $t+u$, and with $p_{t+u}(\mathbf{s},\mathbf{a}) := p(s_0)\prod_{r=0}^{t+u-1}p(s_{r+1}|s_r,a_r)$. Expanding the expectation as an integral and adding and subtracting

$$\mathbb{E}_{p_{h_2}}\left[r(s_{t+u},a_{t+u})\kappa(s_t,\cdot)\zeta_t^{h_1}\right], \tag{A.110}$$

yields

$$\begin{aligned}\nabla_h U(h_1,\cdot) - \nabla_h U(h_2,\cdot) = &\sum_{t=0}^{\infty}\sum_{u=0}^{\infty}\gamma^{t+u}\int r(s_{t+u},a_{t+u})\zeta_t^{h_1}\kappa(s_t,\cdot)p_{t+u}(\mathbf{s},\mathbf{a})\\ &\times\left(\prod_{r=0}^{t+u}\pi_{h_1}(a_r|s_r) - \prod_{r=0}^{t+u}\pi_{h_2}(a_r|s_r)\right)d\mathbf{s}d\mathbf{a}\\ &+\sum_{t=0}^{\infty}\sum_{u=0}^{\infty}\gamma^{t+u}\int r(s_{t+u},a_{t+u})\Sigma^{-1}\left(h_2(s_t)-h_1(s_t)\right)\times\kappa(s_t,\cdot)p_{h_2}(\mathbf{s},\mathbf{a})d\mathbf{s}d\mathbf{a}.\end{aligned} \tag{A.111}$$

Using that $|r(s_{t+u},a_{t+u})| \leq B_r$ and $\|\kappa(s_t,\cdot)\| = 1$ (cf., Assumption 16 and Definition 7

repsectively) we can bound

$$\|\nabla_h U(h_1,\cdot) - \nabla_h U(h_2,\cdot)\| \leq \sum_{t=0}^{\infty}\sum_{u=0}^{\infty} \gamma^{t+u} B_r (I_1 + I_2) \tag{A.112}$$

with

$$I_1 := \int \left\|\zeta_t^{h_1}\right\| |\Delta_\pi(h_1,h_2,s,a)|\, p_{t+u}(\mathbf{s},\mathbf{a}) d\mathbf{s} d\mathbf{a} \tag{A.113}$$

$$I_2 := \int \left\|\Sigma^{-1}\left(h_2(s_t) - h_1(s_t)\right)\right\| p_{h_2}(\mathbf{s},\mathbf{a}) d\mathbf{s} d\mathbf{a}, \tag{A.114}$$

$$\Delta_\pi(h_1,h_2,s,a) := \prod_{r=0}^{t+u} \pi_{h_2}(a_r|s_r) - \prod_{r=0}^{t+u} \pi_{h_1}(a_r|s_r). \tag{A.115}$$

To obtain a bound for I_1 in (A.112) define $h_\lambda = \lambda h_1 + (1-\lambda)h_2$ with $\lambda \in [0,1]$. Next, consider the Taylor expansion of $\prod_{r=0}^{t+u} \pi_h(a_r|s_r)$ as a function of h, which yields

$$\Delta_\pi(h_1,h_2,\mathbf{s},\mathbf{a}) = \sum_{r=0}^{t+u} \left\langle \zeta_r^{h_\lambda} \prod_{r=0}^{t+u} \pi_{h_\lambda}(a_r|s_r)\kappa(s_r,\cdot), h_1 - h_2 \right\rangle \tag{A.116}$$

Thus, the absolute value of Δ_π can be bounded via the Cauchy-Schwartz inequality

$$|\Delta_\pi(h_1,h_2,s,a)| \leq \|h_1 - h_2\| \sum_{r=0}^{t+u} \left\|\zeta_r^{h_\lambda}\right\| \prod_{r=0}^{t+u} \pi_{h_\lambda}(a_r|s_r). \tag{A.117}$$

With this in mind we bound the first integral in (A.112). The following inequalities are explained below.

$$\begin{aligned}
I_1 &= \int p_{t+u}(\mathbf{s},\mathbf{a}) \left\|\zeta_t^{h_1}\right\| |\Delta_\pi(h_1,h_2,s,a)|\, d\mathbf{s} d\mathbf{a} \\
&\leq \|h_1 - h_2\| \int p_{h_\lambda}(\mathbf{s},\mathbf{a}) \left\|\zeta_t^{h_1}\right\| \sum_{r=0}^{t+u} \left\|\zeta_r^{h_\lambda}\right\| d\mathbf{s} d\mathbf{a} \\
&\leq \|h_1 - h_2\|\, \mathbb{E}_{p_{h_\lambda}}\left[\left\|\zeta_t^{h_1}\right\| \sum_{r=0}^{t+u} \left\|\zeta_r^{h_\lambda}\right\|\right] \\
&= \|h_1 - h_2\|\, \mathbb{E}_{p_{h_\lambda}}\left[\left\|\zeta_t^{h_\lambda} + \Sigma^{-1}(h_\lambda(s_t) - h_1(s_t))\right\| \sum_{r=0}^{t+u} \left\|\zeta_r^{h_\lambda}\right\|\right] \\
&\leq \|h_1 - h_2\| \sum_{r=0}^{t+u} \mathbb{E}_{p_{h_\lambda}}\left[\left\|\zeta_t^{h_\lambda}\right\| \left\|\zeta_r^{h_\lambda}\right\|\right] + \|h_1 - h_2\| \sum_{r=0}^{t+u} \|h_\lambda - h_1\| \mathbb{E}_{p_{h_\lambda}}\left[\left\|\Sigma^{-1}\zeta_r^{h_\lambda}\right\|\right] \\
&\leq (t+u+1)\left(\frac{p\,\|h_1 - h_2\|}{\lambda_{\min}(\Sigma)} + \frac{\sqrt{p}\,\|h_1 - h_2\|^2}{(\lambda_{\min}(\Sigma))^{3/2}}\right)
\end{aligned} \tag{A.118}$$

The first inequality results from substituting (A.117) and using the definition of p_{h_λ} in (A.109). Then write the integral as an expectation. The third one states that $\zeta_t^{h_1} = \zeta_t^{h_\lambda} + \Sigma^{-1}(h_\lambda(s_t) - h_1(s_t)$. The next one combines the triangle inequality with the bound (A.107) applied to $h(s) = h_\lambda(s) - h_1(s)$. Finally, we used that $\Sigma^{-1/2}\zeta_t^{h_\lambda}$ and $\Sigma^{-1/2}\zeta_r^{h_\lambda}$ are multivariate independent white Gaussian variables first order moment bounded by $\sqrt{p}$.

To bound I_2 in (A.112), apply again (A.107) to $h(s) = h_2(s) - h_1(s)$. It follows that the norm of the second integral is bounded by $(\lambda_{\min}(\Sigma))^{-1}\|h_1 - h_2\|$, which together with (A.118) can be substituted in (A.112) to conclude the proof, after adding the geometric sum

$$\sum_{t=0}^{\infty}\sum_{u=0}^{\infty}(t+u+1)\gamma^{t+u} = \frac{1+\gamma}{(1-\gamma)^3}. \tag{A.119}$$

□

Lemma 22. *The second and third moments of the estimate $\hat{\nabla}_h U(h,\cdot)$ are bounded by*

$$\mathbb{E}\left[\left\|\hat{\nabla}_h U(h,\cdot)\right\|^2\right] \leq \sigma^2 \quad \textit{and} \quad \mathbb{E}\left[\left\|\hat{\nabla}_h U(h,\cdot)\right\|^3\right] \leq \sigma^3, \tag{A.120}$$

with

$$\sigma = \frac{(3\gamma)^{1/3}}{(1-\gamma)^2}\frac{1}{\lambda_{\min}\left(\Sigma^{1/2}\right)}\left(4\frac{\Gamma(2+p/2)}{\Gamma(p/2)}\right)^{1/4}. \tag{A.121}$$

where $\Gamma(\cdot)$ is the Gamma function.

Proof. Let us start by bounding the cube the norm of the stochastic gradient defined in (6.16).

$$\left\|\hat{\nabla}_h U(h,\cdot)\right\|^3 \leq \frac{1}{8(1-\gamma)^3}\left\|\hat{Q}(s_T,a_T;h) - \hat{Q}(s_T,\bar{a}_T;h)\right\|^3 \|\kappa(s_T,\cdot)\|^3 \left\|\Sigma^{-1}(a_T - h(s_T))\right\|^3. \tag{A.122}$$

Using the fact that $\|\kappa(s_t,\cdot)\| = 1$ (cf., Definition 7) and the fact that the difference between estimates of Q is bounded by $B_r(T_Q + T'_Q)$, (A.122) is upper bounded by

$$\left\|\hat{\nabla}_h U(h,\cdot)\right\|^3 \leq \frac{B_r^3}{8(1-\gamma)^3}(T_Q + T'_Q)^3 \left\|\Sigma^{-1}(a_T - h(s_T))\right\|^3. \tag{A.123}$$

From the independence of T_Q and T'_Q with respect to the state evolution, and the monotonicity of the expectation, it results

$$\mathbb{E}\left[\left\|\hat{\nabla}_h U(h,\cdot)\right\|^3\right] \leq \frac{B_r^3}{8(1-\gamma)^3}\mathbb{E}\left[\left(T_Q + T'_Q\right)^3\right]\mathbb{E}\left[\left\|\Sigma^{-1}(a_T - h(s_T))\right\|^3\right]. \tag{A.124}$$

The sum of two independent geometric variables satisfies

$$P(T_Q + T'_Q = k) = (1-\gamma)^2(k+1)\gamma^k. \tag{A.125}$$

Thus, the third moment is upper bounded by

$$\mathbb{E}\left[(T_Q + T'_Q)^3\right] = \sum_{k=0}^{\infty} k^3(1-\gamma)^2(k+1)\gamma^k = \frac{\gamma(1+14\gamma+8\gamma^2)}{(1-\gamma)^3} \leq \frac{23\gamma}{(1-\gamma)^3}$$

where the last inequality follows from the fact that $\gamma < 1$. On the other hand observe that $\left\|\Sigma^{-1/2}a_T - h(s_T)\right\|^2$ is Chi-squared with parameter p since it is a sum of squares of normal random variables. Hence, the second expectation in (A.124) can be bounded using Jensen's inequality by,

$$\begin{aligned}\mathbb{E}\left[\left\|\Sigma^{-1}(a_T - h(s_T))\right\|^3\right] &\leq \frac{1}{\lambda_{\min}(\Sigma^{1/2})^3}\mathbb{E}\left[\chi_p^{3/2}\right] \leq \frac{1}{\lambda_{\min}(\Sigma^{1/2})^3}\mathbb{E}\left[\chi_p^2\right]^{3/4} \\ &= \frac{1}{\lambda_{\min}(\Sigma^{1/2})^3}\left(4\frac{\Gamma(2+p/2)}{\Gamma(p/2)}\right)^{3/4}\end{aligned} \tag{A.126}$$

Substituting (A.126) and (A.126) in (A.124) yields the the bound for the third moment of the stochastic gradient in (A.120). To validate the bound on the second moment consider $x = \left\|\hat{\nabla}_h U(h,\cdot)\right\|^3$ and observe that since $x^{2/3}$ is a concave function one can reverse Jensen's inequality to obtain

$$\mathbb{E}\left[\left(\left\|\hat{\nabla}_h U(h,\cdot)\right\|^3\right)^{2/3}\right] \leq \mathbb{E}\left[\left\|\hat{\nabla}_h U(h,\cdot)\right\|^3\right]^{2/3} \leq \left(\sigma^3\right)^{2/3}$$

which completes the proof. □

Lemma 23. *Let $e_j = \hat{\nabla}_h U(h_j) - \nabla_h U(h_j)$ and let η_j be such that it satisfies* (6.30)*. Then, the sequence*

$$S_k = \sum_{j=0}^{k} \eta_j e_j, \tag{A.127}$$

converges to a finite limit with probability one.

Proof. By virtue of Theorem 5.4.9 [29]), it suffices to show that S_k is a square integrable margingale and that

$$\lim_{n\to\infty} \sum_{m=1}^{n} \mathbb{E}\left[(S_m - S_{m-1})^2 \middle| \mathcal{F}_m\right] < \infty \quad \text{a.e.} \tag{A.128}$$

Recall that the estimate of the gradient is unbiased, i.e. $\mathbb{E}\left[\hat{\nabla}_h U(h_k,\cdot)\middle|\mathcal{F}_k\right] = \nabla_h U(h_k,\cdot)$, hence we have that $\mathbb{E}\left[e_k\middle|\mathcal{F}_k\right] = 0$. This allows us to write

$$\mathbb{E}\left[S_k\middle|\mathcal{F}_k\right] = S_{k-1} + \mathbb{E}\left[\eta_k e_k\middle|\mathcal{F}_k\right] = S_{k-1}. \tag{A.129}$$

Thus S_k is a martingale. To show that it is square integrable, observe that we can compute squared norm of S_k as

$$\begin{aligned}\|S_k\|^2 = \left\|\sum_{j=0}^{k}\eta_j e_j\right\|^2 &= \eta_k^2\|e_k\|^2 + 2\eta_k e_k^\top \sum_{j=0}^{k-1}\eta_j e_j + \left\|\sum_{j=0}^{k-1}\eta_j e_j\right\|^2 \\ &= \eta_k^2\|e_k\|^2 + 2\eta_k e_k^\top S_{k-1} + \|S_{k-1}\|^2.\end{aligned} \tag{A.130}$$

Take the expectation with respect to the sigma field $\mathcal{F}_k$ and use the fact that $\mathbb{E}\left[e_k\middle|\mathcal{F}_k\right] = 0$ to write

$$\mathbb{E}\left[\|S_k\|^2\middle|\mathcal{F}_k\right] = \eta_k^2\mathbb{E}\left[\|e_k\|^2\middle|\mathcal{F}_k\right] + \|S_{k-1}\|^2. \tag{A.131}$$

The previous expression implies that

$$\mathbb{E}\left[\|S_k\|^2\right] = \eta_k^2\mathbb{E}\left[\|e_k\|^2\right] + \mathbb{E}\left[\|S_{k-1}\|^2\right]. \tag{A.132}$$

Recursively we have that

$$\mathbb{E}\left[\|S_k\|^2\right] = \sum_{j=0}^{k}\eta_j^2\mathbb{E}\left[\|e_j\|^2\right]. \tag{A.133}$$

Since the step sizes are square summable and the second moment of the error is bounded (cf., lemmas 20 and 22) the second moment of S_k is bounded for all k. We next show that (A.128) holds. Observe that by definition of S_k (cf., (A.127)) one can write

$$\sum_{m=1}^{n}\mathbb{E}\left[\|S_m - S_{m-1}\|^2\middle|\mathcal{F}_m\right] = \sum_{m=1}^{n}\mathbb{E}\left[\|\eta_m e_m\|^2\middle|\mathcal{F}_m\right] = \sum_{m=1}^{n}\eta_m^2\mathbb{E}\left[\|e_m\|^2\middle|\mathcal{F}_m\right]. \tag{A.134}$$

Which is bounded for all n as it was previously argued. This completes the proof that $\lim_{k\to\infty} S_k$ converges to a finite random variable with probability one.

□

Bibliography

[1] "Openai gym– continuous mountine car," https://gym.openai.com/envs/MountainCarContinuous-v0/.

[2] A. Argyriou, C. A. Micchelli, and M. Pontil, "When is there a representer theorem? vector versus matrix regularizers," *Journal of Machine Learning Research*, vol. 10, no. Nov, pp. 2507–2529, 2009.

[3] K. B. Ariyur and M. Krstic, *Real-time optimization by extremum-seeking control*. John Wiley & Sons, 2003.

[4] K. J. Arrow and L. Hurwicz, *Studies in linear and nonlinear programming*. CA: Stanford University Press, 1958.

[5] O. Arslan, *Clustering-based robot navigation and control*. University of Pennsylvania, 2016.

[6] O. Arslan and D. E. Koditschek, "Exact robot navigation using power diagrams," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–8.

[7] ——, "Sensor-based reactive navigation in unknown convex sphere worlds," in *submitted to) the 12th International Workshop on the Algoritmic Foundations of Robotics (WAFR)*, 2016.

[8] N. Atanasov, J. Le Ny, N. Michael, and G. J. Pappas, "Stochastic source seeking in complex environments," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 3013–3018.

[9] S.-i. Azuma, M. S. Sakar, and G. J. Pappas, "Stochastic source seeking by mobile robots," *Automatic Control, IEEE Transactions on*, vol. 57, no. 9, pp. 2308–2321, 2012.

[10] J. Barraquand, B. Langlois, and J.-C. Latombe, "Numerical potential field techniques for robot path planning," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 22, no. 2, pp. 224–241, 1992.

[11] J. Barraquand and J.-C. Latombe, "A Monte-Carlo algorithm for path planning with many degrees of freedom," in *Robotics and Automation, 1990. Proceedings., 1990 IEEE International Conference on*. IEEE, 1990, pp. 1712–1717.

[12] D. P. Bertsekas, *Nonlinear programming.* Athena Sci., Belmont, 1999.

[13] S. Bhatnagar, D. Precup, D. Silver, R. S. Sutton, H. R. Maei, and C. Szepesvári, "Convergent temporal-difference learning with arbitrary smooth function approximation," in *Advances in Neural Information Processing Systems*, 2009, pp. 1204–1212.

[14] D. Blackwell, "An analog of the minimax theorem for vector payoffs," *Pacific Journal of Mathematics*, vol. 6, no. 1, pp. 1–8, 1956.

[15] V. S. Borkar, "Stochastic approximation," *Cambridge Books*, 2008.

[16] S. Boyd and L. Vandenberghe, *Convex optimization.* Cambridge university press, 2004.

[17] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[18] T.-H. Chang, A. Nedić, and A. Scaglione, "Distributed constrained optimization by consensus-based primal-dual perturbation method," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1524–1538, 2014.

[19] T. Chen, Q. Ling, and G. B. Giannakis, "An online convex optimization approach to dynamic network resource allocation," *arXiv preprint arXiv:1701.03974*, 2017.

[20] T. Chen, Q. Ling, and G. B. Giannakis, "An online convex optimization approach to proactive network resource allocation," *IEEE Transactions on Signal Processing*, vol. 65, no. 24, pp. 6350–6364, 2017.

[21] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255–312, 2007.

[22] H. Choset, E. Acar, A. A. Rizzi, and J. Luntz, "Exact cellular decompositions in terms of critical points of morse functions," in *Robotics and Automation, 2000. Proceedings. ICRA'00. IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 2270–2277.

[23] H. Choset and P. Pignon, "Coverage path planning: The boustrophedon cellular decomposition," in *Field and Service Robotics.* Springer, 1998, pp. 203–209.

[24] H. M. Choset, *Principles of robot motion: theory, algorithms, and implementation.* MIT press, 2005.

[25] C. I. Connolly, J. Burns, and R. Weiss, "Path planning using Laplace's equation," in *Robotics and Automation, 1990. Proceedings., 1990 IEEE International Conference on.* IEEE, 1990, pp. 2102–2106.

[26] A. De and D. E. Koditschek, "Toward dynamical sensor management for reactive wall-following," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on.* IEEE, 2013, pp. 2400–2406.

[27] M. P. Deisenroth, G. Neumann, J. Peters *et al.*, "A survey on policy search for robotics," *Foundations and Trends® in Robotics*, vol. 2, no. 1–2, pp. 1–142, 2013.

[28] D. V. Dimarogonas, K. J. Kyriakopoulos, and D. Theodorakatos, "Totally distributed motion control of sphere world multi-agent systems using decentralized navigation functions," in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on.* IEEE, 2006, pp. 2430–2435.

[29] R. Durrett, *Probability: theory and examples.* Cambridge university press, 2010.

[30] M. Fazlyab, S. Paternain, V. M. Preciado, and A. Ribeiro, "Interior point method for dynamic constrained optimization in continuous time," in *American Control Conference (ACC), 2016.* IEEE, 2016, pp. 5612–5618.

[31] M. Fazlyab, S. Paternain, V. M. Preciado, and A. Ribeiro, "Prediction-correction interior-point method for time-varying convex optimization," *IEEE Transactions on Automatic Control*, 2017.

[32] D. Feijer and F. Paganini, "Stability of primal–dual gradient dynamics and applications to network optimization," *Automatica*, vol. 46, no. 12, pp. 1974–1981, 2010.

[33] I. Filippidis, "Navigation functions for unknown sphere worlds, general geometries, their inverse problem and combination with formal methods," Diploma Thesis, National Technical University of Athens, 2011.

[34] I. Filippidis and K. J. Kyriakopoulos, "Adjustable navigation functions for unknown sphere worlds," in *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on.* IEEE, 2011, pp. 4276–4281.

[35] I. "Filippidis and K. J. "Kyriakopoulos, "Navigation functions for focally admissible surfaces," in *American Control Conference (ACC), 2013.* IEEE, 2013, pp. 994–999.

[36] I. F. Filippidis and K. J. Kyriakopoulos, "Navigation functions for everywhere partially sufficiently curved worlds," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on.* IEEE, 2012, pp. 2115–2120.

[37] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning.* Springer series in statistics New York, 2001, vol. 1.

[38] S. S. Ge and Y. J. Cui, "New potential functions for mobile robot path planning," *IEEE Transactions on robotics and automation*, vol. 16, no. 5, pp. 615–620, 2000.

[39] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.

[40] V. Z. Grines, E. Y. Gurevich, and O. V. Pochinka, "The energy function of gradient-like flows and the topological classification problem," *Mathematical Notes*, vol. 96, no. 5-6, pp. 921–927, 2014.

[41] S. Hart and A. Mas-Colell, "A general class of adaptive strategies," *Journal of Economic Theory*, vol. 98, no. 1, pp. 26–54, 2001.

[42] E. Hazan, A. Agarwal, and S. Kale, "Logarithmic regret algorithms for online convex optimization," *Machine Learning*, vol. 69, no. 2-3, pp. 169–192, 2007.

[43] M. W. Hirsch, S. Smale, and R. L. Devaney, *Differential equations, dynamical systems, and an introduction to chaos*. Academic press, 2004, vol. 60.

[44] R. A. Howard, *Dynamic programming and Markov processes*. Wiley for The Massachusetts Institute of Technology, 1964.

[45] B. D. Ilhan, A. M. Johnson, and D. E. Koditschek, "Autonomous legged hill ascent," *Journal of Field Robotics*, 2018, to Appear.

[46] A. M. Johnson, M. T. Hale, G. Haynes, and D. E. Koditschek, "Autonomous legged hill and stairwell ascent," in *Safety, Security, and Rescue Robotics (SSRR), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 134–142.

[47] S. Kale, C. Lee, and D. Pál, "Hardness of online sleeping combinatorial optimization problems," *arXiv preprint arXiv:1509.03600*, 2015.

[48] V. Kanade, H. B. McMahan, and B. Bryan, "Sleeping experts and bandits with stochastic action availability and adversarial rewards," in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 272–279.

[49] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," *Int. J. Rob. Res.*, vol. 5, no. 1, pp. 90–98, Apr. 1986. [Online]. Available: http://dx.doi.org/10.1177/027836498600500106

[50] O. Khatib, "Commande dynamique dans l'espace opérationnel des robots manipulateurs en présence d'obstacles," Ph.D. dissertation, 1980.

[51] P. Khosla and R. Volpe, "Superquadric artificial potentials for obstacle avoidance and approach," in *Robotics and Automation, 1988. Proceedings., 1988 IEEE International Conference on*. IEEE, 1988, pp. 1778–1784.

[52] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *Trans. on Sig. Proc.*, vol. 52, no. 8, pp. 2165–2176, 2004.

[53] R. Kleinberg, A. Niculescu-Mizil, and Y. Sharma, "Regret bounds for sleeping experts and bandits," *Machine learning*, vol. 80, no. 2-3, pp. 245–272, 2010.

[54] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[55] D. E. Koditschek, "Strict global Lyapunov functions for mechanical systems," 1988.

[56] D. E. Koditschek, "The control of natural motion in mechanical systems," *Journal of dynamic systems, measurement, and control*, vol. 113, no. 4, pp. 547–551, 1991.

[57] D. E. Koditschek and E. Rimon, "Robot navigation functions on manifolds with boundary," *Advances in Applied Mathematics*, vol. 11, no. 4, pp. 412–442, 1990.

[58] J. Konečnỳ and P. Richtárik, "Semi-stochastic gradient descent methods," *arXiv preprint arXiv:1312.1666*, 2013.

[59] A. Koppel, S. Paternain, C. Richard, and A. Ribeiro, "Decentralized online learning with kernels," *arXiv preprint arXiv:1710.04062*, 2017.

[60] A. Koppel, G. Warnell, E. Stump, and A. Ribeiro, "Parsimonious online learning with kernels via sparse projections in function space," *arXiv preprint arXiv:1612.04111*, 2016.

[61] A. Koppel, G. Warnell, E. Stump, P. Stone, and A. Ribeiro, "Breaking bellman's curse of dimensionality: Efficient kernel gradient temporal difference," *arXiv preprint arXiv:1709.04221*, 2017.

[62] B. H. Krogh, *A generalized potential field approach to obstacle avoidance control*. RI/SME, 1984.

[63] M. Krstić and H.-H. Wang, "Stability of extremum seeking feedback for general nonlinear dynamic systems," *Automatica*, vol. 36, no. 4, pp. 595–601, 2000.

[64] V. Kumar, D. Rus, and S. Singh, "Robot and sensor networks for first responders," *Pervasive Computing, IEEE*, vol. 3, no. 4, pp. 24–33, 2004.

[65] H. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media, 2003, vol. 35.

[66] H. Kushner and E. Sanvicente, "Stochastic approximation of constrained systems with system and constraint noise," *Automatica*, vol. 11, no. 4, pp. 375–380, 1975.

[67] J. Kwon and P. Mertikopoulos, "A continuous-time approach to online optimization," *arXiv preprint arXiv:1401.6956*, 2014.

[68] S. M. LaValle, *Planning algorithms*. Cambridge university press, 2006.

[69] S. M. LaValle and J. J. Kuffner Jr, "Rapidly-exploring random trees: Progress and prospects," 2000.

[70] S. Lee, A. Ribeiro, and M. M. Zavlanos, "Distributed continuous-time online optimization using saddle-point methods," in *Decision and Control (CDC), 2016 IEEE 55th Conference on*. IEEE, 2016, pp. 4314–4319.

[71] G. Lever and R. Stafford, "Modelling policies in mdps in reproducing kernel hilbert space," in *A. I. and Statistics*, 2015, pp. 590–598.

[72] N. Li, C. Zhao, and L. Chen, "Connecting automatic generation control and economic dispatch from an optimization view," *IEEE Transactions on Control of Network Systems*, vol. 3, no. 3, pp. 254–264, 2016.

[73] F. Lingelbach, "Path planning using probabilistic cell decomposition," in *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, vol. 1. IEEE, 2004, pp. 467–472.

[74] G. Lionis, X. Papageorgiou, and K. J. Kyriakopoulos, "Locally computable navigation functions for sphere worlds," in *Robotics and Automation, 2007 IEEE International Conference on*. IEEE, 2007, pp. 1998–2003.

[75] G. Lionis, X. Papageorgiou, and K. J. Kyriakopoulos, "Towards locally computable polynomial navigation functions for convex obstacle workspaces," in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*. IEEE, 2008, pp. 3725–3730.

[76] S.-J. Liu and M. Krstic, "Stochastic source seeking for nonholonomic unicycle," *Automatica*, vol. 46, no. 9, pp. 1443 – 1453, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0005109810002487

[77] S. G. Loizou, "Closed form navigation functions based on harmonic potentials," in *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*. IEEE, 2011, pp. 6361–6366.

[78] S. G. Loizou, "Navigation functions in topologically complex 3-d workspaces," in *American Control Conference (ACC), 2012*. IEEE, 2012, pp. 4861–4866.

[79] S. H. Low and D. E. Lapsley, "Optimization flow control- i: basic algorithm and convergence," *IEEE/ACM Transactions on Networking (TON)*, vol. 7, no. 6, pp. 861–874, 1999.

[80] T. Lozano-Perez, J. L. Jones, E. Mazer, P. O'Donnell, E. W. Grimson, P. Tournassoud, A. Lanusse *et al.*, "Handey: A robot system that recognizes, plans, and manipulates," in *Robotics and Automation. Proceedings. 1987 IEEE International Conference on*, vol. 4. IEEE, 1987, pp. 843–849.

[81] T. Lozano-Pérez and M. A. Wesley, "An algorithm for planning collision-free paths among polyhedral obstacles," *Communications of the ACM*, vol. 22, no. 10, pp. 560–570, 1979.

[82] J. Lygeros, K. H. Johansson, S. N. Simic, J. Zhang, and S. S. Sastry, "Dynamical properties of hybrid automata," *IEEE Transactions on automatic control*, vol. 48, no. 1, pp. 2–17, 2003.

[83] M. Mahdavi, R. Jin, and T. Yang, "Trading regret for efficiency: online convex optimization with long term constraints," *Journal of Machine Learning Research*, vol. 13, no. Sep, pp. 2503–2528, 2012.

[84] D. Maistroskii, "Gradient methods for finding saddle points," *Matekon*, vol. 14, no. 1, pp. 3–22, 1977.

[85] D. Mellinger and V. Kumar, "Minimum snap trajectory generation and control for quadrotors," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, May 2011.

[86] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[87] A. Nedić and A. Ozdaglar, "Subgradient methods for saddle-point problems," *Journal of optimization theory and applications*, vol. 142, no. 1, pp. 205–228, 2009.

[88] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), Tech. Rep., 2011.

[89] Y. "Nesterov and V. "Spokoiny, "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, 2017.

[90] G. Neu and M. Valko, "Online combinatorial optimization with stochastic decision sets and adversarial losses," in *Advances in Neural Information Processing Systems*, 2014, pp. 2780–2788.

[91] W. S. Newman, "High-speed robot control in complex environments," Ph.D. dissertation, Massachusetts Institute of Technology, 1987.

[92] P. Ögren, E. Fiorelli, and N. E. Leonard, "Cooperative control of mobile sensor networks: Adaptive gradient climbing in a distributed environment," *Automatic Control, IEEE Transactions on*, vol. 49, no. 8, pp. 1292–1302, 2004.

[93] S. Paternain, J. A. Bazerque, A. Small, and A. Ribeiro, "Stochastic online policy gradient ascent in reproducing kernel hilbert spaces," *IEEE Transactions on Automatic Control (Submitted)*, 2018.

[94] S. Paternain, D. E. Koditschek, and A. Ribeiro, "Navigation functions for convex potentials in a space with convex obstacles," *IEEE Transactions on Automatic Control*, 2017.

[95] S. Paternain, M. Morari, and A. Ribeiro, "A prediction-correction method for model predictive control," p. (to appear), 2018.

[96] S. Paternain and A. Ribeiro, "Online learning of optimal strategies in unknown environments," in *Decision and Control (CDC), 2015 IEEE 54th Annual Conference on*. IEEE, 2015, pp. 3951–3958.

[97] S. Paternain and A. Ribeiro, "Stochastic artificial potentials for online safe navigation," *arXiv preprint arXiv:1701.00033*, 2016.

[98] S. "Paternain and A. "Ribeiro, "Online learning of feasible strategies in unknown environments," *IEEE Transactions on Automatic Control*, vol. 62, no. 6, pp. 2807–2822, 2017.

[99] Paternain, Santiago and A. Ribeiro, "Online learning of feasible strategies in unknown environments," in *American Control Conference (ACC), 2015*. IEEE, 2015, pp. 4231–4238.

[100] Paternain, Santiago and Ribeiro, Alejandro, "Safe online navigation of convex potentials in spaces with convex obstacles," in *Decision and Control (CDC), 2017 IEEE 56th Annual Conference on*. IEEE, 2017, pp. 2473–2478.

[101] R. Pemantle, "Nonconvergence to unstable points in urn models and stochastic approximations," *The Annals of Prob.*, pp. 698–712, 1990.

[102] M. Pontil, Y. Ying, and D.-X. Zhou, "Error analysis for online gradient descent algorithms in reproducing kernel hilbert spaces," Tech. Report, Dep. of Comp. Sci., Univ. College London, Tech. Rep., 2005.

[103] A. Y. Popkov, "Gradient methods for nonstationary unconstrained optimization problems," *Automation and Remote Control*, vol. 66, no. 6, pp. 883–891, 2005.

[104] M. Rásonyi, L. Stettner *et al.*, "On utility maximization in discrete-time financial market models," *The Annals of Applied Probability*, vol. 15, no. 2, pp. 1367–1395, 2005.

[105] E. Rimon and D. E. Koditschek, "The construction of analytic diffeomorphisms for exact robot navigation on star worlds," *Transactions of the American Mathematical Society*, vol. 327, no. 1, pp. 71–116, 1991.

[106] E. "Rimon and D. E. "Koditschek, "Exact robot navigation using artificial potential functions," *Robotics and Automation, IEEE Transactions on*, vol. 8, no. 5, pp. 501–518, 1992.

[107] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[108] C. Robinson, "Structural stability on manifolds with boundary," *Journal of differential equations*, vol. 37, no. 1, pp. 1–11, 1980.

[109] G. Roussos and K. J. Kyriakopoulos, "Decentralized and prioritized navigation and collision avoidance for multiple mobile robots," in *Distributed Autonomous Robotic Systems*. Springer, 2013, pp. 189–202.

[110] P. E. Rybski, S. A. Stoeter, M. D. Erickson, M. Gini, D. F. Hougen, and N. Papanikolopoulos, "A team of robotic agents for surveillance," in *Proceedings of the fourth international conference on autonomous agents*. ACM, 2000, pp. 9–16.

[111] M. Schmidt, N. L. Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *arXiv preprint arXiv:1309.2388*, 2013.

[112] L. Schwartz, *Théorie des distributions*. Hermann, Paris, 1966.

[113] S. Shalev-Shwartz, "Online learning and online convex optimization," *Foundations and Trends in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2011.

[114] S. E. Shreve and D. P. Bertsekas, "Alternative theoretical frameworks for finite horizon discrete-time stochastic optimal control," *SIAM J. on control and optimization*, vol. 16, no. 6, pp. 953–978, 1978.

[115] S. Smale, "On gradient dynamical systems," *Annals of Mathematics*, vol. 74, no. 1, pp. 199–206, 1961. [Online]. Available: http://www.jstor.org/stable/1970311

[116] S. Sorin, "Exponential weight algorithm in continuous time," *Mathematical Programming*, vol. 116, no. 1-2, pp. 513–528, 2009.

[117] G. S. Sukhatme, A. Dhariwal, B. Zhang, C. Oberg, B. Stauffer, and D. A. Caron, "Design and development of a wireless robotic networked aquatic microbial observing system," *Environmental Engineering Science*, vol. 24, no. 2, pp. 205–215, 2007.

[118] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[119] R. S. Sutton, H. R. Maei, and C. Szepesvári, "A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation," in *Advances in neural information processing systems*, 2009, pp. 1609–1616.

[120] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Adv. in neural information proc. sys.*, 2000, pp. 1057–1063.

[121] M. Takegaki and S. Arimoto, "A new feedback method for dynamic control of manipulators," *Journal of Dynamic Systems, Measurement, and Control*, vol. 103, no. 2, pp. 119–125, 1981.

[122] Y. Tan, W. Moase, C. Manzie, D. Nešić, and I. Mareels, "Extremum seeking from 1922 to 2010," in *Control Conference (CCC), 2010 29th Chinese*. IEEE, 2010, pp. 14–26.

[123] Y. Tan, D. Nešić, and I. Mareels, "On non-local stability properties of extremum seeking control," *Automatica*, vol. 42, no. 6, pp. 889–903, 2006.

[124] H. G. Tanner and A. Kumar, "Formation stabilization of multiple agents using decentralized navigation functions," in *Proceedings of Robotics: Science and Systems*, Cambridge, USA, June 2005.

[125] H. G. Tanner and K. J. Kyriakopoulos, "Nonholonomic motion planning for mobile manipulators," in *IEEE International Conference on Robotics and Automation*, vol. 2. IEEE; 1999, 2000, pp. 1233–1238.

[126] E. Tolstaya, A. Koppel, E. Stump, and A. Ribeiro, "Nonparametric stochastic compositional gradient descent for q-learning in continuous markov decision problems."

[127] H. Uzawa, "Iterative methods for concave programming," *Studies in linear and nonlinear programming*, vol. 6, 1958.

[128] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[129] P. Vincent and Y. Bengio, "Kernel matching pursuit," *Machine Learning*, vol. 48, no. 1, pp. 165–187, 2002.

[130] Y. Viossat and A. Zapechelnyuk, "No-regret dynamics and fictitious play," *Journal of Economic Theory*, vol. 148, no. 2, pp. 825–842, 2013.

[131] C. W. Warren, "Global path planning using artificial potential fields," in *Robotics and Automation, 1989. Proceedings., 1989 IEEE International Conference on.* IEEE, 1989, pp. 316–321.

[132] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[133] H. P. Young, "The evolution of conventions," *Econometrica: Journal of the Econometric Society*, pp. 57–84, 1993.

[134] V. M. Zavala and M. Anitescu, "Real-time nonlinear optimization as a generalized equation," *SIAM Journal on Control and Optimization*, vol. 48, no. 8, pp. 5444–5467, 2010.

[135] D. Zhang and A. Nagurney, "On the stability of projected dynamical systems," *J. Optim. Theory Appl.*, vol. 85, no. 1, pp. 97–124, Apr. 1995. [Online]. Available: http://dx.doi.org/10.1007/BF02192301

[136] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proc. of the twenty-first int. conf. on Machine learning.* ACM, 2004, p. 116.

[137] C. Zhao, U. Topcu, N. Li, and S. Low, "Design and stability of load-side primary frequency control in power systems," *IEEE Transactions on Automatic Control*, vol. 59, no. 5, pp. 1177–1189, 2014.

[138] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *ICML*, 2003, pp. 928–936.