**Stochastic Approximation for Non-expansive
maps:**

Application to Q -Learning Algorithms

by

Jinane Abounadi

Submitted to the Department of Electrical Engineering and Computer
Science

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1998

© Massachusetts Institute of Technology 1998. All rights reserved.

Author
Department of Electrical Engineering and Computer Science
January 9, 1998

Certified by
Dimitri P. Bertsekas
Professor
Thesis Supervisor

Accepted by
Arthur C. Smith
Chair, Department Committee on Graduate Students

Stochastic Approximation for Non-expansive maps: Application to Q -Learning Algorithms

by

Jinane Abounadi

Submitted to the Department of Electrical Engineering and Computer Science
on January 9, 1998, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

We analyze the important class of synchronous and asynchronous variants of fixed point iterations of the form

$$x^{k+1} = x^k + \gamma(k)(F(x^k, \xi^k) - x^k),$$

where F is non-expansive with respect to the sup-norm, $\{\xi^k\}$ is a stochastic sequence, and $\{\gamma(k)\}, \{\xi^k\}$ satisfy the usual stochastic approximation-type of assumptions. These types of iterations fall under the category of stochastic approximation algorithms and can be analyzed using the ODE method based either on Kushner and Clark's Lemma for the synchronous case or Borkar's Theorem for the asynchronous case. However, under this method convergence may be established only if the iterates $\{x^k\}$ are bounded, a fact which is usually difficult to prove. For algorithms which involve non-expansive mappings, we will prove boundedness by developing a novel framework which combines ideas of scaling and properties of Lyapunov functions. The boundedness together with Borkar's stability analysis of ODE's with non-expansive mappings establishes the convergence w.p.1 of the class of algorithms above. An important special subclass of stochastic approximation algorithms with non-expansive mappings is that of Q -learning algorithms, which are simulation-based algorithms for solving dynamic programming problems. The convergence results provide convergence proofs for Q -learning algorithms for discounted cost and stochastic shortest path problems. We also propose two Q -learning algorithms for average cost problems. The first one is based on the relative value iteration algorithm, and we prove its convergence using the ODE method and the general framework developed for algorithms with non-expansive mappings. The second one is based on Bertsekas' relative value iteration algorithm, and we prove its convergence based on the two-time scale asynchronous stochastic approximation ODE analysis of Borkar.

Thesis Supervisor: Dimitri P. Bertsekas
Title: Professor

Acknowledgments

I would like to express my deepest gratitude to my advisor, Professor Dimitri Bertsekas, for his tremendous guidance and support. He has taught me all about Dynamic Programming and Neuro-Dynamic Programming, and the enthusiasm he has for research, as well as the high standards he holds, have had an important impact on my research. I feel very fortunate to have had the opportunity to interact with Professor Vivek Borkar during his visits at LIDS, as my discussions with him have influenced my thesis in a major way. Professor Borkar also served as my thesis reader and provided me with constructive feedback. Professors John Tsitsiklis and Sanjoy Mitter have served on my thesis committee and have been a major source of support during my stay at LIDS. I am grateful to Professor Tsitsiklis for carefully reading my thesis, and making key suggestions that strengthened the results. Professor Mitter has given me valuable insight and direction. I am thankful to my academic advisor, Professor Gallager, for all his advice.

LIDS has provided me with a supportive and stimulating environment. I have benefited tremendously from my interactions with the faculty and the students at LIDS, and in particular I have learned a lot from the Neuro-Dynamic Programming study group led by Professors Bertsekas and Tsitsiklis. I am thankful to my current office-mates P. Marbach, S. Patek, B. Van Roy, and C. Wu, as well as my old office mates, D. Ho, Y. Li, M. Medard, L. Polymenakos, D. Tse, for numerous productive discussions. I also would like to thank Marilyn Pierce, Kathleen O'Sullivan, Amy Briemer, Nancy Young, Fifi Monserrate and Anthony Duggins for all their assistance. Kathleen and Amy have been especially helpful during the last months of my thesis work.

I am largely indebted to various members of my family for their continuous love and support. My parents, Abderrahman and Laaziza Abounadi, have taught me the love of learning from a very young age. My siblings, Nawal, Abdelhadi, Salma, Adil, Rim, Moustapha and Asaad have all been supportive. Abdelhadi's wife, Amal, has sent me numerous encouraging e-mails. Moustapha, my cousin Hisham, Adil, and his wife Becky have all come to help when I needed them. I am especially grateful to my sister Salma for taking wonderful care of my children, Deema and Hilal. I am also thankful to my mother-in-law, Wisam Dahleh, for her help during her visits.

The one person who made it possible for me to complete my Ph.D. thesis is my husband Munzer Dahleh. He has provided me with unconditional loving support in every possible way, and always found a way to give me the right encouragement during the most difficult times. To Munzer, I simply cannot find the right words to express my thanks. Deema and Hilal, have been my source of joy and inspiration.

The research in this thesis was supported by the Center for Intelligent Control Systems grant ARO DAAL03-92-G-0115 and by the National Science Foundation (NSF) Grant DMI-9625489.

Contents

1	Introduction	6
2	Background	12
2.1	Notation and Definitions	12
2.2	Dynamic Programming	14
2.3	Average Cost Problem	22
2.4	Review of Q -Learning Results	29
2.5	Synchronous Stochastic Approximation Algorithms	31
2.6	Asynchronous Stochastic Approximation	41
2.7	Two-time Scale Stochastic Approximation	50
3	Theory for General Framework	57
3.1	Preliminaries	57
3.2	Boundedness Lemmas	58
3.3	Analysis of the Scaled Iteration	61
3.3.1	Analysis of Scaled Iteration - Synchronous	63
3.3.2	Analysis of the Scaled Iteration - Asynchronous	73
3.4	Convergence Theorems	76
3.4.1	Analysis of the Synchronous Iteration	76
3.4.2	Analysis of the Asynchronous Iteration	78
4	Analysis of Discounted Cost and SSP Q-learning	81
4.1	Analysis of SSP Synchronous Q -learning	81

4.2	Analysis of the SSP Asynchronous Q -learning	84
4.3	Summary of Convergence Results	85
5	Q-learning Algorithms for Average Cost Problems	86
5.1	Review of existing proposed algorithms	86
5.2	Average Cost Q -Learning	87
5.3	SSP Average Cost Q -Learning	89
6	Analysis of Average Cost Q-Learning	94
6.1	ODE Analysis	95
6.2	Convergence Theorems	103
6.3	Boundedness for the Synchronous Q -learning algorithm	103
7	SSP Average Cost Q-Learning Analysis	106
7.1	Analysis of SSP1 Algorithm	106
7.2	Analysis of SSP2 Algorithm	108
8	Conclusion	114
A	Mathematical Background	120
B	Asynchronous Relative Value Iteration	123
B.1	Introduction	123
B.2	2-state Markov Chain	123
B.2.1	Value Iteration: Gauss Seidel Version	123
B.2.2	Relaxed Gauss Seidel: GS1	125
B.2.3	Gauss Seidel of Relaxed Jacobi: GS2	126
B.3	3-state Markov Chain	127
B.4	Conclusion	128

Chapter 1

Introduction

Q -learning algorithms have emerged as a powerful simulation tool for solving dynamic programming problems in situations where either the model is not available or the problem must be solved on-line as the data becomes available. Initially, Q -learning was formulated by Watkins [45] in his thesis on reinforcement learning as a model by which certain animals learn based on receiving an immediate reward as well as using past rewards in the form of a learned reward function. Watkins made the connection between this learning model and dynamic programming, and hence formulated a reinforcement learning model as a simulation-based method for solving discounted cost dynamic programming problems. This was very significant as it opened up an interesting field for dynamic programming, as well as an interesting computational method. While the first convergence analysis of the Q -learning algorithm formulated by Watkins is ingenious, it does not provide a general framework for analyzing similar problems. The subsequent papers of Tsitsiklis [40], and of Jaakola, Jordan, and Singh [17] have made the connection between Q -learning algorithms and stochastic approximation. In particular, Q -learning algorithms for discounted cost problems or stochastic shortest path problems are simply the stochastic approximation versions of the well-known value iteration algorithms in dynamic programming. This connection paved the path for a general analysis based on classic stochastic approximation techniques and specific arguments based on contraction or monotonicity properties.

Since the initial interest in Q -learning and the initial results of convergence for

the discounted cost problem and the stochastic shortest path problem, there has been interest in formulating similar algorithms for average cost problems. In fact, Schwartz [34], Singh [36], and Mahadevan [27] have all discussed the need for Q -learning algorithms for average cost problems, and have proposed several algorithms. However, none of them have provided proofs of convergence for these algorithms. In this thesis, we propose two Q -learning algorithms for average cost dynamic programming problems and prove their convergence. The search for a convergence proof led to the development of a general framework for establishing boundedness and proving convergence for a general class of problems. This framework combines ideas from several fields: stochastic approximation analysis via the limiting ODE (Ordinary Differential Equations) technique, asynchronous algorithm analysis, asynchronous stochastic approximation analysis via the limiting ODE technique, and nonlinear analysis of ODE's. Throughout the thesis, we will be considering different notions of convergence: several deterministic notions in Euclidean spaces or function spaces, and the convergence w.p.1 that is usual in stochastic approximation.

The general framework we propose applies to synchronous and asynchronous variants of algorithms of the form

$$x^{k+1} = x^k + \gamma(k)(F(x^k, \xi^k) - x^k), \quad (1.1)$$

with appropriate properties of F , $\{\xi^k\}$, and usual stochastic approximation conditions on $\{\xi^k\}$ and $\{\gamma(k)\}$. The asynchronous model that we choose to use throughout the thesis is based on the formulation in Borkar [10] and is of the following form

$$x_i^{k+1} = x_i^k + \gamma(\nu(k, i)) \left(F_i(x^k, \xi^k) - x_i^k \right) I(i \in Y^k), \quad (1.2)$$

where $\nu(k, i)$ is the number of times component i has been visited by time k , Y^k is the subset of components being updated at time k , and $\{\gamma(k)\}$, $\{Y^k\}$ satisfy reasonable asynchronous stochastic approximation conditions. For the synchronous algorithm above, a powerful technique for analyzing such algorithms is the ODE method given in [21], and for its asynchronous counterpart, a similar technique has been developed

by Borkar [10]. The major idea behind these two techniques is to find a limiting deterministic continuous-time ODE of the stochastic discrete-time processes using interpolation with the appropriate time scaling. Moreover, if the ODE has an asymptotically stable equilibrium point, then the discrete-time iteration converges to this point with probability 1. However, the underlying inconvenience of the ODE limiting techniques is the fact that they require that boundedness be verified apriori. In this thesis we find a methodology for dealing with the boundedness issue. The methodology we propose involves three major steps:

1. Obtaining a related scaled iteration and establishing its convergence.
2. Showing that the original iteration is bounded as a result of the convergence of the scaled iteration.
3. Showing that the boundedness of the original iteration implies convergence by invoking a standard ODE limiting argument.

For each one of the steps above, we will impose appropriate sufficient conditions. Of course, some of these are the usual conditions imposed in stochastic approximation algorithms, such as Lipschitz continuity of the mapping F , the diminishing stepsize, and appropriate conditions on the noise. However, to guarantee boundedness we require that the mapping F be non-expansive with respect to any norm for the synchronous case, and the sup-norm for the asynchronous case. This is a major extension over the existing general stochastic approximation results. The class of fixed-point problems which involve non-expansive mappings is of major importance to various applications. In particular, it includes value iteration algorithms for the different dynamic programming formulations, and hence the methodology outlined above will be used to prove convergence of dynamic programming Q -learning algorithms.

Step 1 of the scheme described above is carried out by choosing the scaling as a function of a Lyapunov surface of an appropriate ODE. Note that we do not need to know the Lyapunov surface itself; all we need to know is that a Lyapunov function for the ODE at hand exists. For that we will use a general converse Lyapunov

Theorem that guarantees the existence of a smooth Lyapunov function if the ODE has an asymptotically stable equilibrium point [47]. Given this scaling scheme, we will be able to show that the scaled iteration has the same deterministic limiting ODE, and hence converges. The argument is similar to the standard limiting ODE argument of Kushner and Clark [21], except that we need to consider a different topology on a different function space, namely the Skorohod topology on $D[0, 1]$ instead of the uniform convergence topology on $C[0, 1]$. Step 2 involves the simple idea of comparing the original iteration and its scaled counterpart and showing that the difference between the two is bounded due to the non-expansiveness of the mapping F . The idea of comparing the two iterations appeared first in Jaakola, Jordan, and Singh [17] in a much more limited setting. Step 3 is a simple application of the convergence results of the ODE limiting arguments as boundedness is already established.

As a last remark, we would like to note that in the stochastic approximation field, the idea of using projection as a way of forcing boundedness is not new. The difference in our approach is that the use of scaling is only a method of proof, and the objective is to establish the boundedness of the original iteration without altering the iterates by forcing them to be bounded. Finally, the thesis has two major objectives:

1. Develop a general framework for analyzing stochastic approximation algorithms involving non-expansive mappings.
2. Formulate Q -learning algorithms for the average cost problem, and prove that they converge w.p.1 to the optimal solution.

The first objective is of important theoretical value as it addresses the issue of boundedness in a new and conceptually intuitive way. It is well known that this has been a major shortcoming to the ODE approach as a tool for analyzing stochastic approximation algorithms. Although the boundedness results obtained rely completely on the non-expansiveness assumption, the line of argument could give insight into other classes of problems and could be used with appropriate modifications to obtain similar important results. The second one is of practical nature as it presents two Q -learning algorithms for the average cost problem, hence extending the existing dis-

counted cost and stochastic shortest path results. How useful Q -learning algorithms have been in practical application is a subject of discussion. In reality, the usage of Q -learning algorithms, as presented here, has not been very popular in practice as it still involves a table look-up representation of the Q values for each state/action pair, which might have prohibitive memory requirements for a major class of problems of interest. This is the usual curse of dimensionality that dynamic programming suffers from. Recently, there has been a boom of experimental research that looks at ways of combining simulation-based dynamic programming algorithms with function approximation. A complete treatment of the results of this field, called Neuro-Dynamic Programming, can be found in [7]. In this approach, instead of storing the Q values in a look-up table, either state aggregation is used to reduce the size of the state space, or a parametrized function would be used to approximate the Q values. This approach has been reported to be successful in an elevator dispatching application [14]. However, analyzing Q -learning algorithms with function approximation is a difficult task in general given the nonlinearity of the problem and the function approximators. The only analytical results in the area of approximate Q -learning are due to Tsitsiklis and Van Roy (see chapter 6 in [7]) and are restricted to two cases:

1. single policy evaluation with linear function approximators,
2. optimal stopping problems with linear function approximators.

Although the focus of this thesis is not approximate Q -learning, it is important to note that using function approximation is essential to the applicability of Q -learning. The goal of this thesis is thus to offer Q -learning algorithms for average cost problems with convergence proofs. Then one can use these with some confidence as the basis of approximate algorithms.

The thesis will be organized as follows: Chapter 2 is a brief background of the models and theories underlying the results of thesis. In particular, it presents Kushner and Clark's limiting ODE results for synchronous algorithms, which are the basis of much of the theory in this thesis. It also adapts Borkar's limiting ODE arguments for asynchronous and two-time scale stochastic approximation algorithms using Kushner

and Clark’s approach. This is especially important as it allows for a uniform line of reasoning for the various proofs of the thesis. Chapter 5 will address the current interest in Q -learning algorithms for average cost problems and review the proposed algorithms and the status of their analysis. Chapter 3 outlines the general framework that we propose in this thesis to deal with the issue of boundedness. Chapter 4 outlines an alternate argument for the analysis of Q -learning algorithms for discounted cost and stochastic shortest path problems using the ODE technique. Kushner and Yin [24] apply the ODE technique to the same problems, but they consider a constrained form of Q -learning algorithms, where the iterates are kept within a bounded interval via projection. They do not address the issue of boundedness in the original algorithms. Chapter 6 outlines the proof of convergence of the proposed Q -learning algorithm for the average cost problem, including the analysis of the ODE corresponding to several variants of the relative value iteration algorithm. The lemmas and theorems in this chapter cover only the synchronous version of the Q -learning algorithm. One of the properties that is essential for establishing boundedness gets violated with asynchronism. We are still looking at ways of including the asynchronous case. Chapter 7 studies a Q -learning algorithm for the average cost problem based on a new relative value iteration algorithm proposed by Bertsekas [4]. The convergence of this algorithm does not follow from the general framework developed in this thesis, but follows directly from the two-time scale stochastic approximation results by Borkar [11], which involve stochastic approximation algorithms with two iterations being updated using two stepsize schedules, with one much faster than the other. Finally, Chapter 8 will summarize the results of the thesis and propose possible extensions and ideas for future work.

Chapter 2

Background

We will first start by introducing the general ideas of dynamic programming, the various formulations, and the computational methods involved. In particular, we will focus on the value iteration algorithms and discuss their important properties. Then we will discuss asynchronous algorithms and stochastic approximation algorithms, and the models and notation will be mostly consistent with the book “Parallel and Distributed Computation” by Bertsekas and Tsitsiklis [5], as these provide the theoretical framework for understanding Q -learning algorithms. We will also describe the Q -learning algorithms for the discounted cost problem and the stochastic shortest path problem formulated by Watkins [45] and review the proof of convergence given by Tsitsiklis [40], which exploits the connection to asynchronous stochastic approximation. Then we will give an overview of the different techniques for analyzing stochastic approximation algorithms, with a primary focus the ODE approach of Kushner and Clark [21] for synchronous stochastic approximation, and the ODE method of Borkar [10] for asynchronous stochastic approximation.

2.1 Notation and Definitions

Let us first introduce the definitions and notation that will be needed.

- Let x be a vector in $\mathbb{R}^n$. The sup-norm of x , denoted by $\|x\|_\infty$ is defined as:

$$\|x\|_\infty = \max_i |x_i|.$$

- Let x be a vector in $\mathbb{R}^n$. The span semi-norm of x , denoted by $\|x\|_s$ is defined as:

$$\|x\|_s = \max_i x_i - \min_i x_i.$$

- Let x be a vector in $\mathbb{R}^n$. The 2-norm of x , denoted by $\|x\|_2$ is defined as:

$$\|x\|_2 = \left(\sum_i x_i^2 \right)^{\frac{1}{2}}.$$

- A map $G : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is *non-expansive* with respect to some norm (or semi-norm) $\|\cdot\|$ if:

$$\|G(x) - G(y)\| \leq \|x - y\| \quad \forall x, y \in \mathbb{R}^n.$$

In the ODE method for analyzing stochastic approximation algorithms, the notion of asymptotic stability plays an important role. We will be using the definition below:

Definition 2.1.1 Let x^* be an equilibrium point of the ODE $\dot{x} = H(x)$, i.e. $H(x^*) = 0$. Let $B(x^*, R)$ denote the ball of radius R centered around x^* . x^* is said to be *asymptotically stable* if and only if for all $\epsilon > 0$, there exists a scalar $R_\epsilon > 0$ such that if $x(0) \in B(x^*, R_\epsilon)$, then $x(t) \in B(x^*, \epsilon)$ for all t , and $x(t)$ converges to x^* .

In the theory of dynamic discrete-time and continuous-time systems, one of the major tools for analyzing the systems is Lyapunov theory. Of course, central to this theory is the idea of Lyapunov function, defined below for ODE systems :

Definition 2.1.2 Let x^* be an equilibrium point of the ODE $\dot{x} = H(x)$, with domain of attraction $\mathcal{D} \in \mathbb{R}^n$. The function $v : \mathcal{D} \rightarrow \mathbb{R}$ is a *Lyapunov function* if

- v is continuous and nonnegative,
- $v(x^*) = 0$ and $v(x) > 0$ for $x \neq x^*$,

- $v(x(t))$ is strictly decreasing as a function of t , where $x(\cdot)$ is a solution of the ODE for any initial condition $x(0) \in \mathcal{D}$ such that $x(0) \neq x^*$.
- If $\{x^n\}$ is a sequence s.t. $\|x^n\| \rightarrow \infty$ as $n \rightarrow \infty$, then $v(x^n) \rightarrow \infty$ as $n \rightarrow \infty$.

For the asynchronous stochastic approximation algorithms, we need to assume that the stepsize $\{\gamma(k)\}$ to be an ‘Ideal Tapering Stepsize’, as defined by Borkar in [10]. We will re-state that definition below:

Definition 2.1.3 *The stepsize $\{\gamma(k)\}$ is said to be an ‘Ideal Tapering Stepsize’ if*

1. $\sum_{k=0}^{\infty} \gamma(k) = \infty$.
2. $\gamma(k+1) \leq \gamma(k)$ for k large enough.
3. There exists $r \in (0, 1)$ such that

$$\sum_{k=0}^{\infty} \gamma^{1+q} < \infty \quad \text{for } q \geq r.$$

4. Let $\Gamma(k) = \sum_{m=0}^k \gamma(m)$, then

$$\begin{aligned} \sup_k \frac{\gamma([rk])}{\gamma(k)} &< \infty \quad \forall r \in (0, 1), \\ \lim_{k \rightarrow \infty} \frac{\Gamma([yk])}{\Gamma(k)} &= 1 \quad \text{uniformly in } y \in [r, 1) \quad \forall r \in (0, 1), \end{aligned}$$

where $[y]$ denotes the integer part of y .

The conditions above are stronger than the usual stochastic approximation conditions, however they are satisfied by most of the common stepsizes used, including $\{\frac{1}{k}\}$, $\{\frac{1}{k} \log(k)\}$, and $\{\frac{\log(k)}{k}\}$.

2.2 Dynamic Programming

Markov Decision Problems (MDP’s) are a natural and powerful way of modeling several classes of problems that arise in major applications, such as queueing, manufacturing, and communication networks. A comprehensive treatment can be found

in [2]. Briefly, an MDP is a controlled Markov chain. At each state several actions are possible, and the probability distribution of the next state depends only on the current state and the action taken at the current time. In general, a policy is a time-dependent mapping from the state space to the control space. The objective is to find the policy that minimizes a certain cost. There are several possible formulations of MDP's, and the theory available is quite extensive. The complexity of the problem depends on the cost to be minimized, on whether the problem is finite-horizon or infinite-horizon, and on whether the state space and control space are finite, countable, or uncountable. The stochastic nature of the problem raises mathematical complications depending on the cost function. In this thesis, we will consider only infinite-horizon stationary MDP's with a finite state space and finite control space. The cost functions will all be in terms of expectations. An MDP is generally defined in terms of the following:

- State space $S = \{1, \dots, n\}$,
- Control space $A = \{1, \dots, m\}$,
- Transition probabilities: $p_{ij}(u)$ denotes the probability that the next state is j given that the current state is i and action u is taken,
- Costs: $g(i, u, j)$ denotes the cost incurred if the current state is i , the action taken is u , and the next state is j . $c(i, u) = \sum_j p_{ij}(u)g(i, u, j)$ is the expected cost at state i if action u is taken.
- Control constraints $U : S \rightarrow 2^A$, where 2^A is the set of all possible subsets of A . $U(i)$ denotes all the possible controls for state i . Without loss of generality, we will assume that $U(i) = A$, for all i , as the control constraints can be accounted for by duplicating controls with the same transition probabilities and costs.

We denote by $\pi = (\mu_0(\cdot), \mu_1(\cdot), \dots)$ a policy, where $\mu_k : S \rightarrow A$ and $\mu_k(i)$ is the action to be taken at state i and time k . We denote the expected infinite-horizon cost starting at state i and for policy π by $J_\pi(i)$. When the policy under consideration is stationary, i.e. $\pi = (\mu(\cdot), \mu(\cdot), \dots)$, it is customary to denote the costs above by

J_μ . Stationary policies are of special interest for stationary problems, as the optimal solutions turn out to be stationary for a large class of problems. In this thesis we will be considering three types of expected costs:

1. Total expected discounted cost: $J_\pi(i) = \lim_{N \rightarrow \infty} \sum_{k=0}^{N-1} E \left[\alpha^k g(i_k, \mu_k(i_k), i_{k+1}) \right]$, where $\alpha \in [0, 1)$ is the discount factor. The expectation is conditional on the initial state being i . The usual assumption here is that the costs per stage are uniformly bounded.
2. Total expected undiscounted cost: $J_\pi(i) = \lim_{N \rightarrow \infty} \sum_{k=0}^{N-1} E [g(i_k, \mu_k(i_k), i_{k+1})]$. The usual assumption imposed here is that there is a zero-cost absorbing state t (termination state) that is eventually reachable from each state w.p.1 under some policy. Policies such that the termination state is reachable from each state w.p.1 are called proper. These problems are labeled stochastic shortest path problems. For convergence results, the least restrictive condition is that there exists a proper policy, and that the cost is infinite for at least one state under all improper policies.
3. Average cost per stage: $J_\pi = \liminf_{N \rightarrow \infty} \frac{1}{N} E \left[\sum_{k=0}^{N-1} g(i_k, \mu_k(i), i_{k+1}) \right]$. The usual assumption imposed here is that the Markov chain corresponding to each stationary policy is unichain. The average cost per stage then for each policy is independent of the initial condition, and so is the optimal average cost per stage.

Several analytic questions arise in the definitions above, such as the existence of the limits and a precise definition of the conditional expectations. A complete study of the infinite horizon problems can be found in [2]. In general, under some reasonable assumptions, the limits above are well defined and solutions to all the problems above involve optimal stationary policies and are characterized via appropriate Bellman's equations. In addition several numerical methods have been proposed to solve Bellman's equations and derive the optimal policies and the optimal value functions, also referred to as optimal cost-to-go functions. The method that has been the basis for

Q -learning algorithms is what is referred to as value iteration, which is a successive approximation algorithm for fixed-point problems. Therefore it is essential to have a complete understanding of the value iteration algorithms for the various dynamic programming formulations above, and understand the properties that allow them to converge.

As shown in [2], the optimal solutions for the three problems described above satisfy Bellman's equations, from which both optimal cost-to-go functions and policies may be derived. We will briefly review the different Bellman's equations as they are the fundamental theoretical basis for all the Dynamic Programming algorithms. For convenience of notation, it is usual to define mappings T and T_μ from $\mathbb{R}^n$ to $\mathbb{R}^n$ for the discounted cost problem as follows:

$$(TJ)(i) = \min_{u \in A} \left[\sum_j p_{ij}(u) (g(i, u, j) + \alpha J(j)) \right] \quad \forall i \in S,$$

and

$$(T_\mu J)(i) = \sum_j p_{ij}(\mu(i)) (g(i, \mu(i), j) + \alpha J(j)) \quad \forall i \in S.$$

For the stochastic shortest path problem, the T and T_μ mappings are defined as follows:

$$\begin{aligned} (TJ)(i) &= \min_{u \in A} \left[\sum_j p_{ij}(u) (g(i, u, j) + J(j)) \right] & \forall i \neq t, \\ J(t) &= 0, \end{aligned}$$

and

$$\begin{aligned} (T_\mu J)(i) &= \sum_j p_{ij}(\mu(i)) (g(i, \mu(i), j) + J(j)) & \text{for } i \neq t, \\ J(t) &= 0, \end{aligned}$$

where t is the zero-cost termination state.

It can be shown for both problems that the cost J_μ for a single stationary policy μ satisfies

$$J_\mu = T_\mu(J_\mu). \tag{2.1}$$

Similarly, the optimal cost-to-go function J^* satisfies

$$J^* = T(J^*). \quad (2.2)$$

Both equations (2.1) and (2.2) are referred to as Bellman's equations for the policy evaluation problem and the optimal policy problem, respectively. For both the discounted cost and stochastic shortest path problems, under the assumptions above, it can be shown [2] that:

1. Bellman's equation has a unique solution J^* ,
2. The optimal policy is stationary, and if μ^* is optimal, then

$$T_{\mu^*} J^* = J^*,$$

3. The successive approximation algorithm

$$J^{k+1} = T J^k \quad (2.3)$$

converges to J^* for any initial choice of J^0 . In the context of dynamic programming, the algorithm above is customarily referred to as value iteration.

The ideas above are the basis for the theory of dynamic programming for infinite-horizon problems, and although we will not go through the arguments here ([2] covers this topic in completeness), we will list the properties of the T mappings that are essential to these arguments. Note that the same properties hold for the T_μ mapping as the cost evaluation of a single policy may be viewed as a special case of the optimization problem.

1. Monotonicity: for both problems, T is monotone, i.e. given $J, J' \in \mathbb{R}^n$,

$$J \leq J' \implies T J \leq T J'.$$

2. Dynamic programming-type property: For any scalar $r > 0$,

$$T(J + re) \leq TJ + re,$$

and

$$T(J - re) \geq TJ - re.$$

3. Unique fixed point property: There is a unique J^* s.t.

$$TJ^* = J^*.$$

4. Contraction: For the discounted cost problem, the T mapping is contractive with respect to the sup norm. For the stochastic shortest path problem, it can be shown that T is a contraction with respect to a weighted sup-norm if all the policies are proper.

Originally, the dynamic programming formulation has been in terms of the cost function, which is a function of the states. In the more recent Q -learning work, it turns out that it is convenient to use ‘quality factors’, which are a function of state-action pairs. In the following section, we will introduce these Q factors and derive other variants of Bellman’s equations for all the problems described above in terms of the Q factors. First, let us consider the stochastic shortest path problem without loss of generality, since the discounted cost problem can be formulated as a stochastic shortest path problem, as shown in [2]). In fact, a discounted cost problem with a discount factor α can be transformed into a stochastic shortest path problem by simply adding an artificial zero-cost termination state t and setting the transition probabilities p_{ij} to αp_{ij} for all pairs (i, j) and letting $p_{it} = (1 - \alpha) \sum_j p_{ij}$ for all i . The transformed problem is a SSP problem where all the policies are proper. Moreover, Bellman’s equations for the two problems are exactly the same, and the resulting value iteration algorithms are also the same. Therefore, from now on we will formulate all

the results for SSP problems only. Define the vector $Q^* \in \mathbb{R}^{n+m}$ as follows:

$$Q^*(i, u) = \sum_j p_{ij}(u) (g(i, u, j) + J^*(j)) \quad \forall (i, u) \in S \times A. \quad (2.4)$$

This identity, together with Bellman's equation for multiple policies (2.2), implies that

$$J^*(i) = \min_u Q^*(i, u) \quad \forall i \in S. \quad (2.5)$$

Combining equations the two equations above, we get that Q^* satisfies:

$$Q^*(i, u) = \sum_{j \neq t} p_{ij}(u) \left(g(i, u, j) + \min_v Q^*(j, v) \right) \quad \forall (i, u) \in S \times A.$$

$Q^*(i, u)$ may be viewed as the expected cost if action u is taken at state i and the optimal policy is observed from then on. Thus the optimal Q values and the optimal policy may be obtained by solving the following form of Bellman's equation:

$$Q(i, u) = \sum_j p_{ij}(u) g(i, u, j) + \sum_j p_{ij}(u) \min_{u'} Q(j, u') \quad \forall (i, u) \in S \times A. \quad (2.6)$$

The equation above has the same properties as Bellman's equation given by (2.2) and can also be solved using the successive approximation algorithm

$$Q^{k+1}(i, u) = \sum_{j \neq t} p_{ij}(u) g(i, u, j) + \sum_{j \neq t} p_{ij}(u) \min_{u'} Q^k(j, u') \quad \forall (i, u) \in S \times A. \quad (2.7)$$

Both forms of the dynamic programming value iteration given by (2.3) and (2.7) involve computing expectations and minima, but the algorithm given by (2.7) has the advantage that the minimum function is inside the expectation. That allows for natural use of simulation methods. For example, the expectation may be estimated by taking several samples of the next state and corresponding immediate costs and averaging them. Stochastic approximation provides an incremental way of carrying out the averaging by introducing a diminishing stepsize and using a single sample in lieu of the expectation. The Robbins-Monro stochastic approximation version of the

algorithm above yields the following synchronous algorithm

$$Q^{k+1}(i, u) = (1 - \gamma(k))Q^k(i, u) + \gamma(k) \left[g(i, u, j(i)) + \min_{u'} Q^k(j(i), u') \right] \quad \text{for all } (i, u) \in S \times A, \quad (2.8)$$

where $j(i)$ is a state obtained via simulation according to the probability distribution $\{p_{ij}(u)\}$. At the termination state t , there is only one possible action u and $Q^k(t, u) = 0$ for all k . The assumption imposed on $\{\gamma(k)\}$ is the usual stochastic approximation assumption

$$\begin{aligned} \sum_{k=0}^{\infty} \gamma(k) &= \infty, \\ \sum_{k=0}^{\infty} \gamma^2(k) &< \infty. \end{aligned}$$

The Q -learning algorithms, proposed originally in Watkins' thesis [45] for discounted cost and stochastic shortest path problems, may be viewed as Gauss-Seidel asynchronous versions of the algorithm above, whereby only one state-action pair is updated at a time. One possible model for stochastic shortest path Q -learning algorithms is

$$Q^{k+1}(i, u) = (1 - \gamma(k, i, u))Q^{k+1} + \gamma(k, i, u) \left[g(i, u, j(i)) + \min_{u'} Q^k(j(i), u') \right] I((i, u) \in Y^k), \quad (2.9)$$

where Y^k is a set-valued random variable, which contains the set of state-action pairs which are to be updated at time k , and the stepsize satisfies the following assumption with probability 1 for all $(i, u) \in S \times A$,

$$\begin{aligned} \sum_{k=0}^{\infty} \gamma(k, i, u) &= \infty, \\ \sum_{k=0}^{\infty} \gamma^2(k, i, u) &< \infty. \end{aligned}$$

The sequence by which the state-action pairs are updated in principle is not important, as long as a certain frequency condition is satisfied. In on-line learning applications, it is natural to follow a simulation trajectory, but issues of sampling over the entire state space pose serious challenges in practice. We will discuss these algorithms later in Section 2.4 when we review the Q -learning results for discounted

cost problems and stochastic shortest path problems. It is also important to note at this point that the asynchronous model above is more general than the model (1.2) addressed in the thesis in that it allows for a more general stepsize. This is natural as the contraction-type properties of discounted cost and stochastic shortest path problems permit more freedom in the choice of the stepsize, while to obtain general convergence results for the entire class of non-expansive mappings it will be necessary to require that the stepsize be a function of the frequency of updates as we will see in Section 2.6.

2.3 Average Cost Problem

Given that the focus of the thesis is the average cost problem, we will go into more detail in reviewing background regarding average cost problems. The dynamic programming derivation of Bellman's equations and the corresponding analysis is more complicated than that of the stochastic shortest path problem. To get a better understanding of the problem, we will first start by considering the single stationary policy case, where the objective is to evaluate the average cost per stage for the policy. Only one action is allowed at each state, and there are no control decisions to be made. This case simplifies to the study of Markov chains, and the computational methods involve linear algebra manipulations, in particular with stochastic matrices. For the single policy case, the problem is simply to compute the average cost per stage of a Markov chain with transition probabilities $\{p_{ij}\}$ and transition costs $\{g(i, j)\}$. The average cost for state i is then

$$J(i) = \lim_{N \rightarrow \infty} \frac{1}{N} E \left[\sum_{k=0}^{N-1} g(x^k, x^{k+1}) | x^0 = i \right],$$

where x^k is the state at time k , and the expectation is conditional on the initial state being i . It turns out that the solution can be given completely in terms of the steady state probabilities. Under the assumption that the Markov chain is unichain, i.e. all the states are either transient or belong to the same recurrent class, the steady state

probabilities are independent of the initial state and satisfy an important equilibrium equation. Let $\pi(i)$ denote the steady state probability of being at state i . Then it is well known in the theory of Markov chains that

$$\pi = \pi P,$$

where $\pi = [\pi(1), \dots, \pi(n)]$ is a row vector, and P is the transition probability matrix. It can then be shown that the average cost per stage is also independent of the initial state and is given by

$$\lambda_\pi = \pi c, \tag{2.10}$$

where $c = [c(1), \dots, c(n)]'$ is a column vector with $c(i) = \sum_j p_{ij}g(i, j)$. Based on the solution of the steady state probabilities (2.10), it is easy to derive a Bellman-type equation for the single policy problem [12]. It is given by the following linear system of equations

$$h = c + Ph - \lambda e, \tag{2.11}$$

where e is a vector in $\mathbb{R}^n$ whose entries all equal 1. $h \in \mathbb{R}^n$ is interpreted as a relative-cost vector, in the sense that $h(i) - h(j)$ is the asymptotic difference between the total undiscounted costs $J_N(i) = \sum_{k=0}^{N-1} E[g(x^k, x^{k+1}) | x^0 = i]$ and $J_N(j) = \sum_{k=0}^{N-1} E[g(x^k, x^{k+1}) | x^0 = j]$. Using basic ideas in linear algebra, it can be shown that a solution (h_π, λ_π) to (2.11) exists, where λ_π is the average cost of the policy given by equation (2.10). Moreover, (h, λ) is a solution if and only if

1. $\lambda = \lambda_\pi,$

2. there exists $r \in \mathbb{R}$ such that $h = h_\pi + re,$

i.e. the solution set is $\{(h, \lambda) = (h_\pi + re, \lambda_\pi) : r \in \mathbb{R}\}$. Based on the linear form of the solution set above, it is easy to see that adding the linear constraint $h(s) = \lambda_\pi$, where s is an arbitrary state will make the solution unique. That results in a new system of n linear equations with n unknowns:

$$(I - P + \tilde{P})h = c, \tag{2.12}$$

where

$$\tilde{P} = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ \vdots & & & \vdots \\ 0 & & 0 & 1 \end{bmatrix}.$$

The solution to the linear system of equations above may be obtained iteratively via the successive approximation algorithm

$$h^{k+1} = c + (P - \tilde{P})h^k.$$

This algorithm is of course a special case of the relative value iteration algorithm for the optimal control problem, and hence its convergence follows from the analysis of relative value iteration [2], which will be discussed later in this section. However, a direct analysis may be carried out using linear algebra, giving an insight to the problem.

In summary, the single policy case is rather intuitive and follows directly from the study of the steady state probabilities of Markov chains. The numerical convergence issues are resolved using arguments from linear algebra. For the multiple policy case, both the derivations of Bellman's equation and the analysis of appropriate algorithms are more complex. In the following formulation of average cost Markov Decision Problems, we will make the assumption that all policies are unichain. As for the discounted cost and SSP problems the solution to the average cost problem is given by Bellman's equation. A good treatment of the subject can be found in [2] and [32]. The solution is in general in terms of a relative cost vector h^* and the optimal average cost λ^* . The relative values can be interpreted as follows: $h^*(i) - h^*(j)$ is the asymptotic difference between the total costs of the optimal policy, starting at states i and j , respectively. h^* and λ^* satisfy Bellman's equation given by

$$\lambda + h(i) = \min_u \left[\sum_{j=1}^n p_{ij}(u) (g(i, u, j) + h(j)) \right] \quad \forall i \in S. \quad (2.13)$$

The equation above is a nonlinear equation with n equations and $n + 1$ unknowns. The solution has the same characteristics as the solution for the single policy problem. It can be shown that a solution (h^*, λ^*) to the system of equations above exists. Moreover, (h, λ) is a solution if and only if

1. $\lambda = \lambda^*$,
2. there exists $r \in \mathbb{R}$, such that $h = h^* + re$,

i.e. the solution set is $\{(h, \lambda) = (h^* + re, \lambda^*) : r \in \mathbb{R}\}$. Again, adding the constraint $\lambda = h(s)$, where s is an arbitrary state makes the solution unique and results in the following fixed point system of n equations with n unknown variables.

$$h(i) = \min_u \left[\sum_{j=1}^n p_{ij}(u) (g(i, u, j) + h(j)) \right] - h(s) \quad \forall i \in S.$$

The corresponding successive approximation algorithm that is based on the fixed point equation (2.13) results in the average cost relative value iteration algorithm

$$h^{k+1}(i) = \min_u \left[\sum_{j=1}^n p_{ij}(u) (g(i, u, j) + h^k(j)) \right] - h^k(s) \quad \forall i \in S.$$

The convergence of the relative value iteration algorithm above requires some additional technical assumptions involving aperiodicities of the transition probability matrices (see [2]). However, a relaxed version of the algorithm which involves a small stepsize converges without the additional technical assumption. The relative value iteration above is the basis for the proposed Q -learning algorithm for the average cost problem.

As for the discounted cost and stochastic shortest path problems, there is an equivalent formulation of Bellman's equation in terms of the Q -values of the state-action pairs. Define a related Markov decision process in the following way: Consider an expanded state space $S' = S \cup \{(i, u) : i \in S, u \in A\}$. Define the transition probabilities and costs on the states in S in the same way as in the original problem. On the new states (i, u) , only one control is possible and transitions are allowed only

to states j in S with transition probabilities $p_{ij}(u)$ and costs $g(i, u, j)$. The states that are artificially added are all transient and therefore do not affect the average cost, and hence this modified problem has the same solution as the original one. Denote by $Q(i, u)$ the differential cost associated with state (i, u) , and by $h(i)$ the differential cost associated with state i . Then, Bellman's equation for the new MDP is given by,

$$\begin{aligned}\lambda + Q(i, u) &= \sum_{j=1}^n p_{ij}(u) (g(i, u, j) + h(j)), & \forall (i, u) \in S \times A, \\ \lambda + h(i) &= \min_u \sum_{j=1}^n p_{ij}(u) (g(i, u, j) + h(j)) & \forall i \in S.\end{aligned}$$

The equations above can be formulated in terms of the Q -values only as follows:

$$\begin{aligned}\lambda + Q(i, u) &= \sum_{j=1}^n p_{ij}(u) (g(i, u, j) + \min_{u'} Q(j, u')) & (i, u) \forall \in S \times A, \\ h(i) &= \min_u Q(i, u) & \forall i \in S.\end{aligned}$$

From the properties of Bellman's equations for the average cost problem, it is well known (see Section 4.2 in [2]) that Bellman's equation has a solution if and only if $\lambda = \lambda^*$. Then the set of solutions is $G = \{Q^* + re : r \in \mathbb{R}\}$, where Q^* is a particular solution. Applying the relative value iteration for average cost problems to the modified MDP yields the following algorithm:

$$Q^{k+1}(i, u) = \sum_{j=1}^n p_{ij}(u) \left(g(i, u, j) + \min_{u'} Q^k(j, u') \right) - Q^k(s, v)e, \quad (2.14)$$

where (s, v) is an arbitrarily chosen fixed state-action pair, and e is the vector in $\mathbb{R}^{n+m}$ whose components all equal 1. In fact, replacing the last term $Q^k(s, v)$ by any scalar function $f(Q^k)$ that satisfies some reasonable conditions gives rise to several variants of the Q -learning algorithm of the form

$$Q^{k+1}(i, u) = \sum_{j=1}^n p_{ij}(u) \left(g(i, u, j) + \min_{u'} Q^k(j, u') \right) - f(Q^k)e. \quad (2.15)$$

Other possible choices for the function f , are $f(Q) = \min_v Q(s, v)$, or $f(Q) = \frac{1}{n} \sum_i \min_v Q(i, v)$. We will see later other possibilities for f that will also work. As for the analysis of the algorithm above, the same reasoning can be used as for the

original relative value iteration algorithms. In fact, as only transient states are added to the modified problem, the unichain assumption is still preserved. Under the additional technical assumption that the Markov chains corresponding to every policy are aperiodic, the iteration in (2.14) converges, and the optimal average cost is given by:

$$\lambda^* = \lim_{k \rightarrow \infty} f(Q^k).$$

Again, the iteration given by equation (2.15), based on the Q -values, will be used as the basis for a stochastic approximation algorithm, since the minimum expression is inside the expectation. Before we proceed with the Q -learning algorithm, we will first establish some notation and present properties that will be useful for the later analysis. Define the T -mapping as follows:

$$(TQ)(i, u) = \sum_{j=1}^n p_{ij}(u) \left(g(i, u, j) + \min_{u'} Q(j, u') \right), \quad (2.16)$$

and define the T' -mapping as:

$$T'(Q) = TQ - f(Q)e. \quad (2.17)$$

The following properties of the mappings above are well known, and they will play a major role in the convergence analysis of the Q -learning algorithms that we propose.

1. T is non-expansive with respect to the sup-norm, i.e.

$$\|T(Q_1) - T(Q_2)\|_\infty \leq \|Q_1 - Q_2\|_\infty \quad \forall Q_1, Q_2 \in \mathbb{R}^{n+m},$$

where $\|Q\|_\infty = \max_{(i,u)} |Q(i, u)|$.

2. T and T' are non-expansive with respect to the span semi-norm $\|\cdot\|_s$, i.e.

$$\|T(Q_1) - T(Q_2)\|_s \leq \|Q_1 - Q_2\|_s \quad \forall Q_1, Q_2 \in \mathbb{R}^{n+m},$$

where $\|Q\|_s = \max_{(i,u)} Q(i, u) - \min_{(i,u)} Q(i, u)$.

3. T satisfies:

$$T(Q + re) = TQ + re \quad \forall r \in \mathbb{R}.$$

To see that property 1 holds, let $Q_1, Q_2 \in \mathbb{R}^{n+m}$, and let $u_l(j)$ be the action that minimizes $Q_l(j, v)$ for $l = 1, 2$. Then for all (i, u) ,

$$\begin{aligned} (TQ_1)(i, u) - (TQ_2)(i, u) &= \sum_j p_{ij}(u) \left[\min_{u'} Q_1(j, u') - \min_v Q_2(j, v) \right], \\ &\leq \sum_j p_{ij}(u) [Q_1(j, u_2(j)) - Q_2(j, u_2(j))], \\ &\leq \|Q_1 - Q_2\|_\infty. \end{aligned}$$

Similarly,

$$\begin{aligned} (TQ_2)(i, u) - (TQ_1)(i, u) &= \sum_j p_{ij}(u) \left[\min_v Q_2(j, v) - \min_{u'} Q_1(j, u') \right], \\ &\leq \sum_j p_{ij}(u) [Q_2(j, u_1(j)) - Q_1(j, u_1(j))], \\ &\leq \|Q_1 - Q_2\|_\infty. \end{aligned}$$

Similarly, to see that property 2 holds, let $Q_1, Q_2 \in \mathbb{R}^{n+m}$, and let $u_l(j)$ be the action that minimizes $Q_l(j, v)$ for $l = 1, 2$. Then for all (i, u) ,

$$\begin{aligned} \max_{(i, u)} \sum_j p_{ij}(u) \left[\min_{u'} Q_1(j, u') - \min_v Q_2(j, v) \right] \\ &\leq \max_{(i, u)} \sum_j p_{ij}(u) [Q_1(j, u_2(j)) - Q_2(j, u_2(j))] \\ &\leq \max_{(i, u)} (Q_1(i, u) - Q_2(i, u)). \end{aligned}$$

Similarly,

$$\begin{aligned} \min_{(i, u)} \sum_j p_{ij}(u) \left[\min_{u'} Q_1(j, u') - \min_v Q_2(j, v) \right] \\ &\geq \min_{(i, u)} \sum_j p_{ij}(u) [Q_1(j, u_1(j)) - Q_2(j, u_1(j))] \\ &\geq \min_{(i, u)} (Q_1(i, u) - Q_2(i, u)). \end{aligned}$$

Therefore,

$$\|TQ_1 - TQ_2\|_s \leq \|Q_1 - Q_2\|_s.$$

Also, $\|T'(Q_1) - T'(Q_2)\|_s = \|TQ_1 - TQ_2\|_s$, therefore

$$\|T'Q_1 - T'Q_2\|_s \leq \|Q_1 - Q_2\|_s.$$

Property 3 follows from simple algebra.

Property 4 is a general property of the dynamic programming operators, and can be easily verified. For all (i, u) ,

$$\begin{aligned} (T(Q + re))(i, u) &= \sum_j p_{ij}(u) \left(g(i, u, j) + \min_{u'} (Q(j, u') + r) \right), \\ &= \sum_j p_{ij}(u) \left(g(i, u, j) + \min_{u'} Q(j, u') \right) + \sum_j p_{ij}(u)r, \\ &= \sum_j p_{ij}(u) \left(g(i, u, j) + \min_{u'} Q(j, u') \right) + r, \\ &= (TQ)(i, u) + r. \end{aligned}$$

The formulation of the relative value iteration algorithm in terms of the Q values will be used as the basis for the Q -learning algorithm. We will see later that the properties above will play an essential role in the algorithm analysis.

2.4 Review of Q -Learning Results

In this section, we will first give a historical view of the lines of argument used to analyze Q -learning algorithms. These basically include the analyses by Watkins [45], Jaakola et al. [17], and Tsitsiklis [41]. Watkins approached the problem from the field of reinforcement learning in psychology, whereby an animal is trained by receiving immediate rewards as a function of behavior. He came up with a model of reinforcement learning based on a simulation variant of value iteration algorithms in dynamic programming and therefore provided a solid basis for a clear formulation and analysis. The analysis results of Watkins and Dayan [46] cover discounted cost problems and stochastic shortest path problems under the assumption that all policies are proper. Although the line of argument they used cannot be generalized to other problems, the connection between reinforcement learning and dynamic programming opened the field of reinforcement learning to the community of dynamic

programming, under the leadership of Bertsekas and Tsitsiklis. In fact, subsequent to the work of Watkins, Tsitsiklis [40] recognized that the analysis of the Q -learning algorithms is best handled by the techniques of stochastic approximation. In fact, the Q -learning algorithms are simply Robbins-Monro versions of asynchronous value iteration algorithms, where the expectations are replaced by one-time samples. The existing stochastic approximation results at the time were not adequate for the analysis of Q -learning for discounted cost and stochastic shortest path problems. The properties that turn out to play a major role in the analysis of the Q -learning algorithms are the same as the ones used in the analysis of the value iteration algorithms, and these are specifically the monotonicity, dynamic programming-type, fixed point, or sup-norm contraction properties discussed in Section 2.2. Motivated by the objective of analyzing Q -learning algorithms, Tsitsiklis [40] studied the convergence of stochastic approximation algorithms which involved either sup-norm contractions or monotonicity along with the dynamic programming-type and fixed point properties. The convergence theorems in Tsitsiklis [40] apply to both the synchronous and asynchronous algorithms introduced in Chapter 1, which may be rewritten as

$$x^{k+1} = x^k + \gamma(k)(T(x^k) + w^k - x^k),$$

and

$$x_i^{k+1} = x_i^k + \gamma(\nu(k, i)) (T_i(x^k) + w_i^k - x_i^k) I(i \in Y^k),$$

respectively, for appropriately defined T , and $\{w^k\}$. In [40], proofs of the convergence results in [40] involve two major parts:

1. Proving that the iterates are bounded w.p.1.
2. Proving convergence to the fixed point w.p.1 of the discrete algorithm if the iterates are bounded.

The boundedness proof is interesting and insightful, but applies only to problems that involve sup-norm contractions and hence could not be applied to Q -learning algorithms for all stochastic shortest path problems. The summary of the results is

as follows:

1. If T is a contraction with respect to the sup-norm, then $\{x^k\}$ converges to the unique fixed point of T , w.p.1.
2. If T is monotone, has a unique fixed point x^* , and satisfies the dynamic programming-type inequality of Section 2.2, then $\{x^k\}$ converges to x^* w.p.1 if it is bounded.

The convergence results above are then directly applicable to Q -learning algorithms for discounted cost and stochastic shortest path problems. Recall from Section 2.2 that discounted cost and stochastic shortest path problems with only proper policies both satisfy the sup-norm contraction property. Therefore, the corresponding Q -learning algorithms converge w.p.1 to the optimal solution, which is the unique solution to Bellman's equations. For the more general class of stochastic shortest path problems, allowing improper policies with infinite costs, the convergence requires the iterates to be bounded. The boundedness proof in [40] shows the difficulty that this problem poses in general, and is to be contrasted with the boundedness proof that the general framework in this thesis provides.

2.5 Synchronous Stochastic Approximation Algorithms

In this section, we will briefly review basic ideas in stochastic approximation, with the objective of establishing the notation, models, and terminology. The field of stochastic approximation started with the pioneering work of Robbins and Monro. The problem was to find an iterative method for finding the root x^* of the equation

$$E[H(x, \xi)] = 0, \tag{2.18}$$

where ξ is a random variable whose probability distribution depends on the parameter x and is denoted by $p_x(\xi)$. At each time step, x^k is available, and a noisy observation of $E[H(x^k, \xi)]$ is made. The algorithm is to compute an updated estimate x^{k+1} of

x^* , and the objective is to have the sequence of estimates $\{x^k\}$ converge to x^* with probability 1. Typically, when the probability distribution of ξ is not known, the noisy observation is simply a sample value. One possible algorithm, based on deterministic numerical algorithms for solving regression equations of the type (2.18), is

$$x^{k+1} = x^k + \gamma(k)(H(x^k, \xi^k)),$$

where $\gamma(k)$ is a time-varying stepsize. Classical stochastic approximation techniques study the conditions under which the stochastic sequence $\{x^k\}$ converges to x^* w.p.1. The stepsize is usually assumed to satisfy the following standard conditions:

Assumption 2.5.1

$$\begin{aligned}\sum_{k=0}^{\infty} \gamma(k) &= \infty, \\ \sum_{k=0}^{\infty} \gamma^2(k) &< \infty.\end{aligned}$$

The second condition forces the sequence $\{\gamma(k)\}$ to decrease to 0, and is used to counteract the noise variation. The first condition insures that the algorithm does not stop midway before it converges to the right value. A very special case of the problem above arises when $H(x, \xi) = \xi - x$, the observations are independent samples $\{\xi^k\}$ of the random variable ξ , and $\gamma(k) = \frac{1}{k+1}$. The problem is then equivalent to finding an iterative scheme for estimating the mean of the random variable ξ based on independent samples. Explicitly the algorithm is

$$x^{k+1} = \left(1 - \frac{1}{k+1}\right)x^k + \frac{1}{k+1}\xi^k.$$

It can be shown by induction that $x^k = \frac{\sum_{l=0}^k \xi^l}{k+1}$, and is therefore the sample mean. The w.p.1 convergence of x^k to x^* follows from the Strong Law of Large Numbers (see [8]). The same result can be derived using stochastic approximation theorems, which follow from the convergence theorems of martingales and submartingales ([8]). These are in turn stochastic analogs of convergence theorems of bounded and monotone deterministic sequences. We will briefly review these theorems in the thesis.

The problem of finding the roots of equation (2.18) covers a large class of problems, including fixed point problems and minimization problems. For minimization problems, the mapping H is generally related to the gradient of the function to be minimized, and these problems were initiated by Kiefer and Wolfowitz and are referred to as Kiefer-Wolfowitz problems. We will not address this class of problems in this thesis.

Of course, the focus of this thesis is on fixed point problems. In this case, if the objective is to find a fixed point of $E[F(x, \xi)]$, then H is defined as $H(x, \xi) = F(x, \xi) - x$, and the algorithm becomes

$$x^{k+1} = (1 - \gamma(k))x^k + \gamma(k)F(x^k, \xi^k).$$

It is usual to denote the noise term by $w^k = F(x^k, \xi^k) - E[F(x^k, \xi)]$, where the expectation is with respect to the probability distribution $p_{x^k}(\xi)$. This leads to a more general form of stochastic approximation algorithms for fixed-point problems:

$$x^{k+1} = (1 - \gamma(k))x^k + \gamma(k)(T(x^k) + w^k), \quad (2.19)$$

where T is a function from $\mathbb{R}^n$ to $\mathbb{R}^n$, and $\{w^k\}$ is a sequence of random variables satisfying $E[w^k | x^k, \dots, x^0, w^{k-1}, \dots, w^0] = 0$. The classical stochastic approximation convergence results typically require $\{w^k\}$ to be zero-mean with bounded conditional variance, and other strong conditions, such as the existence of a smooth Lyapunov function, contraction or monotonicity properties of T . An extensive account of these results can be found in Chapter 4 of the Neuro-dynamic Programming book by Bertsekas and Tsitsiklis [5]. Conceptually, the arguments involve a deterministic-type analysis of nonlinear iterative schemes and a stochastic-type analysis of general martingales. The analysis of nonlinear iterative schemes is handled via different techniques of nonlinear programming

Another totally different approach to analyzing an iteration of the form (2.19) is the ODE technique, which was initially proposed by Ljung [26], and its goal is to study the limit of the discrete iterates $\{x^k\}$ as a function of a suitable continuous

time ODE. This idea was later taken up by Kushner and Clark [21]. They considered interpolated processes on the basis of the iterates $\{x^k\}$ and a decreasing time scale (directly related to the decreasing stepsize), and looked at a family of shifted versions of these processes. Clearly, the limiting process, if it exists, would be closely related to the limiting behavior of the discrete iteration, and if it is possible to characterize it more easily, then it's possible to deduce the limiting behavior of the stochastic discrete iteration. In this case, the existence of a limiting process is guaranteed due to the compactness of this family of processes with respect to the $C[0, 1]$ topology, which is the topology of continuous functions over bounded intervals. The compactness is shown using the Arzela-Ascoli theorem. Simple limiting arguments show that the limiting process is a solution to the ODE $\dot{x} = T(x) - x$. Further reasoning shows that the iterates $\{x^k\}$ converge to the limit point of the ODE if it has a global asymptotically stable equilibrium point. The results above impose few assumptions on the function T and very general assumptions on the noise process $\{w^k\}$. Typically the function T is required to be Lipschitz continuous and $\{w^k\}$ is required to satisfy a condition of the form:

$$\lim_{n \rightarrow \infty} P \left(\sup_{m \geq n} \left| \sum_{k=n}^m \gamma(k) w^k \right| \geq \epsilon \right) = 0,$$

for all $\epsilon > 0$. The condition above is satisfied by general classes of problems, including the cases where the $\{w^k\}$ is a sequence of independent zero-mean random variables with bounded variance, or the noise is Markovian with zero-mean and bounded variance. However, the serious difficulty with the ODE method lies in the fact that compactness requires the iterates to be bounded, which is a condition that in general cannot be verified apriori. We will state here a version of the Kushner and Clark lemma (see Theorem 2.3.1 in [21]), particular to the formulation and assumptions of this thesis. The lemma considers the synchronous model below

$$x^{k+1} = x^k + \gamma(k)(F(x^k, \xi^k) - x^k), \quad (2.20)$$

where F is a map into $\mathbb{R}^n$ defined the appropriate, and $\{\xi^k\}$ is a a sequence of independent identically distributed random variables, defined over a probability space $(\Omega, \mathcal{F}, \mathcal{P})$. This assumption on $\{\xi^k\}$ is not necessary for the results below and can be easily generalized. However, it covers a large class of problems and is sufficient for the Q -learning formulation.

It will be more convenient to state the results in terms of the following more general model:

$$x^{k+1} = x^k + \gamma(k)(T(x^k) + w^k - x^k), \quad (2.21)$$

where T is a map from $\mathbb{R}^n$ to $\mathbb{R}^n$, and $\{w^k\}$ is stochastic process defined over some probability space $(\Omega, \mathcal{F}, \mathcal{P})$. The model of (2.20) is a special case of (2.21) with

$$\begin{aligned} T(x^k) &= E[F(x^k, \xi^k) | \mathcal{F}^k], \\ w^k &= F(x^k, \xi^k) - T(x^k), \end{aligned}$$

where $\mathcal{F}^k = \sigma(x^k, \dots, x^0, \xi^{k-1}, \dots, \xi^0)$. In fact the model of (2.20) has special properties that will be used as basic assumptions for the results that will be presented in terms of the more general model (2.21). These are summarized in the following simple lemma.

Lemma 2.5.1 *Let $\{x^k\}$ be a sequence in $\mathbb{R}^n$ generated according to the iteration*

$$x^{k+1} = x^k + \gamma(k)H(x^k, \xi^k),$$

under the folowing assumptions:

1. *H is Lipschitz continuous in x uniformly in ξ , i.e. there exists L s.t.*

$$\|H(x, \xi) - H(y, \xi)\| \leq L\|x - y\|, \quad \forall x, y, \xi.$$

2. *$\{\xi^k\}$ is a sequence of independent identically distributed random variables defined over $(\Omega, \mathcal{F}, P)$, that takes values on a compact set.*

Then

$$\begin{aligned} h(x^k) &= E[H(x^k, \xi^k) | \mathcal{F}^k], \\ w^k &= H(x^k, \xi^k) - h(x^k), \end{aligned}$$

where $\mathcal{F}^k = \sigma(x^k, \dots, x^0, \xi^{k-1}, \dots, \xi^0)$, satisfy the following properties:

- T is Lipschitz continuous and

$$\|h(x) - h(y)\| \leq L\|x - y\|, \quad \forall x, y.$$

•

$$\begin{aligned} E[w^k | \mathcal{F}^k] &= 0, \\ E[\|w^k\|^2 | \mathcal{F}^k] &\leq A + B\|x^k\|^2, \end{aligned}$$

for some constants A and B .

Proof The proof is a straightforward application of the scalar identity $|r| \leq 1 + r^2$, and the Lipschitz continuity condition. Without loss of generality, the norm in the argument below is assumed to be the 2-norm. Let y be any fixed point in $\mathbb{R}^n$, then for all x ,

$$\begin{aligned} \|h(x)\| &\leq \|h(y)\| + \|h(x) - h(y)\|, \\ &\leq \|h(y)\| + L\|x - y\|, \\ &\leq \|h(y)\| + L\|y\| + L(1 + \|x\|^2), \\ &\leq (L + L\|y\| + \|h(y)\|) + L\|x^k\|^2. \end{aligned} \tag{2.22}$$

Similarly, for all x, ξ ,

$$\begin{aligned} \|H(x, \xi)\| &\leq (L + L\|y\| + \|H(y, \xi)\|) + L\|x^k\|^2, \\ &\leq (L + L\|y\| + \max_{\xi} \|H(y, \xi)\|) + L\|x^k\|^2. \end{aligned} \tag{2.23}$$

Combining inequalities (2.22) and (2.23) yields

$$\|H(x^k, \xi^k) - h(x^k)\| \leq A + B\|x^k\|^2,$$

for appropriately defined constants A and B .

Q.E.D.

Lemma 2.5.2 *Let $\{w^k\}$, $\{A^k\}$, and $\{\gamma(k)\}$ be stochastic sequences defined over some probability space $(\Omega, \mathcal{F}, \mathcal{P})$, and let $\{\mathcal{F}^k\}$ be an increasing families of subfields of $\mathcal{F}$ such that:*

1. x^k is measurable w.r.t. $\mathcal{F}^k$, and $\{A^k\}$ is bounded w.p.1.

2. w^k is measurable w.r.t. $\mathcal{F}^{k+1}$

3.

$$\begin{aligned}\sum_{k=0}^{\infty} \gamma(k) &= \infty \quad \text{w.p.1,} \\ \sum_{k=0}^{\infty} \gamma(k)^2 &< \infty \quad \text{w.p.1.}\end{aligned}$$

4.

$$\begin{aligned}E[w^k | \mathcal{F}^k] &= 0, \\ E[\|w^k\|^2 | \mathcal{F}^k] &\leq A^k.\end{aligned}$$

Then $\sum_{m=0}^{\infty} \gamma(m)w^m$ converges w.p.1. Also

$$\lim_{k \rightarrow \infty} \sum_{m=k}^{m_T(k)} \gamma(m)w^m = 0,$$

for all T , where

$$m_T(k) = \min\{m \geq k : \sum_{l=k}^m \gamma(l) \geq T\}.$$

Proof This follows directly from Proposition VII-2-3 in Neveu [29].

Q.E.D.

The following lemma is a special case of the Kushner and Clark lemma, which has been used as the formal basis for the ODE approach. In particular, the assumptions we impose on the noise are consistent with the framework of the thesis and are sufficient for our objectives.

Lemma 2.5.3 *Let $\{x^k\}$ be a stochastic sequence in $\mathbb{R}^n$ generated according to*

$$x^{k+1} = x^k + \gamma(k) \left(h(x^k) + w^k + n^k \right).$$

where $h(\cdot)$ is a continuous map from $\mathbb{R}^n$ to $\mathbb{R}^n$, $\{w^k\}$ is a stochastic process defined over some probability space $(\Omega, \mathcal{F}, P)$, and $\{n^k\}$ is a given sequence in $\mathbb{R}^n$. Let $\mathcal{F}^k = \sigma(x^k, \dots, x^0, w^{k-1}, \dots, w^0)$. Assume that the following conditions hold:

1. h is Lipschitz continuous in x , i.e. There exist L s.t.

$$\|h(x) - h(y)\| \leq L\|x - y\|, \quad \forall x, y.$$

- 2.

$$\lim_{k \rightarrow \infty} n^k = 0. \tag{2.24}$$

- 3.

$$\begin{aligned} \sum_{k=0}^{\infty} \gamma(k) &= \infty, \\ \sum_{k=0}^{\infty} \gamma(k)^2 &< \infty. \end{aligned}$$

- 4.

$$\begin{aligned} E[w^k | \mathcal{F}^k] &= 0, \\ E[\|w^k\|^2 | \mathcal{F}^k] &\leq A + B\|x^k\|^2, \end{aligned} \tag{2.25}$$

for some constants A and B .

Then, if the iterates $\{x^k\}$ are bounded, and the ODE $\dot{x} = h(x)$ has a unique asymptotically stable equilibrium point x^* , then $\{x^k\}$ converges to x^* w.p.1.

Proof The proof we will present here is the same as the one in Kushner and Clark [21], with less general assumption on the noise. It is based on considering the limiting processes of linearly interpolated processes, where the interpolation intervals $[t_k, t_{k+1})$ are directly related to the stepsizes in the following way

$$t_k = \sum_{m=0}^{k-1} \gamma(m). \tag{2.26}$$

The standard ODE limiting argument studies the asymptotic behavior of the continuous-time processes defined below

$$\begin{aligned}
X_c(t) &= x^k && \text{for } t \in [t_k, t_{k+1}), \\
X_l(t) &= \begin{cases} x^k & \text{for } t = t_k \\ (1 - \frac{t-t_k}{\gamma(k)})x^k + \frac{t-t_k}{\gamma(k)}x^{k+1} & \text{for } t \in [t_k, t_{k+1}) \end{cases}, \\
W_l(t) &= \begin{cases} \sum_{m=0}^{k-1} \gamma(m)w^m & \text{for } t = t_k \\ (1 - \frac{t-t_k}{\gamma(k)})W_l(t_k) + \frac{t-t_k}{\gamma(k)}W_l(t_{k+1}) & \text{for } t \in [t_k, t_{k+1}) \end{cases} \\
N_l(t) &= \begin{cases} \sum_{m=0}^{k-1} \gamma(m)n^m & \text{for } t = t_k \\ (1 - \frac{t-t_k}{\gamma(k)})N_l(t_k) + \frac{t-t_k}{\gamma(k)}N_l(t_{k+1}) & \text{for } t \in [t_k, t_{k+1}) \end{cases}.
\end{aligned}$$

Define the left-shifted versions of all the processes above as follows:

$$\begin{aligned}
X_l^k(t) &= X_l(t + t_k) && \text{for } t \geq -t_k, \\
W_l^k(t) &= W_l(t + t_k) - W_l(t_k) && \text{for } t \geq -t_k, \\
N_l^k(t) &= N_l(t + t_k) - N_l(t_k) && \text{for } t \geq -t_k, \\
X_c^k(t) &= X_c(t + t_k) - G_c(t_k) && \text{for } t \geq -t_k.
\end{aligned}$$

Then simple manipulation shows that $X_l^k(\cdot)$ satisfies the integral equation

$$X_l^k(t) = X_l^k(0) + \int_0^t h(X_l^k(\tau))d\tau + W_l^k(t) + N_l^k(t) + e^k(t),$$

where

$$e^k(t) = \int_0^t h(X_c^k(\tau)) - h(X_l^k(\tau))d\tau.$$

Let $m_T(k) = \min\{m \geq k : \sum_{l=k}^m \gamma(l) \geq T\}$. If $\{x^k\}$ bounded, then Lemma 2.5.2 implies that w.p.1

$$\lim_{k \rightarrow \infty} \sum_{m=k}^{m_T(k)} \gamma(m)w^m = 0.$$

Also it can be easily verified that the assumptions on the stepsizes $\{\gamma(k)\}$ and the noise sequence $\{n^k\}$ imply

$$\begin{aligned}\lim_{k \rightarrow \infty} \sum_{m=k}^{m_T(k)} \gamma(m) n^m &= 0, \\ \lim_{k \rightarrow \infty} \sum_{m=k}^{m_T(k)} \gamma^2(m) &= 0.\end{aligned}$$

As for the convergence of the term $e^k(\cdot)$, a little more work is needed. It basically follows from the fact that the stepsize is square-summable and that the mapping h is Lipschitz continuous. In particular, partition the interval $[0, T]$ into subintervals $[0, t_{k+1} - t_k), [t_{k+1} - t_k, t_{k+2} - t_k), \dots, [t_{m_T(k)-1} - t_k, t_{m_T(k)} - t_k)$, where

$$m_T(k) = \min\{m \geq k : \sum_{l=k}^m \gamma(l) \geq T\}.$$

Then

$$\begin{aligned}\sup_{t \in [0, T]} \|e^k(t)\| &\leq \sum_{m=k}^{m_T(k)} \gamma(m) \sup_{\tau \in [t_m, t_{m+1})} \|h(X_c^m(\tau)) - h(X_l^m(\tau))\| d\tau, \\ &\leq \sum_{m=k}^{m_T(k)} \gamma^2(m) L \sup_k \|x^k\|.\end{aligned}$$

The last term is bounded due to the boundedness assumption on $\{x^k\}$. In addition, $\sum_{m=k}^{m_T(k)} \gamma^2(m) \rightarrow 0$ as $k \rightarrow \infty$, and the uniform convergence of $\{e^k(\cdot)\}$ follows.

Therefore

$$\begin{aligned}W_l^k(\cdot) &\rightarrow 0, \\ N_l^k(\cdot) &\rightarrow 0, \\ e^k(\cdot) &\rightarrow 0.\end{aligned}$$

Hence, w.p.1 any limit point $X(\cdot)$ of $\{X^k(\cdot)\}$ must satisfy the differential equation

$$\dot{X} = h(X).$$

To further prove that the sequence $\{x^k\}$ converges to x^* , we use the same argument as the one given by Kushner and Clark (see Page 24 of [21]). It depends very much on the asymptotic stability of the equilibrium point. For the sake of completeness, we will repeat the argument here. It would suffice to show that $X(t) = x^*$ for all t .

The proof is by contradiction. Suppose that the previous claim is not true. Then for small enough ϵ there is some time t_ϵ such that $\|X(t) - x^*\| \geq \epsilon$ for $t \in (-\infty, t_\epsilon)$. This can be easily seen by using the definition of asymptotic stability given by 2.1.1. Let τ be a time at which $X(\tau) \neq x^*$, and pick $R \in (0, \|X(\tau) - x^*\|)$. Then there exists $r > 0$ s.t. if the $X(0) \in B(x^*, r)$, then $X(t) \in B(x^*, R)$ for $t \geq 0$, where $B(x, R)$ denotes the ball in $\mathbb{R}^n$ of radius R around the point x . Pick any $\epsilon \leq R$. The ODE trajectory starting at $X(\tau)$ must converge to x^* and hence eventually gets within $B(x^*, r)$. Let $Z_r = X(\tau + \tau_r)$ be a point of this trajectory in $B(x^*, r)$, and consider the backward trajectory of the ODE, $\{X(t) : -\infty < t \leq \tau + \tau_r\}$, starting at Z_r . By construction $X(\tau)$ belongs to this trajectory, and if $X(\tau') \in B(x^*, r)$ for some $\tau' < \tau$, then asymptotic stability implies that $X(t)$ must be within $B(x^*, R)$ for $t \geq \tau'$. Of course, that includes τ , and that would contradict the choice of R . Now, consider the shifted processes $\hat{X}^n(\cdot)$

$$\hat{X}^n(t) = X(t + s^n),$$

where $\{s^n\}$ is a decreasing sequence which goes to $-\infty$. $\{\hat{X}^n(\cdot)\}$ is relatively compact in the uniform convergence topology, and hence has a converging subsequence $\hat{X}(\cdot)$. Clearly,

$$\|\hat{X}(t) - x^*\| \geq \epsilon \quad \text{for all } t,$$

for some $\epsilon > 0$. Moreover, $\hat{X}(\cdot)$ must satisfy the same ODE and hence must converge to x^* resulting in a contradiction.

Q.E.D.

2.6 Asynchronous Stochastic Approximation

Most of the work on stochastic approximation algorithms has focused on synchronous algorithms. In fact the interest in asynchronous algorithms has been motivated by problems of distributed computation, and appropriate models have been formulated starting with the paper by Tsistiklis, Bertsekas, and Athans [42] in 1986. The simple

formulation of asynchronous algorithms precisely defined notions of asynchronism in terms of the assignment of the computation of each component to a different processor, communication delays between each pair of processors, and synchronization of the clocks of the various processors, as these are well described in Bertsekas and Tsitsiklis' 1989 book on Distributed Computation [5]. For partially asynchronous stochastic gradient algorithms, where the communication delays between processes are bounded by a constant term, Tsitsiklis [41] has proved convergence using classical stochastic approximation techniques, and has discussed the possibility of using ODE limiting arguments to characterize the asymptotic behavior. Kushner and Yin in [22] and [23] have also considered the ODE approach for weak convergence results. However, an explicit ODE limit for asynchronous models, with a result similar to the Kushner and Clark Lemma, was not obtained until recently (1995), when Borkar developed a general ODE approach for asynchronous stochastic approximation [11] under the assumption that all components get updated according with a uniform minimum frequency, and the stepsize satisfies 'ideal tapering stepsize' conditions (see page 4 of [11], which basically includes the usual s.a. diminishing stepsize conditions and conditions that force the stepsize to depend on the number of times the component has been updated, together with some additional technical assumptions. Borkar establishes that the limiting ODE is time-varying and involves diagonal scaling of the synchronous ODE. The scaling reflects the degree of asynchronism. Under the additional assumption that there is a 'steady state' ratio between the cumulative stepsizes, the limiting ODE turns out to be time-invariant, and the diagonal scaling simply reflects the relative effects of the stepsize. In addition, if the stepsize function is uniform for all the components, the diagonal elements are all equal to 1 and the limiting ODE is simply a slower version of the synchronous ODE. We will focus on this class of algorithms in this thesis, as the analysis of the resulting limiting ODE follows from the stability analysis of ODE's with non-expansive maps of Borkar and Soumyanath [13], and a complete convergence analysis can be obtained.

We will first introduce the asynchronous model used by Borkar, state his main convergence theorem, and briefly discuss its implications. Then we will provide an

alternate proof of the theorem using the line of reasoning of Kushner and Clark. This will allow for a unified method of proof throughout the thesis.

Now, consider the algorithm

$$x_i^{k+1} = x_i^k + \gamma(\nu(k, i)) \left(F_i(x^k, \xi^k) - x_i^k + n_i^k \right) I(i \in Y^k), \quad (2.27)$$

where $\{n^k\}$ is a sequence that goes to 0 as $k \rightarrow \infty$, $\{Y^k\}$ is a set-valued random process that takes values in the collection of all subsets of $\{1, \dots, n\}$, I is the indicator function, $\gamma(\cdot)$ is a function from $\mathbb{N}$ to $\mathbb{R}$, and

$$\nu(k, i) = \sum_{m=0}^k I(i \in Y^m).$$

In this model, the stepsize is a function of the frequency of updates. This results in offsetting the effects of some components being updated more often than others, and essentially results in the same limiting ODE as for the synchronous case. The only difference between the two is that one is slower than the other due to the fact that it takes longer to update all the components. Note that the asynchronous model introduced in equation (2.27) allows for very general Gauss-Seidel iterations. The order in which the components are updated can be arbitrary, as long as the frequency by which each component is visited exceeds a certain threshold Δ . Also, it is worth noting that the stepsize used for updating each component is a function of the number of times the component has been visited so far.

The next lemma is a special case of the general ODE analysis of asynchronous s.a. iterations developed by Borkar [10]. The general analysis allows for delays which are themselves random and simplifies to the lemma stated above.

Lemma 2.6.1 *Let $\{x^k\}$ be a stochastic sequence in $\mathbb{R}^n$ generated according to the Gauss-Seidel asynchronous model described by equation (2.27), where $F(\cdot, \xi)$ is a continuous map from $\mathbb{R}^n$ to $\mathbb{R}^n$ for each ξ , and $\{\xi^k\}$ is a stochastic process defined over some probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{F}^k = \sigma(x^k, \dots, x^0, \xi^{k-1}, \dots, \xi^0, Y^k, \dots, Y^0)$. Let the following conditions hold:*

1. F is Lipschitz continuous in x uniformly in ξ , i.e. there exists L s.t.

$$\|F(x, \xi) - F(y, \xi)\| \leq L\|x - y\|, \quad \forall x, y, \xi$$

2. There is a continuous function $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$, such that

$$E \left[F(x^k, \xi^k) | \mathcal{F}^k \right] = T(x^k) \quad \forall k.$$

3. $\lim_{k \rightarrow \infty} n^k = 0$.

4. There exists $\Delta > 0$ such that for all i ,

$$\liminf_{k \rightarrow \infty} \frac{1}{k+1} \nu(k, i) \geq \Delta \quad w.p.1.$$

5. $\{\gamma(k)\}$ is an ideal tapering stepsize (see Definition 2.1.3).

Then, if the iterates $\{x^k\}$ are bounded, and the ODE $\dot{x} = T(x) - x$ has a global asymptotically stable equilibrium point, $\{x^k\}$ converges to x^* w.p.1.

Proof We will provide a proof of this convergence lemma based on the general analysis in [10], but for consistency we will follow a line argument similar to the one presented by Kushner and Clark [21]. The reasoning is insightful and will be the basis for analysis of the scaled asynchronous algorithm. Define the mappings H and h as follows:

$$H(x, \xi) = F(x, \xi) - x,$$

and

$$h(x) = T(x) - x.$$

Also, define $H^{(i)}$, and $h^{(i)}$ for $i = 1, \dots, n$ as follows

$$H_j^{(i)}(x, \xi) = \begin{cases} H_i(x, \xi) & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases},$$

$$h_j^{(i)}(x) = \begin{cases} h_i(x) & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases}.$$

For simplicity, we will assume that Y^k contains only one element ϕ^k . The general case may be obtained by using an unfolding technique described in [10]. Given these new definitions, the asynchronous algorithm (2.27) may be simplified to

$$x^{k+1} = x^k + \gamma(\nu(k, \phi^k)) H^{(\phi^k)}(x^k, \xi^k). \quad (2.28)$$

Let $w^k = H^{(\phi^k)}(x^k, \xi^k) - h^{(\phi^k)}(x^k)$. Then $\{w^k\}$ is a random process that satisfies

1. w^k is measurable with respect to $\mathcal{F}^{k+1}$,
2. $E[w^k | \mathcal{F}^k] = 0$,
3. $E[\|w^k\|^2 | \mathcal{F}^k] \leq A + B\|x^k\|^2$, for some constants A and B .

Equation (2.28) can be rewritten as

$$x^{k+1} = x^k + \gamma(\nu(k, \phi^k)) [h^{(\phi^k)}(x^k) + w^k].$$

Note that the iteration above looks very much like the synchronous model, except that the function $H^{(\phi^k)}$ is time-varying, and the stepsize may be viewed as a random variable now. But it still satisfies the usual stochastic approximation assumptions, i.e.

$$\begin{aligned} \sum_{k=0}^{\infty} \gamma(\nu(k, \phi^k)) &= \infty \quad \text{w.p.1,} \\ \sum_{k=0}^{\infty} \gamma^2(\nu(k, \phi^k)) &< \infty \quad \text{w.p.1.} \end{aligned}$$

The above follows directly from the assumption on the frequency of updates that

$$\liminf_{k \rightarrow \infty} \frac{1}{k} \sum_{m=0}^{k-1} I(\phi^m = i) \geq \Delta \quad \text{for } \Delta > 0.$$

As is standard in limiting ODE arguments, we need to define piece-wise linear and piece-wise constant interpolated processes based on the iterates $\{x^k\}$ over time

intervals $[t_k, t_{k+1})$, where

$$t_k = \sum_{m=0}^{k-1} \gamma(\nu(m, \phi^m)). \quad (2.29)$$

Before we proceed with these definitions, we need to define the sequence $\{\mu^k\}$ of vectors in $\mathcal{P}(S)$, where

$$\mathcal{P}(S) = \{v \in \mathbb{R}^n : v_i \in [0, 1] \text{ and } \sum_{i=1}^n v_i = 1\},$$

as

$$\mu_i^k = I(\phi^k = i) \quad \text{for } i \in S. \quad (2.30)$$

Then the functions $H^{(\phi^k)}(\cdot, \cdot)$ and $h^{(\phi^k)}(\cdot)$ may be expressed respectively as $\text{diag}(\mu^k)H(\cdot, \cdot)$ and $\text{diag}(\mu^k)h(\cdot)$, where $\text{diag}(\mu)$ is a diagonal matrix whose diagonal elements are the components of μ , i.e.

$$\text{diag}(\mu) = \begin{bmatrix} \mu_1 & \dots & 0 \\ \vdots & \ddots & \\ 0 & & \mu_n \end{bmatrix}.$$

Now the interpolated processes are defined as follows

$$\begin{aligned} \mu(t) &= \mu^k && \text{for } t \in [t_k, t_{k+1}), \\ X_c(t) &= x^k && \text{for } t \in [t_k, t_{k+1}), \\ X_l(t) &= \begin{cases} x^k & \text{for } t = t_k \\ (1 - \frac{t-t_k}{\gamma(\nu(k, \phi^k))})x^k + \frac{t-t_k}{\gamma(\nu(k, \phi^k))}x^{k+1} & \text{for } t \in [t_k, t_{k+1}) \end{cases}, \\ W_l(t) &= \begin{cases} \sum_{m=0}^{k-1} \gamma(\nu(m, \phi^m))w^m & \text{for } t = t_k \\ (1 - \frac{t-t_k}{\gamma(\nu(k, \phi^k))})W_l(t_k) + \frac{t-t_k}{\gamma(\nu(k, \phi^k))}W_l(t_{k+1}) & \text{for } t \in [t_k, t_{k+1}) \end{cases}, \\ N_l(t) &= \begin{cases} \sum_{m=0}^{k-1} \gamma(\nu(m, \phi^m))n^m & \text{for } t = t_k \\ (1 - \frac{t-t_k}{\gamma(\nu(k, \phi^k))})N_l(t_k) + \frac{t-t_k}{\gamma(\nu(k, \phi^k))}N_l(t_{k+1}) & \text{for } t \in [t_k, t_{k+1}) \end{cases}. \end{aligned}$$

Define the left-shifted versions of all the processes above as follows:

$$\begin{aligned}
\mu^k(t) &= \mu(t + t_k) && \text{for } t \geq -t_k \\
X_c^k(t) &= X_c(t + t_k) - X_c(t_k) && \text{for } t \geq -t_k, \\
X_l^k(t) &= X_l(t + t_k) && \text{for } t \geq -t_k, \\
W_l^k(t) &= W_l(t + t_k) - W_l(t_k) && \text{for } t \geq -t_k, \\
N_l^k(t) &= N_l(t + t_k) - N_l(t_k) && \text{for } t \geq -t_k.
\end{aligned} \tag{2.31}$$

We need to invoke some interesting results from [10] in order to study the limiting behavior of the processes above. In particular, we would like to find a ‘meaningful’ and ‘useful’ limit of the sequence $\{\mu^k(\cdot)\}$. As suggested by Borkar in [10], consider the space U of $\mathcal{P}(S)$ -valued continuous-time processes, of the form $\bar{\mu} = \{\mu(t) : t \geq 0\}$, with the coarsest topology that ensures the continuity of the family of maps $\{F^{(p,T)} : T \geq 0, p_i \in L_2[0, T]\}$, where $F^{(p,T)} : U \rightarrow \mathbb{R}^n$, and

$$F^{(p,T)}(\bar{\mu}) = \int_0^T \text{diag}(\mu(\tau))p(\tau)d\tau.$$

Borkar shows that U is compact in the topology defined above (Page 9 of [10]), and furthermore the sequence $\{\mu^k(\cdot)\}$ defined in (2.31) converges to μ^* , where $\mu^*(t) = \frac{1}{n}$ for all t . The convergence is in the sense that for all $p : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that $p_i \in L_2[0, T]$,

$$\int_0^T \text{diag}(\mu^k(\tau))p(\tau)d\tau \rightarrow \int_0^T \text{diag}(\mu^*(\tau))p(\tau)d\tau, \tag{2.32}$$

where the convergence above is component-wise. This type of convergence is exactly what is needed in analyzing the limiting behavior of the processes above. Intuitively, conditions 4 and 5 in Lemma 2.6.1 imply that if, for example, on the average a component gets updated twice as frequently as another, then the stepsize of the update will be half. The time scaling used for the interpolating processes reflects this interplay between the frequency of update and the size of update.

Given the definitions above, simple manipulation shows that for $t \geq -t^k$,

$$\begin{aligned} X_l^k(t) &= X_l^k(0) + \int_0^t \text{diag}(\mu^k(\tau))h(X_c^k(\tau))d\tau + W_l^k(t) + N_l^k(t) \\ &= X_l^k(0) + \int_0^t \text{diag}(\mu^k(\tau))h(X_l^k(\tau))d\tau + W_l^k(t) + N_l^k(t) + e^k(t), \\ &= X_l^k(0) + \int_0^t \text{diag}(\mu^*(\tau))h(X_l^k(\tau))d\tau + W_l^k(t) + N_l^k(t) + e^k(t) + \eta^k(t), \end{aligned}$$

where

$$e^k(t) = \int_0^t \text{diag}(\mu^k(\tau))h(X_c^k(\tau)) - \text{diag}(\mu^k(\tau))h(X_l^k(\tau))d\tau,$$

and

$$\eta^k(t) = \int_0^t \text{diag}(\mu^k(\tau))h(X_l^k(\tau))d\tau - \int_0^t \text{diag}(\mu^*(\tau))h(X_l^k(\tau))d\tau.$$

We will show that $e^k(\cdot) \rightarrow 0$, $W_l^k(\cdot) \rightarrow 0$, and $N_l^k(\cdot) \rightarrow 0$, as $k \rightarrow \infty$ uniformly on any finite interval. As for $\{\eta^k(\cdot)\}$, we will show that it converges to 0 along a converging subsequence of $\{X_l^k(\cdot)\}$.

The convergence of $\{e^k(\cdot)\}$ is due to the fact that the stepsize is square-summable and that the mapping h is Lipschitz continuous. To see this, partition the interval $[0, T]$ into subintervals $[0, t_{k+1} - t_k)$, $[t_{k+1} - t_k, t_{k+2} - t_k)$, $\dots$, $[t_{m_T(k)-1} - t_k, t_{m_T(k)} - t_k)$, where

$$m_T(k) = \min\{m \geq k : \sum_{l=k}^m \gamma(\nu(l, \phi^l)) \geq T\}.$$

Then

$$\begin{aligned} \sup_{t \in [0, T]} \|e^k(t)\| &\leq \sum_{m=k}^{m_T(k)} \gamma(\nu(m, \phi^m)) \\ &\quad \sup_{\tau \in [t_m, t_{m+1})} \|\text{diag}(\mu^m(\tau))h(X_c^m(\tau)) - h(X_l^m(\tau))\|, \\ &\leq \sum_{m=k}^{m_T(k)} \gamma^2(\nu(m, \phi^m))L \sup_k \|x^k\|. \end{aligned}$$

Note that the functions $\text{diag}(\mu^k)h(\cdot)$ basically pick out one component of the function h , and hence are also Lipschitz continuous with the same Lipschitz constant. The last term is bounded due to the boundedness assumption on $\{x^k\}$. In addition, $\sum_{m=k}^{m_T(k)} \gamma^2(\nu(m, \phi^m)) \rightarrow 0$ as $k \rightarrow \infty$, and the uniform convergence of $\{e^k(\cdot)\}$ follows.

Similarly, $W^k = \sum_{m=0}^{k-1} \gamma(\nu(m, \phi^m))w^m$ is a zero-mean bounded increment Martingale with respect to $\mathcal{F}^k$, and hence by Lemma 2.5.2, it follows that $\{W^k\}$ converges

w.p.1, and hence

$$\lim_{k \rightarrow \infty} \sum_{m=k}^{m_T(k)} \gamma(\nu(m, \phi^m)) w^m = 0 \quad \text{w.p.1.} \quad (2.33)$$

Equation (2.33) implies that $\{W_l^k(\cdot)\}$ converges to 0 uniformly on any finite interval.

Because $\{n^k\}$ converges to 0, it follows that

$$\lim_{k \rightarrow \infty} \sum_{m=k}^{m_T(k)} \gamma(\nu(m, \phi^m)) n^m = 0,$$

and hence $\{N_l^k(\cdot)\}$ converges to 0 uniformly on any finite interval.

Now, the sequence $\{X_l^k\}$ is bounded and equi-continuous over any finite interval, and hence is relatively compact by the Arzela-Ascoli theorem. Therefore it must have a converging subsequence. Let us consider a converging subsequence and label it $\{X_l^k\}$ by abuse of notation. Let $X(\cdot)$ be the limit point of the converging subsequence $\{X_l^k\}$. Along this subsequence,

$$\begin{aligned} \eta^k(t) = & \overbrace{\int_0^t \text{diag}(\mu^k(\tau)) h(X(\tau)) d\tau - \int_0^t \text{diag}(\mu^*) h(X(\tau)) d\tau}^{\eta_1^k} \\ & + \overbrace{\int_0^t \text{diag}(\mu^*) h(X(\tau)) d\tau - \int_0^t \text{diag}(\mu^*) h(X_l^k(\tau)) d\tau}^{\eta_2^k} \\ & + \overbrace{\int_0^t \text{diag}(\mu^k(\tau)) h(X_l^k(\tau)) d\tau - \int_0^t \text{diag}(\mu^k(\tau)) h(X(\tau)) d\tau}^{\eta_3^k}. \end{aligned} \quad (2.34)$$

For a given $\epsilon > 0$, there is K_1 such that $\|\eta_1^k\| < \frac{\epsilon}{3}$ for $k \geq K_1$ as $h(X(\cdot))$ is continuous and hence each of its components belongs to $L_2[0, t]$. Similarly, there is K_2 such that $\|\eta_2^k\| < \frac{\epsilon}{3}$ for $k \geq K_2$ due to the convergence of $\{X_l(\cdot)\}$ to $X(\cdot)$. The Lipschitz continuity of $h(\cdot)$ implies that

$$\|\eta_3^k\| \leq L \int_0^t \|X_l^k(\tau) - X(\tau)\|,$$

and hence there is K_3 such that $\|\eta_3^k\| < \frac{\epsilon}{3}$ for $k \geq K_3$. Pick $K = \max(K_1, K_2, K_3)$. It follows that $\|\eta^k(t)\| < \epsilon$ for $k > K$. Given that $\{e^k(\cdot)\}$ are continuous, the uniform

convergence over finite intervals of $\{e^k(\cdot)\}$ to 0 follows.

Therefore w.p.1 any limit point $X(\cdot)$ of $\{X^k(\cdot)\}$ must satisfy the differential equation

$$\begin{aligned}\dot{X} &= \text{diag}(\mu^*)h(X), \\ &= \frac{1}{n}h(X).\end{aligned}$$

The convergence of $\{x^k\}$ to x^* w.p.1 follows from the asymptotic stability of the ODE using the same argument as the one used for the proof of the synchronous convergence result of Section 2.5.

Q.E.D.

2.7 Two-time Scale Stochastic Approximation

In this section, we will review the results of Borkar's paper [11], which uses an ODE argument to analyze stochastic algorithms involving two coupled iterations with different stepsizes. This work is the basis for the analysis of the second Q -learning algorithm that we propose. Although the results in [11] are for synchronous algorithms only, it is straightforward to extend them to asynchronous algorithms using ideas from Section 2.6. We will present the results of [11], as well as a brief proof based on the Kushner and Clark line of reasoning. We will also give the analogous asynchronous results and outline a corresponding proof. Let $\{x^k\}$ and $\{y^k\}$ be sequences in $\mathbb{R}^n$ and $\mathbb{R}^l$, respectively, which are generated as follows

$$\begin{aligned}x^{k+1} &= x^k + \gamma(k) \left(h(x^k, y^k) + w_1^k \right), \\ y^{k+1} &= y^k + \beta(k) \left(G(x^k, y^k) + w_2^k \right),\end{aligned}\tag{2.35}$$

under the following set of assumptions:

Assumption 2.7.1

1. h and G are Lipschitz continuous.

2. The stepsizes satisfy the following:

$$\begin{aligned}\sum_{k=0}^{\infty} \gamma(k) &= \infty \\ \sum_{k=0}^{\infty} \gamma(k)^2 &< \infty\end{aligned}\tag{2.36}$$

$$\begin{aligned}\sum_{k=0}^{\infty} \beta(k) &= \infty \\ \sum_{k=0}^{\infty} \beta(k)^2 &< \infty\end{aligned}\tag{2.37}$$

and

$$\limsup_{k \rightarrow \infty} \frac{\beta(k)}{\gamma(k)} = 0.\tag{2.38}$$

3. Let $\mathcal{F}^k = \sigma(x^k, \dots, x^0, y^k, \dots, y^0, w^{k-1}, \dots, w^0)$, then

$$\begin{aligned}E[w_1^k | \mathcal{F}^k] &= 0, \\ E[\|w_1^k\|^2 | \mathcal{F}^k] &\leq A_1 + B_1 \|x^k\|^2 + D_1 \|y^k\|^2.\end{aligned}\tag{2.39}$$

Similarly,

$$\begin{aligned}E[w_2^k | \mathcal{F}^k] &= 0, \\ E[\|w_2^k\|^2 | \mathcal{F}^k] &\leq A_2 + B_2 \|x^k\|^2 + D_2 \|y^k\|^2,\end{aligned}\tag{2.40}$$

for some constants A_2, B_2 , and D_2 .

4. For all $y \in \mathbb{R}^l$, the ODE

$$\dot{x}(t) = h(x(t), y)\tag{2.41}$$

has a global asymptotically stable equilibrium point $g(y)$ such that the map g is Lipschitz continuous.

5. The ODE

$$\dot{y}(t) = G(g(y(t)), y(t))\tag{2.42}$$

has a global asymptotically stable equilibrium point y^* .

Let us briefly discuss the assumptions above: The Lipschitz continuity assumption, along with conditions (2.36) and (2.37) on the stepsizes are just the usual. Condition (2.38) is a key condition for the two-time scale analysis. It states that the second

iteration is much slower than the first one as the size of the update is determined by the smaller stepsize $\beta(k)$. Intuitively, this condition says the first iteration in x sees the y component as a constant, while the second iteration in y sees the first iteration (which converges faster) as having reached its limit point $g(y)$, which is the global equilibrium point of the ODE (2.41). The limiting behavior of the second iteration is thus governed by the ODE (2.42), which has a global asymptotically stable equilibrium point y^* . This in turn implies that $\{(x^k, y^k)\}$ should converge to $(g(y^*), y^*)$ w.p.1. In fact, this is the result of Theorem 1.1 from [11], which is restated below:

Lemma 2.7.1 *Consider the coupled sequence $\{(x^k, y^k)\}$ generated by (2.35) under Assumption 2.7.1. If the iterates are bounded, then $\{(x^k, y^k)\}$ converges to $(g(y^*), y^*)$ w.p.1.*

Proof Again, we will simply adapt the proof in [11] to a Kushner and Clark argument. Again, we define the piece-wise linear interpolated processes $X_l(\cdot)$ and $Y_l(\cdot)$ based on the iterates $\{x^k\}$ and $\{y^k\}$ over the interpolation intervals $[t_k, t_{k+1})$, along with their shifted versions $\{X_l^k(\cdot)\}$ and $\{Y_l^k(\cdot)\}$ in the standard way. First, we will consider the faster time scale $\{\gamma(k)\}$, i.e.

$$t_k = \sum_{m=0}^{k-1} \gamma(m).$$

We can easily characterize the limiting processes of $X(\cdot)$ and $Y(\cdot)$ of $\{X_l^k(\cdot)\}$ and $\{Y_l^k(\cdot)\}$ using Lemma 2.5.3 if we re-write the iterations in (2.35) as

$$\begin{aligned} x^{k+1} &= x^k + \gamma(k) \left[h(x^k, y^k) + w_1^k \right], \\ y^{k+1} &= y^k + \gamma(k) \left[\frac{\beta(k)}{\gamma(k)} G(x^k, y^k) + w_2^k \right]. \end{aligned} \tag{2.43}$$

Note that each one of the iterations above is a special case of the stochastic iteration of Lemma 2.5.3. In particular, the sequences $\{w_1^k\}$ and $\{w_2^k\}$ in both iterations satisfy condition (2.25) of the lemma, and similarly the sequence $\{\frac{\beta(k)}{\gamma(k)} G(x^k, y^k)\}$ satisfies

condition (2.24) of the lemma. Therefore, the limiting processes $X(\cdot)$ and $Y(\cdot)$ satisfy

$$\begin{aligned}\dot{X}(t) &= h(X(t), Y(t)), \\ \dot{Y}(t) &= 0.\end{aligned}$$

The above of course implies that $Y(t) = Y$ for $t \geq 0$, for some constant Y , and $X(\cdot)$ satisfies the ODE $\dot{X}(t) = h(X(t), Y)$. The assumption imposed on the ODE (2.41), along with the argument at the end of Section 2.6, imply that

$$X(t) = g(Y) \quad \text{for all } t > 0.$$

Therefore,

$$(x^k, y^k) \rightarrow (g(y), y) \quad \text{as } k \rightarrow \infty, \tag{2.44}$$

for some constant vector y .

Again, let us define the piece-wise linear interpolated processes $\bar{Y}_l(\cdot)$ based on the iterates $\{y^k\}$ over the interpolation intervals $[t_k, t_{k+1})$, along with its shifted versions $\{\bar{Y}_l^k(\cdot)\}$ in the standard way. But now the interpolation intervals are over the slower time scale $\{\beta(k)\}$, i.e.

$$t_k = \sum_{m=0}^{k-1} \beta(m).$$

Let us re-write the second iteration as follows

$$y^{k+1} = y^k + \beta(k) \left[G(g(y^k), y^k) + n^k + w_2^k \right], \tag{2.45}$$

where

$$n^k = G(x^k, y^k) - G(g(y^k), y^k).$$

Iteration (2.45) is a special case of the stochastic algorithm of Lemma 2.5.3. Particularly, the sequence $\{w_2^k\}$ satisfies condition (2.25) of the lemma, and $\{n^k\}$ converges to 0 due to (2.44) and the fact that G and g are Lipschitz continuous. Therefore, the

limiting process $\bar{Y}(\cdot)$ of $\{Y_l^k(\cdot)\}$ satisfies the ODE:

$$\dot{\bar{Y}} = G(g(y), y).$$

Hence

$$y^k \rightarrow y^* \quad \text{as } k \rightarrow \infty,$$

and

$$x^k \rightarrow g(y^*) \quad \text{as } k \rightarrow \infty.$$

Q.E.D.

Lemma 2.7.1 can be easily extended to analyze coupled asynchronous stochastic approximation iterations with different stepsizes. In particular, let $\{x^k\}$ and $\{y^k\}$ be sequences in $\mathbb{R}^n$ and $\mathbb{R}^l$, respectively, which are generated as follows:

$$\begin{aligned} x_i^{k+1} &= x_i^k + \gamma(\nu_1(k, i)) (h_i(x^k, y^k) + w_{1i}^k) I(i = \phi_1^k) \quad \text{for } i = 1, \dots, n, \\ y_j^{k+1} &= y_j^k + \beta(\nu_2(k, j)) (G_j(x^k, y^k) + w_{2j}^k) I(j = \phi_2^k) \quad \text{for } j = 1, \dots, l, \end{aligned}$$

where $\{\phi_1^k\}$ and $\{\phi_2^k\}$ are random processes that take values over the sets $S_1 = \{1, \dots, n\}$ and $S_2 = \{1, \dots, l\}$, respectively, and

$$\begin{aligned} \nu_1(k, i) &= \sum_{m=0}^k I(i = \phi_1^m), \\ \nu_2(k, j) &= \sum_{m=0}^k I(j = \phi_2^m). \end{aligned}$$

In addition, we impose the following set of assumptions:

Assumption 2.7.2

1. h and G are Lipschitz continuous.
2. There exist $\Delta_1 > 0$ and $\Delta_2 > 0$ such that for all $i = 1, \dots, n$, and $j = 1, \dots, l$,

$$\begin{aligned} \liminf_{k \rightarrow \infty} \frac{1}{k+1} \nu_1(k, i) &\geq \Delta_1 \quad \text{w.p.1,} \\ \liminf_{k \rightarrow \infty} \frac{1}{k+1} \nu_2(k, j) &\geq \Delta_2 \quad \text{w.p.1.} \end{aligned} \tag{2.46}$$

3. The stepsizes $\{\gamma(k)\}$ and $\{\beta(k)\}$ are ideal tapering stepsizes (see Definition 2.1.3), and they satisfy

$$\limsup_{k \rightarrow \infty} \frac{\beta(k)}{\gamma(k)} = 0. \quad (2.47)$$

4. Let $\mathcal{F}^k = \sigma(x^k, \dots, x^0, y^k, \dots, y^0, w^{k-1}, \dots, w^0)$, then

$$\begin{aligned} E[w_1^k | \mathcal{F}^k] &= 0, \\ E[\|w_1^k\|^2 | \mathcal{F}^k] &\leq A_1 + B_1 \|x^k\|^2 + D_1 \|y^k\|^2, \end{aligned} \quad (2.48)$$

for some constants A_1, B_1 , and D_1 . Similarly,

$$\begin{aligned} E[w_2^k | \mathcal{F}^k] &= 0, \\ E[\|w_2^k\|^2 | \mathcal{F}^k] &\leq A_2 + B_2 \|x^k\|^2 + D_2 \|y^k\|^2, \end{aligned} \quad (2.49)$$

for some constants A_2, B_2 , and D_2 .

5. For all $y \in \mathbb{R}^l$, the ODE

$$\dot{x}(t) = h(x(t), y) \quad (2.50)$$

has a global asymptotically stable equilibrium point $g(y)$ such that the map g is Lipschitz continuous.

6. The ODE

$$\dot{y}(t) = G(g(y(t)), y(t)) \quad (2.51)$$

has a global asymptotically stable equilibrium point y^* .

The assumptions above are basically the same as the ones for the synchronous case, with the additional frequency condition on the updates of the different components. The same intuition applies to the analysis of this problem, and the extension to Lemma 2.7.1 is simply:

Lemma 2.7.2 *Consider the coupled sequence $\{(x^k, y^k)\}$ generated by (2.35) under Assumption (2.7.2). If the iterates are bounded, then $\{(x^k, y^k)\}$ converges to $(g(y^*), y^*)$ w.p.1.*

Proof The proof of Lemma 2.7.2 follows easily from a combination of the proofs of Lemma 2.7.1 and Lemma 2.6.1. The only additional fact that will be needed to complete the argument is

$$\lim_{k \rightarrow \infty} \frac{\beta(\nu_2(k, \phi_2^k))}{\gamma(\nu_1(k, \phi_1^k))} = 0 \quad \text{w.p.1.}$$

It is easy to verify that this follows from conditions (2.46) and (2.47). In addition the limiting ODE's for the processes $X_l(\cdot)$ and $Y_l(\cdot)$ will be

$$\dot{x} = \frac{1}{n} h(x(t), y) \quad \text{for a fixed } y,$$

and

$$\dot{y} = 0,$$

respectively. The limiting ODE for the process $\bar{Y}_l(\cdot)$ will be

$$\dot{y} = \frac{1}{l} G(g(y(t)), y(t)).$$

Q.E.D.

Chapter 3

Theory for General Framework

3.1 Preliminaries

In this chapter, we will state and prove the basic results of this thesis, which are divided into two parts: The boundedness lemmas and the convergence analysis of appropriately scaled iterations, both synchronous and asynchronous. The boundedness lemmas will rely heavily on the non-expansiveness condition, with respect to some norm for the synchronous case, and the sup-norm for the asynchronous case. The analysis of the scaled iteration depends on the fact that the projection is picked to depend on a Lyapunov surface of an appropriate ODE.

The subsequent ideas rely heavily on a map that projects any point onto a bounded set $\mathcal{B}$ via scaling. The projection is carried out each time the point leaves a larger set $\mathcal{C}$. Let $\partial\mathcal{B}$ and $\bar{\mathcal{B}}$ denote the boundary and closure of $\mathcal{B}$, respectively. (i.e. $\bar{\mathcal{B}} = \mathcal{B} \cup \partial\mathcal{B}$). Similar definitions apply for $\partial\mathcal{C}$ and $\bar{\mathcal{C}}$. We say that $\mathcal{B}$ is totally enclosed by $\mathcal{C}$ if $\mathcal{B} \subset \mathcal{C}$ and $\inf_{x \in \partial\mathcal{B}, y \in \partial\mathcal{C}} \|x - y\| = d > 0$, (i.e. there is a nonzero minimum distance between the two sets). The map $\Pi_{\mathcal{B},\mathcal{C}}$ is defined as follows:

Definition 3.1.1 *Let $\mathcal{B}$ be an open ball in $\mathbb{R}^n$ around the origin, and $\mathcal{C}$ be an open, connected, and bounded subset of $\mathbb{R}^n$ which totally contains $\mathcal{B}$. We define the mapping $\Pi_{\mathcal{B},\mathcal{C}}$ from $\mathbb{R}^n$ to $\bar{\mathcal{B}}$ as follows:*

$$\Pi_{\mathcal{B},\mathcal{C}}(x) = \gamma_{\mathcal{B},\mathcal{C}}(x) \cdot x,$$

where $\gamma_{\mathcal{B},\mathcal{C}} : \mathbb{R}^n \rightarrow (0, 1]$ is given by

$$\gamma_{\mathcal{B},\mathcal{C}}(x) = \begin{cases} 1 & \text{if } x \in \mathcal{C} \\ \max\{\beta > 0 : \beta x \in \bar{\mathcal{B}}\} & \text{if } x \notin \mathcal{C} \end{cases}$$

It is easy to check that the map above is a well defined real-valued function. In fact for all x , $\gamma_{\mathcal{B},\mathcal{C}}(x)$ is uniquely defined and $\gamma_{\mathcal{B},\mathcal{C}}(x) \leq 1$. Given that $\mathcal{B}$ is an open ball, the map $\Pi_{\mathcal{B},\mathcal{C}}$ is actually a projection, but the decision to project depends on whether the point is outside a bigger ball $\mathcal{C}$. We will make use of this projection map extensively in our analysis.

3.2 Boundedness Lemmas

The first boundedness lemma is inspired by a lemma in Jaakola, Jordan, and Singh [17], which guarantees convergence of the original iteration as long as a scaled version converges under the strong condition of homogeneity. It turns out that if the mapping involved is non-expansive with respect to some norm or span semi-norm, the boundedness in norm or span semi-norm of the original iteration follows from the convergence of a scaled version. The boundedness in norm will be used along with the ODE analysis to prove convergence.

Lemma 3.2.1 *Let $\mathcal{B}$ be an open and bounded subset of $\mathbb{R}^n$ containing the origin, and let $\mathcal{C}$ be an open and bounded subset of $\mathbb{R}^n$ containing $\mathcal{B}$. Consider the algorithm*

$$x^{k+1} = G^k(x^k, \xi^k). \tag{3.1}$$

Let the following assumptions hold:

1. *Noise conditions: $\{\xi^k\}$ is a random process defined over the probability space $(\Omega, P, \mathcal{F})$.*
2. *Non-expansiveness in x with respect to some norm:*

$$\|G^k(x, \xi) - G^k(y, \xi)\| \leq \|x - y\| \quad \forall x, y, \xi.$$

3. *Convergence of the scaled iteration: The sequence $\{\tilde{x}^k\}$ generated by*

$$\tilde{x}^{k+1} = G^k(\Pi_{\mathcal{B},c}(\tilde{x}^k), \xi^k), \quad \tilde{x}^0 = x^0,$$

converges to some vector $x^ \in \mathcal{B}$ w.p.1.*

Then $\{x^k\}$ is bounded w.p.1.

Proof By assumption, there exists a null set Ω_0 such that if $\omega \notin \Omega_0$, then $\{\tilde{x}^k\}$ converges to some $x^* \in \mathcal{B}$. Consider any $\omega \notin \Omega_0$. Then for every neighborhood $\mathcal{N}$ of x^* , there exists N , such that $\tilde{x}^k \in \mathcal{N}$ for $k \geq N$. As $x^* \in \mathcal{B}$ and $\mathcal{B}$ is open, there exists a neighborhood $\mathcal{N}$ of x^* that is totally contained in $\mathcal{B}$. Therefore, there exists a large enough N , such that $\tilde{x}^k \in \mathcal{B}$ for $k \geq N$. In other words, there exists N large enough such that

$$\gamma_{\mathcal{B},c}(\tilde{x}^k) = 1 \quad \text{for } k \geq N, \quad (3.2)$$

and hence

$$\tilde{x}^{k+1} = G^k(\tilde{x}^k, \xi^k) \quad \text{for } k \geq N. \quad (3.3)$$

Therefore, for $k \geq N$,

$$\begin{aligned} \|x^{k+1} - \tilde{x}^{k+1}\| &= \|G^k(x^k, \xi^k) - G^k(\tilde{x}^k, \xi^k)\| \\ &\leq \|x^k - \tilde{x}^k\| \\ &\leq \|x^N - \tilde{x}^N\|. \end{aligned}$$

As $\{\tilde{x}^k\}$ is bounded, it follows that $\{x^k\}$ is bounded. The above is true for all $\omega \notin \Omega_0$. Therefore, $\{x^k\}$ is bounded w.p.1.

Q.E.D.

The second lemma gives another set of conditions under which a weaker result can be obtained. Fortunately, this will be sufficient for the analysis of the Q -learning algorithm for the average cost problem. In this case, we replace the non-expansive property by non-expansiveness with respect to the span semi-norm.

Lemma 3.2.2 *Let $\mathcal{B}$ be an open and bounded subset of $\mathbb{R}^n$ containing the origin, and let $\mathcal{C}$ be an open and bounded subset of $\mathbb{R}^n$ containing $\mathcal{B}$. Consider the algorithm*

$$x^{k+1} = G^k(x^k, \xi^k). \quad (3.4)$$

Let the following assumptions hold:

1. *Noise conditions: $\{\xi^k\}$ is a random process defined over the probability space $(\Omega, P, \mathcal{F})$.*
2. *Non-expansiveness in x with respect to the span semi-norm:*

$$\|G^k(x, \xi) - G^k(y, \xi)\|_s \leq \|x - y\|_s \quad \forall x, y, \xi.$$

3. *Convergence of scaled iteration:*

$$\tilde{x}^{k+1} = G^k(\Pi_{\mathcal{B}, \mathcal{C}}(\tilde{x}^k), \xi^k), \quad \tilde{x}^0 = x^0,$$

converges to some vector $x^ \in \mathcal{B}$ w.p.1.*

Then $\{\|x^k\|_s\}$ is bounded w.p.1.

Proof The proof is very similar to the proof of Lemma 3.2.1, as semi-norms also satisfy the triangular inequality. The same reasoning implies that except on a set of measure 0, $\{\tilde{x}^k\}$ converges to x^* . Let N be large enough so that $\tilde{x}^k \in \mathcal{B}$ for $k \geq N$. Then

$$\tilde{x}^{k+1} = G^k(\tilde{x}^k, \xi^k) \quad \text{for } k \geq N. \quad (3.5)$$

Therefore for $k \geq N$,

$$\begin{aligned} \|x^{k+1} - \tilde{x}^{k+1}\|_s &\leq \|G^k(x^k, \xi^k) - G^k(\tilde{x}^k, \xi^k)\|_s \\ &\leq \|\tilde{x}^k - x^k\|_s \\ &\leq \|\tilde{x}^N - x^N\|_s. \end{aligned}$$

But $\{\|\tilde{x}^k\|_s\}$ is bounded, implying that $\{\|x^k\|_s\}$ is bounded. The above is true except for a set of measure 0, therefore $\{\|x^k\|_s\}$ is bounded w.p.1.

Q.E.D.

3.3 Analysis of the Scaled Iteration

Under appropriate assumptions, the previous section guarantees boundedness of the original iteration as long as the scaled iteration converges. When applied to the synchronous and asynchronous algorithms given by (2.20) and (2.27), the boundedness lemmas above raise an interesting possibility, which is to use ODE types of arguments to establish the convergence of the scaled iterations. The issue of boundedness is taken care of by construction. However, this gives rise to two technical challenges:

- showing that a limiting ODE exists,
- insuring that the limiting ODE, if it exists, is easy to analyze.

The two difficulties above are of course tightly related, and it is easy to see that the method of scaling impacts them both closely. Any general scaling scheme could make the problem intractable. If the scaling is not done appropriately, the scaled iteration might not converge. The iterates could for example keep hitting the boundary infinitely often, and never converge, or the scaling could generate additional fixed points at the boundary that the iterates might converge to.

In particular, the scaled version of the synchronous and asynchronous algorithms of equations (2.20) and (2.27) are respectively given by

$$\begin{aligned}\tilde{x}^{k+1} &= z^k + \gamma(k) \left(F(z^k, \xi^k) - z^k \right), \\ z^{k+1} &= \Pi_{\mathcal{B}, \mathcal{C}}(\tilde{x}^{k+1}),\end{aligned}\tag{3.6}$$

and

$$\begin{aligned}\tilde{x}_i^{k+1} &= z_i^k + \gamma(\nu(k, i)) \left(F_i(z^k, \xi^k) - z_i^k \right) I\{i \in Y^k\}, \\ z^{k+1} &= \Pi_{\mathcal{B}, \mathcal{C}}(\tilde{x}^{k+1}),\end{aligned}\tag{3.7}$$

where $\Pi_{\mathcal{B},\mathcal{C}}$ is given by Definition 3.1.1. The ODE we will be referring to in the following is $\dot{x} = T(x) - x$, where T is defined as in Lemma 2.5.3. We pick $\mathcal{C}$ to be a set enclosed by a level set of a Lyapunov function of the ODE $\dot{x} = T(x) - x$. The existence of such a Lyapunov function is guaranteed under the assumption that the ODE has a global asymptotically stable equilibrium point x^* by a converse Lyapunov theorem. In fact, it turns out that a necessary and sufficient condition for an equilibrium point x^* to be asymptotically stable is the existence of a Lyapunov function on its domain of attraction $\mathcal{D}$ (see [48]). Moreover, using some smoothing techniques Wilson showed that if x^* is asymptotically stable, then there exists a C^∞ Lyapunov function on $\mathcal{D}$ (see Theorem 3.2 in [47]). Let v be a possible Lyapunov function. Then $\mathcal{C} = \{x \in \mathbb{R}^n : v(x) < C\}$. The constant C should be large enough so that $\mathcal{C}$ encloses x^* , the origin, and a ball $B(0, R)$ around it. In fact, this can be done thanks to the following lemma, which insures that the level sets can be picked arbitrarily far away from the equilibrium point, i.e. they can be picked to completely enclose an open ball $B(0, R)$ around the origin with $R > \|x^*\|_2$. For convenience, we will assume that the ODE has a unique and stable equilibrium point, i.e. the domain of attraction is $\mathbb{R}^n$.

Lemma 3.3.1 *Let $\dot{x} = h(x)$ be an ODE with a global asymptotically stable equilibrium point x^* , and let v be a smooth Lyapunov function for the ODE. For any $R > 0$, there is $C > 0$ such that the ball of radius R centered around x^* $B(x^*, R)$ is in the interior of the level set $L = \{x \in \mathbb{R}^n : v(x) \leq C\}$.*

Proof Consider the closure of $B(x^*, R)$, which is denoted by $\bar{B}(x^*, R)$. $\bar{B}(x^*, R)$ is a closed and bounded subset of $\mathbb{R}^n$, and hence is compact. As v is continuous, the minimum and maximum of v over $\bar{B}(x^*, R)$ is attained on $\bar{B}(x^*, R)$. Let $C_{max} = \max_{x \in \bar{B}(x^*, R)} v(x)$. But $v(x) > 0$ for some $x \in \bar{B}(x^*, R)$, and therefore $C_{max} > 0$. Pick any level set of the form $L = \{x \in \mathbb{R}^n : v(x) \leq C_{max} + r\}$, where $r > 0$. Such a level set will fully contain $\bar{B}(x^*, R)$.

Q.E.D.

The analysis of the scaled iterations (3.6) and (3.7) will also involve an ODE argument using the same compactness technique as the ODE limiting arguments of Kushner

and Clark [21] and Borkar [10]. Before we proceed with the analysis, we need a last remark, which justifies the choice of the scaling sets $\mathcal{B}$ and $\mathcal{C}$ above.

Remark 3.3.1 *The vector field of the ODE $\dot{x} = h(x)$ is transversal to the level sets, implying that if $x \in \partial\mathcal{C}$ (where $\mathcal{C} = \{x \in \mathbb{R}^n : v(x) < C\}$ for some C), then $(x + \Delta h(x)) \in \mathcal{C}$ for small enough $\Delta > 0$.*

Intuitively, if the stepsize is small enough, we can think of the algorithm as starting at the boundary of $\mathcal{B}$ and moving around initially in $\mathcal{C}$. As it approaches the boundary of $\mathcal{C}$, it gets pushed back to the interior of $\mathcal{C}$ thanks to the fact that the vector field of the ODE on the boundary points inward and inspite of the noise term.

3.3.1 Analysis of Scaled Iteration - Synchronous

The iteration (3.6) can be rewritten as

$$z^{k+1} = z^k + \gamma(k) \left(F(z^k, \xi^k) - z^k \right) + g^k,$$

for an appropriately defined sequence $\{g^k\}$. Note that $g^k = 0$, unless there is rescaling. Our goal is to show that the algorithm above converges to the unique equilibrium point x^* of the ODE $\dot{x} = T(x) - x$ w.p.1 under the same exact assumptions as Lemma 2.5.3, by showing that the iterates asymptotically track the same ODE.

Without loss of generality, consider the deterministic iteration

$$z^{k+1} = z^k + \gamma(k) \left(h(z^k) + w^k \right) + g^k, \tag{3.8}$$

where w^k is some noise term, and

$$g^k = \Pi_{\mathcal{B}, \mathcal{C}} \left[z^k + \gamma(k) \left(h(z^k) + w^k \right) \right] - \left[z^k + \gamma(k) \left(h(z^k) + w^k \right) \right],$$

and $\mathcal{B}$ and $\mathcal{C}$ satisfy the conditions above. For consistency, let $\tilde{x}^k = z^k + \gamma(k) \left(h(z^k) + w^k \right)$. Also let the following two sets of assumptions hold:

Assumption 3.3.1

$$\begin{aligned}\sum_{k=0}^{\infty} \gamma(k) &= \infty, \\ \sum_{k=0}^{\infty} \gamma^2(k) &< \infty.\end{aligned}$$

Assumption 3.3.2

1. *There exists D s.t. $\|w^k\| \leq D$ for all k ,*
2. *$\lim_{k \rightarrow \infty} \sum_{m=k}^{m_T(k)} \gamma(m) w^m = 0$ for all T , where*

$$m_T(k) = \min\{m \geq k : \sum_{l=k}^m \gamma(l) \geq T\}.$$

3. *h is Lipschitz continuous, i.e.*

$$\|h(x) - h(y)\| \leq L\|x - y\|$$

4. *The ODE $\dot{x} = h(x)$ has a global asymptotically stable equilibrium point,*
5. *$\mathcal{C}$ is the interior of a level set of the Lyapunov function of the ODE above, which contains the origin and x^* ,*
6. *$\mathcal{B}$ is an open ball totally enclosed in $\mathcal{C}$.*

To proceed further, we need to define piece-wise linear or piece-wise constant interpolated processes, based on the iterates $\{z^k\}$. As usual in limiting ODE arguments, the interpolated processes are over time intervals $[t_k, t_{k+1})$, where $t_k = \sum_{m=0}^{k-1} \gamma(m)$.

$$\begin{aligned}Z_l(t) &= \begin{cases} z^k & \text{for } t = t_k \\ (1 - \frac{t-t_k}{\gamma(k)})z^k + \frac{t-t_k}{\gamma(k)}\tilde{x}^{k+1} & \text{for } t \in [t_k, t_{k+1}) \end{cases}, \\ Z_c(t) &= z^k & \text{for } t \in [t_k, t_{k+1}), \\ G_c(t) &= \sum_{m=0}^{k-1} g^m & \text{for } t \in [t_k, t_{k+1}), \\ W_l(t) &= \begin{cases} \sum_{m=0}^{k-1} \gamma(m) w^m & \text{for } t = t_k \\ (1 - \frac{t-t_k}{\gamma(k)})W_l(t_k) + \frac{t-t_k}{\gamma(k)}W_l(t_{k+1}) & \text{for } t \in [t_k, t_{k+1}) \end{cases}.\end{aligned}\tag{3.9}$$

Note that $Z_l(\cdot)$ is a right-continuous (i.e. if t is a discontinuity point, then $x(t^-) \neq x(t^+)$ and $x(t^+) = x(t)$) piece-wise linear process which is continuous everywhere except at times t_k corresponding to $g^k \neq 0$. Similarly, $Z_c(\cdot)$ and $G_c(\cdot)$ are piece-wise constant processes. Moreover, $G(\cdot)$ is a constant function everywhere except for the discontinuities at times t_k corresponding to $g^k \neq 0$. Define the left-shifted versions of all the processes above as follows:

$$\begin{aligned}
Z_l^k(t) &= Z_l(t + t_k) & \text{for } t \geq -t_k, \\
W_l^k(t) &= W_l(t + t_k) - W_l(t_k) & \text{for } t \geq -t_k, \\
Z_c^k(t) &= Z_c(t + t_k) & \text{for } t \geq -t_k, \\
G_c^k(t) &= G_c(t + t_k) - G_c(t_k) & \text{for } t \geq -t_k.
\end{aligned} \tag{3.10}$$

Then, simple manipulation shows that for $t \geq -t^k$,

$$\begin{aligned}
Z_l^k(t) &= Z_l^k(0) + \int_0^t h(Z_c^k(\tau)) d\tau + W_l^k(t) + G_c^k(t) \\
&= Z_l^k(0) + \int_0^t h(Z_l^k(\tau)) d\tau + W_l^k(t) + G_c^k(t) + e^k(t),
\end{aligned}$$

where

$$e^k(t) = \int_0^t [h(Z_c^k(\tau)) - h(Z_l^k(\tau))] d\tau.$$

Assumption 3.3.2 on $\{w^k\}$ implies that $\{W_l^k(\cdot)\}$ converges to 0 uniformly on finite intervals. The remaining goal is to show that the other two terms $\{G_c^k(\cdot)\}$ and $\{e^k(\cdot)\}$ also converge to 0 in some uniform way. The appropriate lemmas and proofs are presented below.

Consider the sequences of functions $\{Z_l^k(\cdot)\}$, $\{G_c^k(\cdot)\}$, and $\{e^k(\cdot)\}$. Then the following result holds:

Lemma 3.3.2 *$\{e^k(\cdot)\}$ converges to 0 uniformly on any finite interval of the form $[0, T]$.*

Proof The lemma above results from the fact that the size of the interpolation interval and the size of the interpolation jump ($\|z^{k+1} - z^k\|$) are both proportional to

the stepsize $\gamma(k)$. Explicitly,

$$\begin{aligned}\|e^k(t)\| &\leq \int_0^t \|h(Z_c(\tau)) - h(Z_l(\tau))\| d\tau, \\ &\leq \int_0^t L \|Z_c(\tau) - Z_l(\tau)\| d\tau.\end{aligned}$$

The first inequality is due to the Cauchy-Schwartz inequality, and the second one is due to the Lipschitz continuity of T . However, for uniform convergence, we're interested in $\sup_{t \in [0, T]} \|e^k(t)\|$, which is bounded from above by $\int_0^T L \|Z_c(\tau) - Z_l(\tau)\| d\tau$. The interval $[0, T]$ can be partitioned into subintervals $[0, t_{k+1} - t_k)$, $[t_{k+1} - t_k, t_{k+2} - t_k)$, $\dots$, $[t_{m_T(k)-1} - t_k, t_{m_T(k)} - t_k)$, where

$$m_T(k) = \min\{m \geq k : \sum_{l=k}^m \gamma(l) \geq T\}.$$

It is easy to see that

$$\begin{aligned}\sup_{t \in [0, T]} \|e^k(t)\| &\leq \sum_{m=k}^{m_T(k)} \gamma(m) L \sup_{\tau \in [t_m, t_{m+1})} \|Z_l(\tau) - Z_c(\tau)\| \\ &\leq \sum_{m=k}^{m_T(k)} \gamma(m) L (\tau - t_m) \|h(z^m) + w^m\| \\ &\leq \sum_{m=k}^{m_T(k)} \gamma^2(m) L D',\end{aligned}$$

where $D' = \sup_{z \in \mathcal{C}} \|h(z)\| + D$. The second inequality above is due to the fact that for $\tau \in [t_m, t_{m+1})$,

$$\begin{aligned}Z_l(\tau) - Z_c(\tau) &= \frac{\tau - t_m}{\gamma(m)} (h(z^m) + w^m) \\ &\leq h(z^m) + w^m.\end{aligned}$$

But $\sum_{m=k}^{m_T(k)} \gamma^2(m) \rightarrow 0$ as $k \rightarrow \infty$ because $\sum_{k=0}^{\infty} \gamma^2(k) < \infty$ by Assumption 3.3.1.

Q.E.D.

As for the processes $Z_l^k(\cdot)$ and $G_c^k(\cdot)$, the topology of uniform convergence on finite intervals is not applicable due to the fact that they are not continuous. However, they are right-continuous, and an appropriate function space to consider is $D[0, 1]$, which is basically the set of right-continuous functions on finite intervals. The topology that extends the idea of uniform convergence to the $D[0, 1]$ space is the Skorohod topology (see Appendix A) and we need to use the equivalent of the Arzela-Ascoli theorem to

use compactness in order to guarantee the existence of limit points. In order to look at limit points of the sequences $\{Z_l^k(\cdot)\}$ and $\{G_c^k(\cdot)\}$ we first need to verify that they are relatively compact (i.e. they have compact closure). A theorem that guarantees relative compactness of a set A in the Skorohod topology is given in ([8], P. 119), and states that A is relatively compact if and only if the following two conditions hold:

1.

$$\sup_{x \in A} \sup_{t \in [0, T]} |x(t)| < \infty.$$

2.

$$\begin{aligned} \lim_{\delta \rightarrow 0} \sup_{x \in A} \sup_{t_1 \leq t \leq t_2, t_2 - t_1 \leq \delta} \min\{|x(t) - x(t_1)|, |x(t_2) - x(t)|\} &= 0, \\ \lim_{\delta \rightarrow 0} \sup_{x \in A} \sup_{t_1, t_2 \in [0, \delta]} |x(t_1) - x(t_2)| &= 0, \\ \lim_{\delta \rightarrow 0} \sup_{x \in A} \sup_{t_1, t_2 \in [T - \delta, T]} |x(t_1) - x(t_2)| &= 0. \end{aligned}$$

The first condition is simply the uniform boundedness condition. The second one is a generalization of the equi-continuity condition in $C[0, 1]$.

In particular, we would like to show that the sequences $\{Z_l^k(\cdot)\}$ and $\{G_c^k(\cdot)\}$ are relatively compact. Note that the process $Z_l^k(\cdot)$ has a discontinuity at time t if and only if $G_c^k(\cdot)$ has a discontinuity at time t . The following lemma states further that the points of discontinuity cannot get arbitrarily close to each other, in other words there is a minimum distance between any two consecutive discontinuity points.

Lemma 3.3.3 *There exists $\Delta > 0$ such that if t and t' are two distinct discontinuity points $G_c(\cdot)$ and $Z_l(\cdot)$, then $|t - t'| \geq \Delta$.*

Proof Suppose that there is a discontinuity at t_k for some k . Then $g^{k-1} \neq 0$, and z^k lies on the boundary of the ball $\mathcal{B}$. Let d be the minimum distance between the boundaries of $\mathcal{B}$ and $\mathcal{C}$. Note that $d > 0$ because the two sets are closed and disjoint. We claim that $z^{k+1}, z^{k+3}, \dots, z^{k+m(k)}$ are all inside $\mathcal{C}$, where

$$m(k) = \max\{j : \sum_{i=0}^j \gamma(k+i) \leq \frac{d}{D'}\},$$

and

$$D' = \sup_{z \in \mathcal{C}} \|h(z)\| + \sup_k \|w^k\|.$$

To see this, notice that if

$$\gamma(k) \leq \frac{d}{D'},$$

then

$$\|\tilde{x}^{k+1} - z^k\| \leq d.$$

Therefore

$$\tilde{x}^{k+1} \in \mathcal{C},$$

and

$$z^{k+1} = \tilde{x}^{k+1}.$$

That implies that $g^k = 0$, and there is no discontinuity at time t_{k+1} . An induction argument shows that if

$$\sum_{i=0}^{j-1} \gamma(k+i) \leq \frac{d}{D'},$$

then

$$z^{k+i} \in \mathcal{C} \quad \text{for} \quad i = 1, \dots, j.$$

This implies that

$$z^{k+i} \in \mathcal{C} \quad \text{for} \quad i = 1, \dots, m(k). \tag{3.11}$$

In other words, $g^{k+i} = 0$ for $i = 0, \dots, m(k)$. This implies that $G_c(t) = G_c(t_k)$ for $t \in [t_k, t_k + \frac{d}{D'})$. Similarly, the process $Z_l(\cdot)$ does not have any discontinuities at any times t_j , where $k < j \leq m(k)$, and hence is continuous in the interval $[t_k, t_k + \frac{d}{D'})$. The statement of the lemma follows by letting $\Delta = \frac{d}{D'}$.

Q.E.D.

The next lemma establishes the relative compactness of the sequences above. It follows directly from Lemma 3.3.3

Lemma 3.3.4 *The sequences $\{Z_l^k(\cdot)\}$ and $\{G_c^k(\cdot)\}$ are relatively compact.*

Proof The boundedness of $\{Z_l^k(\cdot)\}$ follows directly from the boundedness of the

iterates $\{z^k\}$. $\{g^k\}$ are bounded uniformly due to the boundedness of $\{z^k\}$ and $\{w^k\}$, and the Lipschitz continuity of T . For ease of notation, let $\sup_k \|g^k\| = D_g$. As there is a minimum distance Δ between any two consecutive discontinuities of $G_c^k(\cdot)$, it follows that there can be at most $\frac{T}{\Delta}$ jumps in a time interval of size T . Therefore $\{G_c^k(\cdot)\}$ are bounded uniformly by $\frac{T}{\Delta} \sup_k \{g^k\}$. To verify the relative compactness of $\{G_c(\cdot)\}$, we now only need to verify the second set of conditions needed for relative compactness.

Note that if δ is small enough, then any interval of the form $[t_1, t_2]$ s.t. $(t_2 - t_1) < \delta$ has at most one discontinuity. Therefore, for $t_1 \leq t \leq t_2$, either $G_c^k(t_1) = G_c^k(t)$ or $G_c^k(t) = G_c^k(t_2)$, implying that for small enough δ , for all $t_1, t_2 \in [0, T]$, s.t. $t_2 - t_1 \leq \delta$,

$$\min\{|G_c^k(t) - G_c^k(t_1)|, |G_c^k(t_2) - G_c^k(t)|\} = 0.$$

Therefore,

$$\lim_{\delta \rightarrow 0} \sup_k \min\{|G_c^k(t) - G_c^k(t_1)|, |G_c^k(t_2) - G_c^k(t)|\} = 0,$$

and the first part of the second condition for relative compactness is satisfied. At the end points of the interval $[0, T]$, the argument is a little bit more involved, as the distance between the end points and discontinuity points may get arbitrarily close. We need to look at the expression

$$\sup_k \sup_{\substack{|t_1 - t_2| < \delta \\ t_1, t_2 \in [0, \delta)}} \{|G_c^k(t_1) - G_c^k(t_2)|\}.$$

Note that for δ small enough, there may be at most one discontinuity in the interval $[0, \delta)$, and the size of the jump is bounded by $\gamma_\delta D'$ where $\gamma_\delta = \sup\{\gamma(k) : \gamma(k) \leq \delta\}$. Note that $\gamma_\delta \rightarrow 0$ as $\delta \rightarrow 0$. Therefore,

$$\lim_{\delta \rightarrow 0} \sup_k \sup_{\substack{|t_1 - t_2| < \delta \\ t_1, t_2 \in [0, \delta)}} \{|G_c^k(t_1) - G_c^k(t_2)|\} = 0.$$

A similar argument applies to the upper boundary, and establishes that

$$\limsup_{\delta \rightarrow 0} \sup_k \sup_{\substack{|t_1 - t_2| < \delta \\ t_1, t_2 \in [T - \delta, T]}} \{|G_c^k(t_1) - G_c^k(t_2)|\} = 0.$$

The same conditions can be verified for $\{Z_l^k(\cdot)\}$ using the facts that there is a minimum positive distance between any two discontinuity points, and that the continuity points have derivatives that are uniformly bounded due to the Lipschitz continuity of T .

Q.E.D.

Given that the sequences $\{Z_l^k(\cdot)\}$ and $\{G_c^k(\cdot)\}$ are relatively compact in the Skorohod topology, it follows that any sequence has a limiting subsequence. Let $\{n_k\}$ be a subsequence along which both $\{Z_l^k(\cdot)\}$ and $\{G_c^k(\cdot)\}$ have limit points. Denote the limit points by $Z_l(\cdot)$ and $G_c(\cdot)$, then these limit points satisfy the integral equation

$$Z(t) = Z(0) + \int_0^t h(Z(\tau))d\tau + G(t).$$

The following are properties of $Z(\cdot)$ and $G(\cdot)$.

1. $G(\cdot)$ is piece-wise constant and right-continuous and there is a minimum positive distance between any two consecutive discontinuity points.
2. $Z(\cdot)$ is right-continuous and there is a minimum positive distance between any two consecutive discontinuity points.

It is easy to see that the properties above hold using the fact that convergence in the Skorohod topology implies that there is a sequence of continuous nondecreasing functions $\lambda^k : [0, T] \rightarrow [0, T]$ such that $Z_l^k(\lambda^k(t)) \rightarrow Z(t)$ uniformly in t , and $\lambda^k(t) \rightarrow t$ uniformly in t . The following lemma states that the limiting process $G(\cdot)$ cannot have any discontinuities for $t > 0$ and follows from a simple limiting argument. The intuition behind the argument is that the number of discontinuities of the sequences $\{Z_l^k(\cdot)\}$ and $\{G_c^k(\cdot)\}$ are finite and uniformly bounded in k due to Lemma 3.3.5, and any discontinuity point τ in the limiting process must be the limit point of

discontinuity points in the corresponding sequence.

Lemma 3.3.5 *If t and t' are two distinct discontinuity points of $G(\cdot)$ and $Z(\cdot)$, then $|t - t'| \geq \Delta$, where Δ is the minimum distance between distinct discontinuity points of $\{G_c(\cdot)\}$ and $\{Z_l(\cdot)\}$.*

Proof The result follows from the fact that if $\{Z^k(\cdot)\}$ converges to $\{Z(\cdot)\}$ in the Skorohod topology, then any discontinuity point of $\{Z(\cdot)\}$ is a limit point of discontinuity points of $\{Z^k(\cdot)\}$.

Q.E.D.

In order to derive the ODE limit, we need one last lemma, which states that the limiting process $G(\cdot)$ must be constant, and hence must equal 0, and that the limiting process $Z(\cdot)$ must be continuous.

Lemma 3.3.6 *The limiting process $G(\cdot)$ is constant for all t , and the process $Z(\cdot)$ is continuous for all t , and satisfies the differential equation $\dot{Z} = h(Z)$.*

Proof The proof is by contradiction. Assume that $G(\cdot)$ and $Z(\cdot)$ have a discontinuity point at some time $\tau > 0$. Then

$$\|G(\tau^+) - G(\tau^-)\| = \Delta_\tau > 0,$$

and

$$\|Z(\tau^+) - Z(\tau^-)\| = \Delta'_\tau > 0.$$

Let $\tau_k = \lambda^k(\tau)$ be the sequence of times such that

$$G_c^k(\tau_k) \rightarrow G(\tau),$$

and

$$Z_l^k(\tau_k) \rightarrow Z(\tau).$$

Because of right-continuity it follows that

$$\begin{aligned} G_c^k(\tau_k^+) &\rightarrow G(\tau^+) = G(\tau), \\ Z_l^k(\tau_k^+) &\rightarrow Z(\tau^+) = Z(\tau). \end{aligned}$$

Similarly, it can be easily shown that

$$\begin{aligned} G_c^k(\tau_k^-) &\rightarrow G(\tau^-), \\ Z_l^k(\tau_k^-) &\rightarrow Z(\tau^-). \end{aligned}$$

This implies that given some small ϵ , for large enough k ,

$$\|G_c^k(\tau_k^+) - G_c^k(\tau_k^-)\| \geq \Delta_\tau,$$

and

$$\|Z_l^k(\tau_k^+) - Z_l^k(\tau_k^-)\| \geq \Delta'_\tau,$$

i.e. $\{\tau_k\}$ are jump points for the processes $\{G_c^k(\cdot)\}$ and $\{Z_l^k(\cdot)\}$ respectively, which implies that

$$Z_l^k(\tau_k^-) \in \mathcal{C}_c, \text{ the complement of } \mathcal{C},$$

and

$$Z_l^k(\tau_k^+) \in \partial\mathcal{B}.$$

Both sets $\mathcal{C}_c$ and $\partial\mathcal{B}$ are closed, therefore the following must hold

$$\begin{aligned} Z(\tau^-) &\in \mathcal{C}_c, \\ Z(\tau^+) &\in \partial\mathcal{B}, \end{aligned} \tag{3.12}$$

But Lemma 3.3.5 implies that there are no discontinuities in $[\tau - \Delta, \tau)$, and hence $G_c(\cdot)$ is constant over $[\tau - \Delta, \tau)$. Therefore the ODE $\dot{Z} = h(Z)$ holds for $t \in [\tau - \Delta, \tau)$. Given that $h(Z)$ is transversal to $\partial\mathcal{C}$ and that $Z(\cdot)$ is in $\mathcal{C}$, it follows that $Z(t)$ for $t \in [\tau - \Delta, \tau)$ must stay strictly in $\mathcal{C}$, and therefore $Z(\tau^-)$ must also belong to $\mathcal{C}$, contradicting equation (3.12).

Q.E.D.

The analysis above results in the following important lemma, which yields the limiting ODE of the scaled iteration (3.8) under pretty general assumptions.

Lemma 3.3.7 *Consider the algorithm given by (3.8). Let assumptions 3.3.1 and 3.3.2 hold. Then the left-shifted interpolated processes defined in (3.10) converge in the Skorohod topology over any interval $D[T_1, T_2]$ to a continuous process that satisfies the ODE $\dot{Z} = h(Z)$ over $(-\infty, \infty)$.*

Proof Lemma 3.3.6 applies to any finite interval $[T_1, T_2]$ with $T_1 < T_2$. Hence the limit process $Z(\cdot)$ satisfies the ODE over $(-\infty, \infty)$.

Q.E.D.

We need one last step to establish that the discrete-time sequence $\{z^k\}$ indeed converges to the unique equilibrium point of the ODE $\dot{x} = h(x)$. This is really a corollary of Lemma 3.3.7. To prove it, we use the standard argument of Kushner & Clark's, repeated in Section 2.6 in the convergence proof of asynchronous stochastic approximation algorithms.

Corollary 3.3.1 *The scaled algorithm given by (3.8), under assumptions 3.3.1 and 3.3.2 converges to the unique equilibrium point of $\dot{x} = h(x)$.*

3.3.2 Analysis of the Scaled Iteration - Asynchronous

The analysis of the asynchronous scaled iteration combines ideas from sections 2.6 and 3.3.1. Working with the simpler case where only one component is updated at a time and using the notation of Section 2.6, iteration 3.7 can be rewritten as

$$\begin{aligned}\tilde{x}^{k+1} &= z^k + \gamma(\nu(k, \phi^k)) \left(H^{(\phi^k)}(z^k, \xi^k) \right), \\ z^{k+1} &= \Pi_{B,C}(\tilde{x}^{k+1}),\end{aligned}$$

where $\Pi_{B,C}$ are defined as in Section 3.3.1, and the ODE of interest here is $\dot{x} = \frac{1}{n}h(x)$.

Again, without loss of generality, consider the deterministic iteration

$$z^{k+1} = z^k + \gamma(\nu(k, \phi^k)) \left(h^{(\phi^k)}(z^k) + w^k \right) + g^k, \quad (3.13)$$

where

$$g^k = \Pi_{\mathcal{B}, \mathcal{C}} \left[z^k + \gamma(\nu(k, \phi^k)) (h^{(\phi^k)} + w^k) \right] - \left[z^k + \gamma(\nu(k, \phi^k)) (h^{(\phi^k)} + w^k) \right],$$

and $\mathcal{B}$ and $\mathcal{C}$ are picked in the same manner as in the previous section, with $\mathcal{C}$ being the interior of a Lyapunov set of the ODE $\dot{x} = \frac{1}{n}h(x)$. In addition to assumptions 3.3.1 and 3.3.2, we need to impose the usual asynchronous assumption on the stepsize:

Assumption 3.3.3 *There exists $\Delta > 0$ such that for all i ,*

$$\lim_{k \rightarrow \infty} \frac{1}{k+1} \nu(k, i) \geq \Delta \quad w.p.1.$$

Define the piece-wise linear and piece-wise constant processes as in (3.9), together with their left shifts as in (3.10). All the results of the previous section will hold using the same arguments with slight modifications to take into account the time-varying nature of the algorithm (3.13). Let $\{t_k\}$ be the usual interpolation times for asynchronous algorithms given by (2.29), and $\{\mu^k\}$ be defined as in (2.30). Define the following continuous-time processes

$$\begin{aligned} \mu(t) &= \mu^k && \text{for } t \in [t_k, t_{k+1}), \\ Z_c(t) &= z^k && \text{for } t \in [t_k, t_{k+1}), \\ G_c(t) &= \sum_{m=0}^{k-1} g^m && \text{for } t \in [t_k, t_{k+1}), \\ Z_l(t) &= \begin{cases} z^k & \text{for } t = t_k, \\ (1 - \frac{t-t_k}{\gamma(k)})z^k + \frac{t-t_k}{\gamma(k)}\tilde{x}^{k+1} & \text{for } t \in [t_k, t_{k+1}), \end{cases} \\ W_l(t) &= \begin{cases} \sum_{m=0}^{k-1} \gamma(\nu(m, \phi^m))w^m & \text{for } t = t_k, \\ (1 - \frac{t-t_k}{\gamma(\nu(k, \phi^k))})W_l(t_k) + \frac{t-t_k}{\gamma(\nu(k, \phi^k))}W_l(t_{k+1}) & \text{for } t \in [t_k, t_{k+1}). \end{cases} \end{aligned} \tag{3.14}$$

Similarly, define the left-shifted processes as follows:

$$\begin{aligned}
X_l^k(t) &= X_l(t + t_k) && \text{for } t \geq -t_k, \\
W_l^k(t) &= W_l(t + t_k) - W_l(t_k) && \text{for } t \geq -t_k, \\
X_c^k(t) &= X_c(t + t_k) - G_c(t_k) && \text{for } t \geq -t_k, \\
\mu^k(t) &= \mu(t + t_k) && \text{for } t \geq -t_k.
\end{aligned} \tag{3.15}$$

Then, simple manipulation results in the following integral equation for $t \geq -t^k$,

$$Z_l^k(t) = Z_l^k(0) + \int_0^t \text{diag}(\mu^*)h(Z_l^k(\tau))d\tau + W_l^k(t) + G_c^k(t) + e^k(t) + \eta^k(t),$$

where μ^* is the limit of $\{\mu^k(\cdot)\}$ in the sense of Section 2.6 (see equation (2.32)), and

$$\begin{aligned}
e^k(t) &= \int_0^t \text{diag}(\mu^k(\tau))h(X_c^k(\tau)) - \text{diag}(\mu^k(\tau))h(X_l^k(\tau))d\tau, \\
\eta^k(t) &= \int_0^t \text{diag}(\mu^k(\tau))h(X_l^k(\tau))d\tau - \int_0^t \text{diag}(\mu^*)h(X_l^k(\tau))d\tau.
\end{aligned}$$

The convergence of $\{W_l^k(\cdot)\}$ to 0 follows from part (2) of Assumption 3.3.2 and Assumption 3.3.3. The convergences of $\{e^k(\cdot)\}$ and $\{\eta^k(\cdot)\}$ to 0 follow from the same arguments as those of Section 2.6.

To complete the analysis, we need to verify that lemmas 3.3.3, 3.3.4, 3.3.5, and 3.3.6 all extend to the asynchronous case. Lemma 3.3.3 holds for the asynchronous algorithm, and the proof goes through with $D' = \sup_{z \in C} \max_{i=1, \dots, n} \|h^{(i)}(z)\| + \sup_k \|w^k\|$. The arguments for the other 3 lemmas hold without any modifications as they follow directly from the properties of the sequences given by Lemma 3.3.3. Therefore, under the assumptions above, the asynchronous iteration given by 3.13 tracks the time-invariant $\dot{x} = \frac{1}{n}h(x)$. This result is summarized in the following lemma:

Lemma 3.3.8 *Consider the algorithm given by (3.13). Let assumptions 3.3.1, 3.3.2, and 3.3.3 hold. Then the left-shifted interpolated processes defined in (3.15) converge in the Skorohod topology over any interval $D[T_1, T_2]$ to a continuous process that satisfies the ODE $\dot{x} = \frac{1}{n}h(x)$ over $(-\infty, \infty)$.*

Similarly, the standard Kushner & Clark's argument for the case of a global asymptotically stable equilibrium point results in the following corollary:

Corollary 3.3.2 *The algorithm given by (3.13), under assumptions 3.3.1, 3.3.2, and 3.3.3 converges to the unique equilibrium point of $\dot{x} = h(x)$.*

The results above are all deterministic. To apply them to the stochastic approximation algorithm studied in this thesis, we will verify that the assumptions hold true on all Ω except on a set of measure 0. In particular the only condition that needs to be verified separately is the uniform boundedness of the noise terms, which will follow from the conditions of non-expansiveness for every sample path and the boundedness of the scaled iterates. In the next section, we will combine the results of Sections 3.2 and 3.3 to state some general convergence theorems.

3.4 Convergence Theorems

3.4.1 Analysis of the Synchronous Iteration

The goal is to state a general convergence theorem for the stochastic approximation algorithm given by equation (2.20) under the most general conditions that guarantee boundedness apriori.

Theorem 3.4.1 *Let $\{x^k\}$ be generated by the algorithm in (2.20), i.e.*

$$x^{k+1} = x^k + \gamma(k)(F(x^k, \xi^k) - x^k),$$

where $\{\xi^k\}$ is defined over some probability space $(\Omega, \mathcal{F}, \mathcal{P})$. Let the following assumption hold:

1. *$F(x, \xi)$ is non-expansive in x with respect to some norm for all ξ .*
2. *$E[F(x^k, \xi^k) | \mathcal{F}^k] = T(x^k)$ for some Lipschitz continuous function T , where*

$$\mathcal{F}^k = \sigma(x^k, \dots, x^0, \xi^{k-1}, \dots, \xi^0).$$

3. $\dot{x} = T(x) - x$ has a global asymptotically stable equilibrium point x^* ,

4. $\sum_{k=0}^{\infty} \gamma(k) = \infty$,

5. $\sum_{k=0}^{\infty} \gamma(k)^2 < \infty$.

Then the sequence $\{x^k\}$ converges to x^* w.p.1.

Proof The theorem is an application of Lemma 3.2.1 and Corollary 3.3.1. Now, consider the scaled iteration

$$\begin{aligned}\tilde{x}^{k+1} &= z^k + \gamma(k) (F(z^k, \xi^k) - z^k), \\ z^{k+1} &= \Pi_{\mathcal{B}, \mathcal{C}}(\tilde{x}^{k+1}).\end{aligned}\tag{3.16}$$

where $\mathcal{B}$ and $\mathcal{C}$ are chosen in accordance with Assumption 3.3.2, which is possible due to the assumption on the asymptotic stability of the ODE. The scaled algorithm (3.16) given above is a special case of the general scaled algorithms analyzed in Section 3.3 with

$$w^k = F(z^k, \xi^k) - E[F(z^k, \xi^k) | \mathcal{F}^k],$$

and

$$h(z) = T(z) - z,$$

with

$$\mathcal{F}^k = \sigma(\tilde{x}^k, \dots, \tilde{x}^0, z^k, \dots, z^0, \xi^{k-1}, \dots, \xi^0).$$

Note that $\{w^k\}$ is a random process, which is bounded uniformly in k w.p.1 due to the fact that $\{\|z^k\|\}$ is bounded and $F(x, \xi)$ is non-expansive in x for all ξ . The boundedness of $\{w^k\}$ implies that

$$E[\|w^k\|^2 | \mathcal{F}^k] \leq A \quad \text{for all } k,$$

for some deterministic bound A . Hence, the standard Martingale convergence results apply and imply that $\{\sum_{m=0}^{k-1} \gamma(m) w^m\}$ converges except on a set of measure 0.

Therefore,

$$\lim_{k \rightarrow \infty} \sum_{m=k}^{m_T(k)} \gamma(m) w^m = 0 \quad \text{w.p.1.} \quad (3.17)$$

It is easy to verify that the rest of the conditions of Assumptions 3.3.1 and 3.3.2 follow from the assumptions of the theorem. Therefore,

$$\{\tilde{x}^k\} \rightarrow x^* \quad \text{w.p.1.} \quad (3.18)$$

Now Lemma 3.2.1 applies with G^k defined as

$$G^k(x^k, \xi^k) = (1 - \gamma(k))x^k + \gamma(k)(F(x^k, \xi^k)).$$

It is easy to verify that for all k , G^k is a convex combination of two non-expansive functions, and is therefore itself non-expansive. Due to the fact that $\xi^k \in S$ for all k , and hence comes from a finite set, it follows that conditions 1-3 of the lemma are satisfied. Condition 4 follows from equation (3.18). Therefore $\{x^k\}$ is bounded w.p.1, and the standard ODE results (Lemma 2.5.3) apply and imply that $\{x^k\}$ converges to x^* w.p.1.

Q.E.D.

3.4.2 Analysis of the Asynchronous Iteration

The objective of this section is to develop the analogous results to Section 3.4.1 for asynchronous stochastic approximation algorithms. The results of this section do not cover asynchronous algorithms with delays and apply only to Gauss-Seidel algorithms.

Theorem 3.4.2 *Let $\{x^k\}$ as generated by the algorithm in (2.27), i.e.*

$$x_i^{k+1} = x_i^k + \gamma(\nu(k, i)) (F_i(x^k, \xi^k) - x_i^k) I\{i = \phi^k\},$$

where $\{\xi^k\}$ and $\{\phi^k\}$ are defined over a probability space $(\Omega, \mathcal{F}, P)$. Let

$$\mathcal{F}^k = \sigma(x^k, \dots, x^0, \xi^{k-1}, \dots, \xi^0, \phi^k, \dots, \phi^0).$$

Let the following assumptions hold:

1. $F(x, \xi)$ is non-expansive in x with respect to the sup norm for all ξ .
2. $E[F(x^k, \xi^k) | \mathcal{F}^k] = T(x)$ for some Lipschitz continuous function T .
3. $\dot{x} = T(x) - x$ has a global asymptotically stable equilibrium point.
4. Update rule and stepsize size conditions: Each component must get updated at a minimum average frequency, i.e. there exists $\Delta > 0$ such that for all i ,

$$\liminf_{k \rightarrow \infty} \frac{1}{k+1} \nu(k, i) \geq \Delta \quad w.p.1.$$

5. $\{\gamma(k)\}$ is an ideal tapering stepsize (see Definition 2.1.3).

Then the sequence $\{x^k\}$ converges to x^* w.p.1.

Proof As in the synchronous case, the theorem is a direct application of Lemma 3.2.1 and Corollary 3.3.2. In this case the scaled iteration is given by

$$\begin{aligned} \tilde{x}^{k+1} &= z^k + \gamma(\nu(k, \phi^k)) \left(H^{(\phi^k)}(z^k, \xi^k) \right), \\ z^{k+1} &= \Pi_{B, C}(\tilde{x}^{k+1}), \end{aligned} \tag{3.19}$$

where $H^{(i)}$, B , C are all defined as in Sections 2.6 and 3.3.2. Clearly (3.19) is a special case of the scaled asynchronous algorithms analyzed in Section 3.3.2. It is easy to verify that all the assumptions of 3.3.8 hold. In particular the boundedness of $\{w^k\}$ follows from the non-expansiveness of $H^{(i)}$ for $i = 1, \dots, n$ and the boundedness of $\{z^k\}$. The second condition on $\{w^k\}$ holds except on a set of measure 0, and the argument is basically the same as in the previous section as $E[w^k | \mathcal{F}^k] = 0$, and $E[\|w^k\|^2 | \mathcal{F}^k]$ is bounded by a deterministic bound. Therefore $\{\tilde{x}^k\}$ converges to x^* , the unique equilibrium point of the ODE.

The boundedness of $\{x^k\}$ will follow from Lemma 3.2.1 if all the assumptions hold. Let the noise sequence be the pair $\{(\xi^k, \phi^k)\}$, and define

$$G_i^k(x^k, \xi^k, \phi^k) = x_i^k + \gamma(\nu(k, i)) \left(F_i(x^k, \xi^k) - x_i^k \right) I\{i = \phi^k\}.$$

It is easy to verify that the G^k 's defined above satisfy all the assumptions of Lemma 3.2.1, and as $\{\tilde{x}^k\}$ converges to x^* , it follows that $\{x^k\}$ is bounded. Given the boundedness of $\{x^k\}$, it follows from Lemma 2.6.1 that $\{x^k\}$ converges w.p.1 to x^* , the global equilibrium point of ODE of $\dot{x} = h(x)$.

Q.E.D.

Chapter 4

Analysis of Discounted Cost and SSP Q -learning

As mentioned in Section 2.2, discounted cost problems can be formulated as stochastic shortest path problems. We will therefore only analyze stochastic shortest path (SSP) problems in this chapter. As we saw in Chapter 2, the optimal costs for discounted cost problems and SSP problems satisfy Bellman's equation given by (2.6). The Q -learning algorithm, originally formulated by Watkins [45], is an asynchronous version of the Robbins-Monro version of algorithm (2.8), and we will formulate here both the synchronous and asynchronous versions using the notation of Sections 2.5 and 2.6, and show that their convergences follow directly from theorems 3.4.1 and 3.4.2.

4.1 Analysis of SSP Synchronous Q -learning

In fact equation (2.8) can be rewritten as

$$\begin{aligned} Q^{k+1} &= Q^k + \gamma(k) [F(Q^k, \xi^k) - Q^k] \\ &= Q^k + \gamma(k) [T(Q^k) - Q^k + w^k], \end{aligned} \tag{4.1}$$

where $\{\xi^k\}$ is a sequence of vector-valued random variables over S , subject to the probability distribution $\text{Prob}(\xi_{iu}^k = j) = p_{ij}(u)$ for all k ,

$$F(Q, \xi)(i, u) = g(i, u, \xi_{iu}) + \min_v Q(\xi_{iu}, v),$$

$$(TQ)(i, u) = \sum_j p_{ij}(u) F(Q, j),$$

and

$$w^k = F(Q^k, \xi^k) - T(Q^k).$$

The assumptions imposed on the algorithm above are the usual stochastic approximation conditions on the stepsize:

Assumption 4.1.1

$$\begin{aligned} \sum_{k=0}^{\infty} \gamma(k) &= \infty, \\ \sum_{k=0}^{\infty} \gamma^2(k) &< \infty. \end{aligned}$$

The mappings F and T above have been extensively studied in [2], and discussed in Chapter 2. The following is a list of properties of the mappings F and T that play an important role in the analysis of the deterministic value iteration algorithms as well as the Q -learning algorithms.

1. T is non-expansive with respect to the sup-norm for all stochastic shortest path problems.
2. T is contractive with respect to some weighted sup-norm for stochastic shortest path problems in the case where all the policies are proper [6], and in particular is contractive with respect to the sup-norm for discounted cost problems.
3. For discounted cost problems and SSP problems, where all the policies are proper, the ODE $\dot{Q} = TQ - Q$, converges to the global fixed point Q^* of the mapping T for any initial condition. Moreover, the equilibrium point Q^* is asymptotically stable.
4. For SSP problems, under the assumption that there exists at least one proper policy, and that all improper policies have infinite cost for at least one state,

the ODE $\dot{Q} = TQ - Q$, converges to a unique equilibrium point Q^* , and the equilibrium point is asymptotically stable.

Properties 1-2 are well known results and their proofs can be found in [2] and [6]. Property 3 can be easily verified. Property 4 follows from the fact that T is a contraction mapping. In fact, an obvious potential function is $v(Q) = \|Q - Q^*\|$ where $\|\cdot\|$ is the norm with respect to which T is a contraction. Property 5 follows from the analysis of Bellman's equation for SSP problems in [2], and from the analysis of ODE maps involving non-expansive mappings in [13]. The SSP analysis implies that the set of fixed points for the mapping T contains a unique point Q^* . As T is non-expansive, it follows that any solution trajectory $Q(t)$ converges to Q^* . Moreover, Borkar's analysis in [13] implies that $\|Q(t) - Q^*\|_\infty$ is non-increasing, establishing that Q^* is a global asymptotically stable equilibrium point for the ODE.

Similarly, the F -mapping has the following properties:

1. F is sup-norm non-expansive in Q for all ξ .
2. For the synchronous algorithm, let $\mathcal{F}^k = \sigma(x^k, \dots, x^0, \xi^{k-1}, \dots, \xi^0, \phi^k, \dots, \phi^0)$, then

$$E[F(Q^k, \xi^k) | \mathcal{F}^k] = TQ^k.$$

Property 1 can be shown by the following argument: Let $Q_1, Q_2 \in \mathbb{R}^{n+m}$, then

$$\begin{aligned} F(Q_1, \xi)(i, u) - F(Q_2, \xi)(i, u) &= \min_{u'} Q_1(\xi_{iu}, u') - \min_{u'} Q_2(\xi_{iu}, u') \\ &\leq Q_1(\xi_{iu}, u_2) - Q_2(\xi_{iu}, u_2) \\ &\leq \max_{(i, u) \in \mathcal{S} \times A} |Q_1(i, u) - Q_2(i, u)| \\ &\leq \|Q_1 - Q_2\|_\infty, \end{aligned}$$

where u_2 achieves the minimum in $\min_{u'} Q_2(\xi_{iu}, u')$. A symmetric argument shows that

$$F(Q_2, \xi)(i, u) - F(Q_1, \xi)(i, u) \leq \|Q_1 - Q_2\|.$$

Property 2 follows directly from the definitions of F and T .

The properties above are enough to satisfy all the assumptions of Theorem 3.4.1, and immediately imply the following convergence theorem.

Theorem 4.1.1 *Let $\{Q^k\}$ be the sequence generated by the iteration in equation (4.1), then $\{Q^k\}$ converges to Q^* , w.p.1.*

Proof Follows immediately from Theorem 3.4.1.

Q.E.D.

4.2 Analysis of the SSP Asynchronous Q -learning

The asynchronous version of (4.1) is what is usually referred to as the Q -learning algorithm, and using the formulation in Section 2.6 it is expressed as

$$Q^{k+1}(i, u) = Q^k(i, u) + \gamma(\nu(k, \phi^k)) \left[(F(Q^k, \xi^k))(i, u) - Q^k(i, u) \right] I\{(i, u) = \phi^k\}, \quad (4.2)$$

where $\{\xi^k\}$ is as defined above, and $\{\phi^k\}$ is a random process taking values over $S \times A$. Again, the assumptions imposed on the algorithm are the usual asynchronous stochastic approximation assumptions on the stepsize

Assumption 4.2.1 1.

$$\liminf_{k \rightarrow \infty} \frac{1}{k+1} \nu(k, i) \geq \Delta \quad \text{for some } \Delta > 0.$$

2. $\{\gamma(k)\}$ is an ideal tapering stepsize (see Definition 2.1.3).

The same properties hold for the asynchronous algorithm, with $\mathcal{F}^k = \sigma(x^k, \dots, x^0, \xi^{k-1}, \dots, \xi^0, \phi^k, \dots)$ for Property 2. Again, the assumptions of Theorem 3.4.2 follow directly from the properties of the F and T mappings.

Theorem 4.2.1 *Let $\{Q^k\}$ be the sequence generated by the iteration in equation (4.2), then $\{Q^k\}$ converges to Q^* , w.p.1.*

Proof Follows immediately from Theorem 3.4.2.

Q.E.D.

4.3 Summary of Convergence Results

The results above cover three types of problems:

1. discounted cost problems,
2. stochastic shortest path problems where all the policies are proper,
3. stochastic shortest path problems where there is a proper policy, and all improper policies have infinite cost.

The convergence results given by theorems 4.1.1 and 4.2.1 provide alternate proofs for the convergence of Q -learning for discounted cost problems and stochastic shortest path problems, where all policies are proper. This is a slight extension over Tsitsiklis' result for general stochastic shortest path problems, which imposes the boundedness condition for the SSP problems where not all policies are proper. However, it is important to note that the general theory developed in this thesis does not deal with delays, while Tsitsiklis' proof of convergence of Q -learning algorithms allows for delays.

Chapter 5

Q -learning Algorithms for Average Cost Problems

5.1 Review of existing proposed algorithms

In the section, we will briefly discuss the Q -learning algorithms that have been proposed in the literature. There are basically three other authors that have looked into Q -learning for average cost problems, and none of them has proved the convergence of the proposed algorithm. Schwartz [34] proposes a two-time scale algorithm, which he denotes by R -learning, which carries out a value iteration-type of step to update the values of the state-action pairs, and updates an estimate of the optimal average cost using the immediate reward along with an adjustment factor. The idea behind the R -learning algorithm is to obtain a good estimate of the average cost while searching for the optimal policy via a value iteration-type of update. Although Schwartz presents some intuitive arguments to justify his algorithm along with some experimental results, he does not provide any analysis and there is no guarantee that the algorithm converges. Singh [36] presents two Q -learning algorithms for the average cost problem: one is a slight modification of Schwartz's R -learning algorithm, which updates the estimate of the optimal average cost at every step. The other one updates the estimate of the average cost in a fashion similar to the Jalali and Ferguson's deterministic asynchronous algorithm for average cost problems [18]. Singh

provides some simulation results on a medium-size problem but does not provide a convergence analysis. Finally Mahadevan [27] gives a nice discussion about average cost problems, and the need to consider the average cost criterion with an emphasis on the difference between gain-optimal policies and bias-optimal policies. He carries out extensive numerical experiments on R -learning and shows the possible problems that the algorithm can run into. However, he does not provide much insight into the analysis. Note that all the above authors have stayed away from the relative value iteration algorithm as the basis of an asynchronous stochastic approximation algorithm due to the fact that the relative value iteration algorithm for the average cost problem does not converge asynchronously (see [3]). However, a discussion in Appendix B shows that the introduction of a small enough stepsize or a diminishing stepsize may make the asynchronous algorithm convergent.

The two Q -learning algorithms which we propose in this thesis are based on two different value iterations for the average cost problem. The first one is based on a Robbins-Monro stochastic approximation variant of the relative value iteration algorithm. The second one is based on a stochastic approximation variant of the value iteration proposed by Bertsekas [4].

5.2 Average Cost Q -Learning

The minimization in the relative value iteration algorithm (2.14) is inside the expectation. This suggests that a stochastic version of the algorithm above might work. In fact, the synchronous and asynchronous stochastic variants of (2.14) yield algorithms, which we will refer to, respectively, as the synchronous and asynchronous average cost Q -learning algorithms. The synchronous Q -learning algorithm is

$$\begin{aligned}
 Q^{k+1}(i, u) &= Q^k(i, u) + \gamma(k) \left(g(i, u, \xi_{iu}^k) + \min_{u'} Q^k(\xi_{iu}^k, u') - f(Q^k) - Q^k(i, u) \right), \\
 &\text{for all } (i, u) \in S \times A,
 \end{aligned}
 \tag{5.1}$$

where $\{\xi^k\}$ is a sequence of vector-valued random variables over S , subject to the probability distribution $\text{Prob}(\xi_{iu}^k = j) = p_{iu}(j)$ for all k , $f : \mathbb{R}^{n+m} \rightarrow \mathbb{R}$ is a scalar

function, which is subject to conditions that will be specified later, the stepsize $\{\gamma(k)\}$ satisfies the usual stochastic approximation conditions

Assumption 5.2.1

$$\begin{aligned}\sum_{k=0}^{\infty} \gamma(k) &= \infty, \\ \sum_{k=0}^{\infty} \gamma^2(k) &< \infty.\end{aligned}$$

The asynchronous Q -learning algorithm is

$$\begin{aligned}Q^{k+1}(i, u) &= Q^k(i, u) + \gamma(\nu(k, i, u)) \\ &\quad \left(g(i, u, \xi_{iu}^k) + \min_{u'} Q^k(\xi_{iu}^k, u') - f(Q^k) - Q^k(i, u) \right) I((i, u) \in Y^k) \\ &\quad \text{for all } (i, u) \in S \times A,\end{aligned}\tag{5.2}$$

where ξ^k and f are as described above, $\{Y^k\}$ is a random process taking on values in the collection of all possible subsets of $S \times A$ and $\nu(k, i, u) = \sum_{i=0}^k I((i, u) \in Y^k)$ is the number of times the state-action pair (i, u) has been visited at time k . As is usual in asynchronous stochastic approximation algorithms, it will be necessary to impose the following assumptions on the stepsize and the update rule:

Assumption 5.2.2

1. *Each state-action pair must be updated infinitely often:*

$$\liminf_{k \rightarrow \infty} \frac{\nu(k, i, u)}{k+1} \geq \Delta \quad w.p.1 \quad \forall (i, u) \in S \times A.$$

2. *$\{\gamma(k)\}$ is an ideal tapering stepsize (see Definition 2.1.3).*

For ease of notation, define the mappings F, F', T, T' as follows:

$$\begin{aligned}F(Q, \xi)(i, u) &= g(i, u, \xi_{iu}) + \min_{u'} Q(\xi_{iu}, u'), \\ F'(Q, \xi) &= F(Q, \xi) - f(Q)e, \\ (TQ)(i, u) &= \sum_{j=1}^n p_{ij}(u) \left[g(i, u, j) + \min_{u'} Q(j, u') \right], \\ T'Q &= TQ - f(Q)e.\end{aligned}$$

As was discussed in Section 2.3 the mappings T , and T' have important properties that play an important role in the analysis of the relative iteration algorithm. The

mappings F and F' have similar properties that will be essential to the analysis of the Q -learning algorithms above. We will re-state them below without repeating the arguments.

1. $F(Q, \xi)$ is non-expansive with respect to the sup-norm for all ξ .
2. F' non-expansive in Q with respect to span semi-norm for all ξ .
3. Let $\mathcal{F}^k = \sigma(Q^k, \dots, Q^0, \xi^{k-1}, \dots, \xi^0, Y^k, \dots, Y^0)$. Then

$$E \left[F'(Q^k, \xi^k) | \mathcal{F}^k \right] = T'(Q).$$

In Chapter 6, we will carry out the analysis of both the synchronous and asynchronous Q -learning algorithms based on the properties above.

5.3 SSP Average Cost Q -Learning

The other Q -learning algorithm we will analyze is based on a value iteration algorithm for average cost problems proposed by Bertsekas [4], which is based on the connection between average cost problems and stochastic shortest path problems. This connection is presented thoroughly in volume I [2]. In this section, we define the problem exactly in the same way as in Section 2.3, with the following additional assumptions.

1. Every stationary policy is unichain.
2. There exists a single recurrent state s under every policy.

Under these assumptions, the recurrent state s may be viewed as a reference state for each cycle, and the average cost of any policy is simply the ratio of the expected cost and the expected number of stages (expected time) between consecutive visits of the state s . The expected cost and time between consecutive visits are solutions to two separate stochastic shortest path problems, defined over the same transition probability model. The first one retains the same cost structure, while for the second one, the cost of any transition is simply 1. The idea of viewing the solution of the

average cost problem as resulting from two stochastic shortest path problems is rather simple, but generalizing it to the optimization problem with multiple policies is not so simple. This is done nicely by Bertsekas in [4], and we will briefly describe the ideas behind this SSP view of the average cost problem, as they are the basis for the Q -learning algorithm that we propose.

We will define a general class of stochastic shortest path problems whose cost structure is parametrized by λ , which we will refer to as the λ -problems. The parameters for each problem are as follows:

- The state space is $S' = S \cup \{s'\}$, where s' is an artificially added termination state (i.e. zero-cost and absorbing).
- The control constraint set is the same as before for all states in S . There is only one possible control at state s' .
- The transition probabilities are: For all $(i, u) \in S \times A$,

$$\begin{aligned} p'_{ij}(u) &= p_{ij}(u) & \text{if } j \neq s, \\ p'_{ij}(u) &= 0 & \text{if } j = s, \\ p'_{ij}(u) &= p_{is}(u) & \text{if } j = s'. \end{aligned}$$

- The costs are defined as follows: For all $(i, u) \in S \times A$,

$$\begin{aligned} g'(i, u, j) &= g(i, u, j) - \lambda & \text{if } j \neq s, \\ g'(i, u, j) &= 0 & \text{if } j = s, \\ g'(i, u, j) &= g(i, u, j) - \lambda & \text{if } j = s'. \end{aligned}$$

The recurrence assumption on s implies that under every policy, the termination state is reached with probability 1. Therefore, for every λ , the λ -problem is a stochastic shortest path problem that satisfies the condition that all policies are proper. Denote by h_λ the solution of the SSP λ -problem. Then, Bellman's equation is

$$h_\lambda = Th_\lambda - \lambda e, \tag{5.3}$$

where the T -mapping is defined by

$$\begin{aligned}(Th)(i) &= \min_u \sum_{j \neq s} p_{ij}(u) [g(i, u, j) + h(j)] \quad \text{for } i = 1, \dots, n, \\ h(s') &= 0.\end{aligned}$$

In fact, it can be shown that h_λ is concave and piecewise-linear in λ for each component. Intuitively, this is because h_λ is clearly continuous, and for each policy the cost for the SSP λ -problem is linear in λ with a negative slope. Therefore, the cost of the optimal policy is piece-wise linear and concave. Any linear piece corresponds to a range of λ with the same optimal policy. Therefore, the algorithm,

$$\lambda^{k+1} = \lambda^k + \beta h_{\lambda^k}(s), \tag{5.4}$$

converges for a small enough constant stepsize β . The analysis above motivates the value iteration algorithm proposed by Bertsekas [4], which is a combination of the value iteration algorithm for Bellman's equation (5.3) and (5.4).

$$\begin{aligned}h^{k+1} &= Th^k - \lambda^k e, \\ \lambda^{k+1} &= \lambda^k + \beta(k) h^k(s).\end{aligned} \tag{5.5}$$

The convergence of the algorithm above is established in [4] under the condition that the stepsize $\beta(k)$ be kept within some interval of positive numbers. The SSP Q -learning algorithm is basically a stochastic Gauss-Seidel version of (5.5).

Let Q be defined as in the regular average cost Q -learning algorithm. We will define two SSP average cost Q -learning algorithms below: the first one, which will be denoted by SSP1, does not involve any projections, and its analysis will require imposing the boundedness condition. The second one, which will be denoted by SSP2, forces the iterates to be bounded by use of projection. The SSP1 algorithm is

as follows:

$$\begin{aligned}
Q^{k+1}(i, u) &= Q^k(i, u) + \gamma(\nu(k, i, u)) \\
&\quad \left[\left(g(i, u, \xi_{iu}^k) + \min_{u'} Q^k(\xi_{iu}^k, u') \right) I(\xi_{iu}^k \neq s) - \lambda^k - Q^k(i, u) \right] I((i, u) = \phi^k), \\
\lambda^{k+1} &= \lambda^k + \beta(k) \min_u Q^k(s, u),
\end{aligned} \tag{5.6}$$

where $\{\xi^k\}$, $\{\phi^k\}$ are defined in the same way as for the regular average cost Q -learning algorithm.

The SSP2 algorithm is basically the same as the SSP1 algorithm, except that the iterates $\{\lambda^k\}$ are kept bounded by projection onto an interval:

$$\begin{aligned}
Q^{k+1}(i, u) &= Q^k(i, u) + \gamma(\nu(k, i, u)) \\
&\quad \left[\left(g(i, u, \xi_{iu}^k) + \min_{u'} Q^k(\xi_{iu}^k, u') \right) I(\xi_{iu}^k \neq s) - \lambda^k - Q^k(i, u) \right] I((i, u) = \phi^k), \\
\lambda^{k+1} &= \Gamma_{[-K, K]} \left(\lambda^k + \beta(k) \min_u Q^k(s, u) \right),
\end{aligned} \tag{5.7}$$

where K is a positive constant such that $-K \leq \lambda^* \leq K$, and $\Gamma_{[-K, K]}(\cdot)$ is the projection map onto the interval $[-K, K]$. Thus this algorithm assumes the knowledge of upper and lower bounds of λ^* . Such bounds are provided by $K = \max_{i, u, j} g(i, u, j)$.

For both algorithms above, we need to impose the appropriate asynchronous two-time scale assumptions on the stepsizes.

Assumption 5.3.1

1. *There is $\Delta > 0$ such that*

$$\liminf_{k \rightarrow \infty} \frac{\nu(k, i, u)}{k+1} \geq \Delta \quad w.p.1 \quad \forall (i, u) \in S \times A.$$

2. *The stepsizes $\{\gamma(k)\}$ and $\{\beta(k)\}$ are ideal tapering stepsizes (see Definition 2.1.3), and they satisfy*

$$\limsup_{k \rightarrow \infty} \frac{\beta(k)}{\gamma(k)} = 0.$$

Our goal is to show that the Q -learning algorithm converges w.p.1 to $(Q_{\lambda^*}, \lambda^*)$, where λ^* is the optimal average cost, and $Q(\lambda^*)$ is a solution to the average cost Bellman's

equation, which satisfies $\min_u Q_{\lambda^*}(s, v) = 0$. The analysis of the SSP1 algorithm is a straightforward application of the two-time scale asynchronous convergence results of Lemma 2.7.2. While the analysis of the SSP2 algorithm will not follow directly from the two-time scale analysis, the same arguments can be tailored to take into account the projection of $\{\lambda^k\}$ and will yield the same results. The boundedness of the sequence $\{Q^k\}$ will follow from Tsitsiklis' boundedness results for stochastic shortest path Q -learning algorithms [40].

Chapter 6

Analysis of Average Cost Q -Learning

In this chapter, we will analyze the average cost Q -learning algorithm, defined in Section 5.2, using the ODE technique. We will prove convergence of the synchronous algorithm without the apriori assumption on boundedness. However, the convergence proof of the asynchronous algorithm will require the boundedness assumption. For both convergence proofs, it is necessary to carry out the analysis of the ODE associated with the average cost algorithm. We will show that it has a global asymptotically stable equilibrium point. This by itself implies that both the synchronous and asynchronous algorithms converge based on Lemmas 2.5.3 and 2.6.1 if the boundedness assumption is imposed. Although the boundedness lemmas of Chapter 3 are not directly applicable to the average cost problem, we will be able to use Lemma 3.2.2, which exploits the non-expansiveness of the average cost mapping with respect to the span semi-norm to prove that the iterates remain bounded with probability one. In summary, this chapter has three objectives:

1. analyzing the average cost ODE,
2. proving convergence w.p.1 of the synchronous and asynchronous Q -learning algorithms under the boundedness assumption,
3. proving the boundedness w.p.1 of the synchronous Q -learning algorithm.

Before proceeding, we will rewrite the general form for the average cost Q -learning algorithm based on the notation and definitions of Chapter 5. The synchronous algorithm is

$$\begin{aligned} Q^{k+1} &= Q^k + \gamma(k) [F(Q^k, \xi^k) - f(Q^k)e - Q^k] \\ &= Q^k + \gamma(k) [T(Q^k) - f(Q^k)e - Q^k + w^k], \end{aligned} \quad (6.1)$$

where $w^k = F(Q^k, \xi^k) - T(Q^k)$. Similarly, the asynchronous algorithm is

$$Q^{k+1}(i, u) = Q^k(i, u) + \gamma(\nu(k, \phi^k)) [(F(Q^k, \xi^k))(i, u) - f(Q^k) - Q^k(i, u)] I\{(i, u) = \phi^k\}, \quad (6.2)$$

where the assumptions are the same as in Chapter 5. The assumptions on f will be explored in the next section.

6.1 ODE Analysis

The ODE associated with algorithms (6.1) and (6.2) is given by

$$\begin{aligned} \dot{Q} &= TQ - f(Q)e - Q \\ &= T'Q - Q. \end{aligned} \quad (6.3)$$

Its analysis does not follow directly from the analysis of ODEs with non-expansive maps of Borkar [9] because of the additional term $f(Q)e$. However, we will be able to exploit the two properties

1. non-expansiveness of T' with respect to the span semi-norm,
2. non-expansiveness of T with respect to the sup-norm,

as well as other dynamic programming-type of properties to show that the ODE (6.3) converges to a global asymptotically stable equilibrium point. We will do this through a series of lemmas. The first lemma relates the trajectory of the ODE above to the ODE $\dot{Q} = TQ - Q$.

Lemma 6.1.1 *Let T be non-expansive w.r.t. to the span semi-norm. Let $x(t)$ be the trajectory of $\dot{x} = T(x) - x - f(x)e$, with initial condition $x(0) = x_0$, where f is a continuous map from $\mathbb{R}^n$ to $\mathbb{R}$. Let $y(t)$ be the trajectory of $\dot{y} = T(y) - y$, with initial condition $y(0) = x_0$, then*

$$x(t) = y(t) + r(t)e \quad \forall t. \quad (6.4)$$

Moreover if F and f satisfy the following conditions:

$$T(x + re) = T(x) + re \quad \forall x, r, \quad (6.5)$$

$$f(x + re) = f(x) + rf(e) \quad \forall x, r, \quad (6.6)$$

then $r(t)$ satisfies the ODE

$$\dot{r} = -r(t)f(e) - f(y(t)), \quad r(0) = 0. \quad (6.7)$$

Proof Using the variation of constants formula (see Appendix A), we get:

$$\begin{aligned} x(t) &= x_0 e^{-t} + \int_0^t e^{-(t-\tau)} T(x(\tau)) d\tau - \left[\int_0^t e^{-(t-\tau)} f(x(\tau)) d\tau \right] e \\ y(t) &= x_0 e^{-t} + \int_0^t e^{-(t-\tau)} T(y(\tau)) d\tau \end{aligned}$$

Therefore,

$$x(t) - y(t) = \int_0^t e^{-(t-\tau)} (T(x(\tau)) - T(y(\tau))) d\tau - \left[\int_0^t e^{-(t-\tau)} f(x(\tau)) d\tau \right] e$$

Also

$$\max_i (x_i(t) - y_i(t)) \leq \int_0^t e^{-(t-\tau)} \max_i (T_i(x(\tau)) - T_i(y(\tau))) d\tau - \int_0^t e^{-(t-\tau)} f(x(\tau)) d\tau$$

and

$$\min_i (x_i(t) - y_i(t)) \geq \int_0^t e^{-(t-\tau)} \min_i (T_i(x(\tau)) - T_i(y(\tau))) d\tau - \int_0^t e^{-(t-\tau)} f(x(\tau)) d\tau$$

subtracting the equations above from each other yields,

$$\begin{aligned}\|x(t) - y(t)\|_s &\leq \int_0^t e^{-(t-\tau)} \|T(x(\tau)) - T(y(\tau))\|_s d\tau, \\ &\leq \int_0^t e^{-(t-\tau)} \|x(\tau) - y(\tau)\|_s d\tau.\end{aligned}\tag{6.8}$$

The last equation is due to the fact that F is non-expansive with respect to the span semi-norm. Applying Gronwall's lemma (see A.2 with $b = 0$ and $c = 1$) to the last inequality implies that for $t \geq 0$

$$\|x(t) - y(t)\|_s \leq 0,$$

and hence

$$\|x(t) - y(t)\|_s = 0 \quad \forall t \geq 0.$$

Therefore,

$$x(t) = y(t) + r(t)e,\tag{6.9}$$

for some scalar function $r(\cdot)$ on $\mathbb{R}^+$ which satisfies $r(0) = 0$. If the functions F and f satisfy the conditions (6.5) and (6.6), respectively, then taking the derivative of both sides of equation (6.9) yields:

$$\begin{aligned}\dot{r}(t) &= \dot{x}(t) - \dot{y}(t) \\ &= (Tx(t) - x(t) - f(x(t))e) - (Ty(t) - y(t)) \\ &= -f(e)r(t) - f(y(t)),\end{aligned}\tag{6.10}$$

with the initial condition $r(0) = 0$.

Q.E.D.

The conditions (6.5) and (6.6) on the functions F and f are not restrictive at all, and the result of the lemma will be useful for the analysis of the average cost ODE as the conditions on F and f will be naturally satisfied.

Next, we will restate the result by Borkar and Soumayanath [13] on the stability of ODE's involving sup-norm non-expansive mappings.

Lemma 6.1.2 *Let T be non-expansive with respect to the sup-norm. Assume that G , the set of fixed points of T , is nonempty, then the trajectory $y(t)$ of the ODE $\dot{y} = T(y) - y, y(0) = y_0$ satisfies the following conditions:*

1. $\|y(t) - z\|_\infty$ is a non-increasing function of t , for all $z \in G$.
2. $y(t)$ converges to some $y^* \in G$, and the equilibrium point y^* it converges to depends on the initial condition.

Proof The lemma above is basically Theorem 4.1 in [13].

Q.E.D.

Now, it will be convenient to carry out the transformation $x(t) = Q(t) - \lambda^* e$, where λ^* is the optimal average cost for the problem. Then it is easy to verify that

$$\dot{x} = Tx - \lambda^* e - x - f(x)e. \quad (6.11)$$

In order to be able to show that the ODE (6.11) is stable, it will be necessary to impose further assumptions on the scalar function f , and these are summarized below:

Assumption 6.1.1

1. f is Lipschitz continuous, i.e. $|f(x) - f(y)| \leq L_f \|x - y\|_\infty$ for some $L_f \in \mathbb{R}$.
2. $f(x + re) = f(x) + rf(e)$, for $r \in \mathbb{R}$.
3. $f(e) > 0$.

Some examples of functions f that satisfy the assumptions above are:

1. $f(Q) = Q(s, v)$, for some fixed action-state pair (s, v) .
2. $f(Q) = \frac{1}{d} \sum_{(i,u) \in S \times A} Q(i, u)$, where d is the total number of state-action pairs.
3. $f(Q) = \min_v Q(s, v)$, where some fixed state s .

4. $f(Q) = \frac{1}{n} \sum_{i \in S} \min_u Q(i, u)$, where n is the total number of states.

Under the assumptions above, we are able to prove that the ODE (6.11) is stable and that it has a global asymptotically stable equilibrium point.

Lemma 6.1.3 *Consider the average cost ODE given by (6.11). If f satisfies Assumption 6.1.1, then the ODE has a unique asymptotically stable equilibrium point x^* , which satisfies $f(x^*) = 0$.*

Proof Let $G = \{y \in \mathbb{R}^{n+m} : Ty - \lambda^*e = y\}$, then from the properties of Bellman's equation for the average cost problem, we know that the identity

$$y = Ty - \lambda^*e \tag{6.12}$$

has several fixed points. i.e. G is nonempty, and moreover if y^* is a fixed point of (6.12), then $G = \{y^* + re : r \in \mathbb{R}\}$. Now, consider the related ODE

$$\dot{y} = Ty - \lambda^*e - y, \quad y(0) = x_0. \tag{6.13}$$

In this case, lemma 6.1.2 applies as $Ty - \lambda^*e$ is non-expansive with respect to the sup-norm, and implies that the solution trajectory $y(t)$ converges to some $y^* \in G$. Let G' denote the set of equilibrium points of the ODE (6.11). Then,

$$G' = \{y \in G : f(y) = 0\}. \tag{6.14}$$

The identity above can be easily verified. In fact, if

$$y \in \{y \in G : f(y) = 0\},$$

then

$$y = Ty - \lambda^*e,$$

and

$$f(y) = 0.$$

Therefore,

$$y = Ty - \lambda^* e - f(y)e, \quad \text{and hence} \quad y \in G'.$$

Also, if

$$y \in G',$$

then

$$y = Ty - \lambda^* e - f(y)e.$$

But, from the properties of Bellman's equation for the average cost problem, the equation $y = Ty - \lambda e$ admits a solution if and only if $\lambda = \lambda^*$. Therefore,

$$f(y) = 0 \quad \text{and hence} \quad y \in G.$$

Applying Lemma 6.1.1 to the ODE's (6.11) and (6.13) yields that

$$x(t) = y(t) + r(t)e,$$

where

$$\dot{r}(t) = -f(e)r(t) - f(y(t)) \quad \text{with initial condition } r(0) = 0.$$

The variation of constants formula (see A.1) gives $r(t)$ explicitly as

$$r(t) = - \int_0^t e^{-f(e)(t-\tau)} f(y(\tau)) d\tau. \tag{6.15}$$

Intuitively, if $y(t)$ converges to y^* , for some $y^* \in G$, then $r(t)$ will converge to

$$\begin{aligned} r^* &= - \int_0^t e^{-f(e)(t-\tau)} f(y^*) d\tau \\ &= - \frac{f(y^*)}{f(e)}. \end{aligned} \tag{6.16}$$

Equation (6.16) follows directly from (A.3) in Appendix A, and it implies that $x(t)$ converges to

$$x^* = y^* - \frac{f(y^*)}{f(e)} e.$$

As observed earlier, x^* satisfies $f(x^*) = 0$.

Next, we would like to show that G' contains a unique point. Let x_1^* and x_2^* be the limit points of the ODE (6.11) corresponding to the initial conditions $x_1(0)$ and $x_2(0)$, respectively. Similarly, Let y_1^* and y_2^* be the limit points of the ODE (6.13) corresponding to the initial conditions $x_1(0)$ and $x_2(0)$, respectively. We have just shown that:

$$\begin{aligned} x_1^* &= y_1^* - \frac{f(y_1^*)}{f(e)}e \\ x_2^* &= y_2^* - \frac{f(y_2^*)}{f(e)}e. \end{aligned}$$

But due to the properties of the solutions of the average cost Bellman's equation, $y_2^* = y_1^* + se$ for some $s \in \mathbb{R}$. The second condition on F from Lemma 6.1.3 implies that

$$\begin{aligned} x_2^* &= y_1^* + se - \frac{f(y_1^*)}{f(e)}e - \frac{sf(e)}{f(e)}e, \\ &= x_1^*. \end{aligned} \tag{6.17}$$

So far, we have shown that $x(t)$ converges to a limit point for any initial condition. G' contains a unique point x^* , and hence is a global equilibrium point. To further prove that this equilibrium point is asymptotically stable, we need to show that $\|x(t) - x^*\|$ can be made arbitrarily small by choosing sufficiently small $\|x(0) - x^*\|$. The following holds:

$$x(t) - x^* = (y(t) - y^*) - \left[\int_0^t e^{-f(e)(t-\tau)} [f(y(\tau)) - f(y^*)] d\tau \right] e. \tag{6.18}$$

Therefore,

$$\begin{aligned} \|x(t) - x^*\|_\infty &\leq \|y(t) - y^*\| + \int_0^t e^{-(t-\tau)} |f(y(\tau)) - f(y^*)| d\tau \\ &\leq \|y(t) - y^*\| + L_f \int_0^t e^{-(t-\tau)} \|y(t) - y^*\| d\tau, \\ &\leq \left(1 + \frac{L_f}{f(e)}\right) \|y(0) - y^*\|. \end{aligned} \tag{6.19}$$

The first inequality follows from the triangle inequality, the second one follows from the Lipschitz continuity of f , and the third one follows from Lemma 6.1.2, which states that $\|y(t) - y^*\|$ is non-increasing in t . But recall that $y(0) = x(0)$. Moreover, define the map $L : \mathbb{R}^n \rightarrow \mathbb{R}^n$ as $L(z) = y^*(z)$, where $y^*(z)$ is the limit point of the ODE (6.13) with initial condition $y(0) = z$. It is easy to verify that the map L is continuous, and that $L(x^*) = x^*$ as x^* is an equilibrium point for both ODE's (6.11) and (6.13). The objective is to show that $\|y(0) - y^*\|$ can be made arbitrarily small by making $\|y(0) - x^*\|$ small enough.

$$\begin{aligned}\|y(0) - y^*\| &\leq \|y(0) - x^*\| + \|y^* - x^*\| \\ &\leq \|y(0) - x^*\| + \|L(y(0)) - L(x^*)\|.\end{aligned}$$

The last term is due to the fact that x^* is an equilibrium point of the ODE (6.13) and hence $L(x^*) = x^*$. Due to the continuity of L , given any $\epsilon > 0$, there is $\delta > 0$ such that

$$\|L(y(0)) - L(x^*)\| < \frac{\frac{\epsilon}{2}}{(1 + \frac{L_f}{f(e)})}$$

if

$$\|y(0) - x^*\| < \delta.$$

Let $R_\epsilon = \min(\delta, \frac{\frac{\epsilon}{2}}{(1 + \frac{L_f}{f(e)})})$. Then equation (6.19) implies that $\|x(t) - x^*\| < \epsilon$ for $t \geq 0$ if $\|x(0) - x^*\| < R_\epsilon$, which establishes the asymptotic stability of x^* .

Q.E.D.

The analysis of the ODE (6.3) follows immediately from the results above, as

$$Q(t) = x(t) + \lambda^* e.$$

It also has a unique equilibrium point $Q^* = x^* + \lambda^* e$, and the asymptotic stability follows directly from the fact that $\|Q(t) - Q^*\| = \|x(t) - x^*\|$ for all $t \geq 0$. The above is summarized in the following corollary:

Corollary 6.1.1 *If f satisfies the conditions of Lemma 6.1.3, then the ODE (6.3) has a global asymptotically stable equilibrium point Q^* , and $f(Q^*) = \lambda^*$.*

6.2 Convergence Theorems

If we impose the boundedness condition, then Corollary 6.1.1 along with Lemma 2.5.3 and Lemma 2.6.1 imply the convergence w.p.1 of a general class of both synchronous and asynchronous algorithms Q -learning algorithms for the average cost problem. For the synchronous algorithm, the convergence theorem is as follows:

Theorem 6.2.1 *Let $\{Q^k\}$ be generated according to the synchronous Q -learning algorithm (6.1), and let assumptions 5.2.1 and 6.1.1 hold. If $\{Q^k\}$ is bounded, then $\{Q^k\}$ converges w.p.1. to Q^* , the unique solution to Bellman's equation with $f(Q) = \lambda^*$.*

Similarly, for the asynchronous algorithm, we have

Theorem 6.2.2 *Let $\{Q^k\}$ be generated according to the asynchronous Q -learning algorithm (6.1), and let assumptions 5.2.2 and 6.1.1 hold. If $\{Q^k\}$ is bounded, then $\{Q^k\}$ converges w.p.1. to Q^* , the unique solution to Bellman's equation with $f(Q) = \lambda^*$.*

In the next section, we will show that for the synchronous algorithm, the iterates $\{Q^k\}$ remain bounded w.p.1. if we impose an additional non-restrictive assumption on f .

6.3 Boundedness for the Synchronous Q -learning algorithm

In addition to Assumption 6.1.1, let us impose on f

Assumption 6.3.1

1. f is non-expansive, i.e. $|f(Q_1) - f(Q_2)| \leq \|Q_1 - Q_2\|_\infty$.
2. $f(Q + re) = f(Q) + rf(e)$, for $r \in \mathbb{R}$.
3. $f(e) = 1$.

$$4. \min_{(i,u)} Q(i, u) \leq f(Q) \leq \max_{(i,u)} Q(i, u).$$

Lemma 6.3.1 *Let $\{Q^k\}$ be generated according the synchronous average cost Q -learning algorithm (6.1). If in addition to Assumption 6.1.1, f also satisfies Assumption 6.3.1, then $\{Q^k\}$ is bounded w.p.1.*

Proof Lemma 3.2.2 along with Corollary 3.3.1 imply the boundedness of $\{Q^k\}$ in the span semi-norm. It is straightforward to see that Corollary 3.3.1 applies to the scaled version of (6.1) as all the assumptions follow directly from the properties of F' mapping and the stability of the corresponding ODE. Given that the scaled version converges w.p.1, Lemma 3.2.2 applies with

$$G^k(Q, \xi) = Q + \gamma(k) [F(Q, \xi) - f(Q)e - Q].$$

The assumptions of the lemma on $\{G^k(\cdot)\}$ follow directly from the properties of the F' mapping listed in Chapter 5 and from the convergence of the scaled iteration. Therefore the span semi-norms of the iterates are bounded. In other words, there is a B such that $\|Q^k\|_s \leq B$. Let $D = \max(\|Q^0\|, \max_{(i,u,j)} g(i, u, j) + B)$. We claim that $\{\|Q^k\|\}$ is bounded by D . The argument will be by induction. The claim is clearly true for $\|Q^0\|$. Suppose it's true for some k , then the assumptions imposed on f imply that:

$$\begin{aligned} |\min_u Q^k(\xi_{iu}^k, u') - f(Q^k)| &\leq \|Q^k\|_s \\ &= B. \end{aligned} \tag{6.20}$$

Using Cauchy-Schwartz inequality we get that for all (i, u) ,

$$\begin{aligned} |Q^{k+1}(i, u)| &\leq (1 - \gamma(k))\|Q^k\| + \gamma(k) (\max_{(i,u,j)} g(i, u, j) + B) \\ &\leq D, \end{aligned}$$

thus completing the induction argument.

Q.E.D.

Lemma 6.3.1 along with Theorem 6.2.1 directly imply the following theorem

Theorem 6.3.1 *Let $\{Q^k\}$ be generated according to the average cost Q -learning algorithm (6.2). If f satisfies assumptions 6.1.1 and 6.3.1, then $\{Q^k\}$ converges w.p.1. to Q^* , the unique solution to Bellman's equation with $f(Q^*) = \lambda^*$.*

A similar result may not be obtained for the asynchronous case from the analysis above because the span semi-norm property no longer holds. We might have to exploit some other properties of the asynchronous algorithm to be able to prove boundedness of the iterates.

Chapter 7

SSP Average Cost Q -Learning

Analysis

For both the SSP1 and SSP2 algorithms given by (5.6) and (5.7), respectively, we assumed that at each time k , only one state-action pair is updated. This is simply to simplify the notation. The more general case, which allows several components to be updated at the same time, can be reduced to this one via a method of unfolding described in [10]. Therefore, we will carry out the analysis for the simpler form without any loss of generality. In the first section of this chapter, we will analyze the SSP1 algorithm, and in the second section we will analyze the SSP2 algorithm.

7.1 Analysis of SSP1 Algorithm

The convergence of the SSP1 algorithm, under the boundedness condition, follows directly from Lemma 2.7.2. To see this, define the mappings $T : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^{n+m}$ and $T' : \mathbb{R}^{n+m+1} \rightarrow \mathbb{R}^{n+m}$ as follows:

$$\begin{aligned}(TQ)(i, u) &= \sum_{j \neq s} p_{ij}(u) \left[g(i, u, j) + \min_{u'} Q(j, u') \right], \\ (T'(Q, \lambda))(i, u) &= (TQ)(i, u) - \lambda.\end{aligned}$$

Similarly, define the mappings $F(\cdot, \cdot)$, and $F'(\cdot, \cdot, \cdot)$, and $f(\cdot)$ as follows:

$$\begin{aligned} (F(Q, \xi))(i, u) &= \left[g(i, u, \xi_{iu}) + \min_{u'} Q(\xi_{iu}, u') \right] I(\xi_{iu} \neq s), \\ (F'(Q, \xi, \lambda))(i, u) &= (F(Q, \xi))(i, u) - \lambda \\ f(Q) &= \min_u Q(s, u). \end{aligned}$$

The SSP1 algorithm can be re-written as:

$$\begin{aligned} Q^{k+1}(i, u) &= Q^k(i, u) + \gamma(\nu(k, i, u)) \\ &\quad \left[T'(Q^k, \lambda^k) - Q^k(i, u) + w^k \right] I((i, u) = \phi^k), \\ \lambda^{k+1} &= \lambda^k + \beta(k)f(Q^k), \end{aligned} \tag{7.1}$$

where

$$w^k = F'(x^k, \xi^k, \lambda^k) - T'(x^k, \xi^k, \lambda^k).$$

Note that (7.1) is a special case of the general class of algorithms analyzed in Lemma 2.7.2. The second iteration is not stochastic, and f does not depend explicitly on λ . The corresponding limiting ODE's are:

$$\begin{aligned} \dot{Q}(t) &= T'(Q(t), \lambda) - Q(t), \\ &= T(Q(t), \lambda) - Q(t), \end{aligned} \tag{7.2}$$

where λ is treated as a fixed constant. Similarly,

$$\dot{\lambda} = f(q(\lambda)), \tag{7.3}$$

where $q(\lambda) = Q_\lambda$ is the unique asymptotically stable equilibrium point of the ODE (7.2). The following convergence follows directly as long as all the assumptions of the lemma can be verified.

Theorem 7.1.1 *Let $\{(Q^k, \lambda^k)\}$ be generated by the SSP1 algorithm (5.6). If the iterates are bounded, then $\{(Q^k, \lambda^k)\}$ converges to $(Q_{\lambda^*}, \lambda^*)$, the optimal solution of the average cost problem which satisfies $f(Q_{\lambda^*}) = 0$, w.p.1.*

Proof We need simply verify the assumption of Lemma 2.7.2. The Lipschitz continuity conditions on T' and f follow from their definition. The conditions on the stepsizes follow Assumption 5.3.1. The noise conditions on $\{w^k\}$ follows directly from the definition and the Lipschitz continuity of T and F . Now, we are left with only the task of verifying the stability conditions of the ODE's (7.2) and (7.3). For fixed λ , the mapping $T'(Q, \lambda)$ is a contraction mapping with respect to some weighted sup-norm, as it is the dynamic programming mapping for a stochastic shortest path problem where all the policies are proper [6]. Therefore, the equation

$$Q = T'(Q, \lambda)$$

has a unique asymptotically stable equilibrium point, which we denote by Q_λ . Moreover, the analysis of the SSP view of the average cost problem in Chapter 5 states that each component of Q_λ is concave and piece-wise linear in λ , and hence is Lipschitz continuous (the number of linear pieces is finite as the number of policies is finite). The second ODE (7.3) has a unique asymptotically stable equilibrium point, as $f(Q_\lambda) = h_\lambda(s)$ is concave and piece-wise linear. In fact the unique equilibrium point is λ^* (see Vol. 2 of [2]). λ^* is asymptotically stable because the trajectory solution $\lambda(t)$ is basically piecewise exponential.

Q.E.D.

7.2 Analysis of SSP2 Algorithm

Using the notation of the previous section, we re-write the SSP2 algorithm (5.7) as follows:

$$\begin{aligned} Q^{k+1}(i, u) &= Q^k(i, u) + \gamma(\nu(k, i, u)) \\ &\quad \left[T'(Q^k, \lambda^k) - Q^k(i, u) + w^k \right] I((i, u) = \phi^k), \\ \lambda^{k+1} &= \Gamma_{[-K, K]} \left(\lambda^k + \beta(k) f(Q^k) \right), \end{aligned} \tag{7.4}$$

where

$$w^k = F'(x^k, \xi^k, \lambda^k) - T'(x^k, \xi^k, \lambda^k).$$

Lemma 2.7.2 does not apply directly because the second term involves a projection. Before we proceed to show that same limiting behavior holds for (7.4), we first start by establishing the boundedness of the iterates.

Lemma 7.2.1 *The sequence $\{(Q^k, \lambda^k)\}$ is bounded w.p.1.*

Proof $\{\lambda^k\}$ is bounded, as it is constrained in the interval $[-K, K]$. The boundedness of $\{Q^k\}$ follows basically from Theorem 1 in [40]. As T is contractive with respect to some weighted sup-norm, it follows that there is $\alpha \in (0, 1)$ and $D > 0$, such that

$$\max_{(i,u)} |(TQ)(i, u)| \leq \alpha \max_{(i,u)} |Q(i, u)| + D.$$

This implies that for all $\lambda \in [-K, K]$,

$$\max_{(i,u)} |(T'(Q, \lambda))(i, u)| \leq \alpha \max_{(i,u)} |Q(i, u)| + D'$$

where $D' = D + K$. Note that although the map T' depends on λ , the parameters α and D' are independent of λ , and the line of argument in [40] goes through.

Q.E.D.

Lemma 7.2.2 *The sequence $\{(Q^k, \lambda^k)\}$ generated by the SSP2 algorithm (5.6) converges to (Q_λ, λ) for some λ w.p.1.*

Proof Let us re-write the second iteration in (7.4) as follows:

$$\lambda^{k+1} = \lambda^k + \gamma(\nu(k, \phi^k)) \frac{e^k}{\gamma(\nu(k, \phi^k))}, \quad (7.5)$$

where

$$e^k = \Gamma_{[-K, K]}(\lambda^k + \beta(k)f(Q^k)) - \lambda^k.$$

The entire ODE analysis of two-time scale algorithms will apply if we can show that $\frac{e^k}{\gamma(\nu(k, \phi^k))} \rightarrow 0$ as $k \rightarrow \infty$. Consider a sample path over which $\{\|Q^k\|\}$ is bounded. Then $\{|f(Q^k)|\}$ is also bounded, and hence there is B s.t. $|f(Q^k)| < B$ for all k . At each time step, there are three possible outcomes for e^k :

1. If $\lambda^k + \beta(k)f(Q^k) \in [-K, K]$, then

$$e^k = 0.$$

2. If $\lambda^k + \beta(k)f(Q^k) > K$, then

$$\begin{aligned} 0 < e^k &= K - \lambda^k \\ &< \beta(k)f(Q^k). \end{aligned}$$

3. If $\lambda^k + \beta(k)f(Q^k) < -K$, then

$$\begin{aligned} 0 > e^k &= -K - \lambda^k, \\ &> \beta(k)f(Q^k). \end{aligned}$$

Under the three scenarios above, $\{e^k\}$ satisfies

$$|e^k| \leq \beta(k)B \quad \text{for all } k.$$

But

$$\lim_{k \rightarrow \infty} \frac{\beta(k)}{\gamma(\nu(k, \phi^k))} = 0 \quad \text{w.p.1,}$$

because of the assumptions on the step size. Therefore

$$\lim_{k \rightarrow \infty} \frac{e^k}{\gamma(\nu(k, \phi^k))} = 0.$$

The rest of the argument remains exactly the same as in Section 2.7, and implies that the limiting ODE's of the processes $Q_l(\cdot)$ and $\lambda_l(\cdot)$, defined in a similar way as in Section 2.7 are respectively,

$$\begin{aligned} \dot{Q} &= \frac{1}{d} (T(Q) - \lambda e - Q), \quad \text{for fixed } \lambda, \\ \dot{\lambda} &= 0, \end{aligned}$$

where d is the total number of state-action pairs. The asymptotic stability of the unique equilibrium point Q_λ of the first ODE implies the convergence of the sequence to (Q_λ, λ) .

Q.E.D.

The next step is to show that the limit point (Q_λ, λ) is indeed $(Q_{\lambda^*}, \lambda^*)$, as stated in the following theorem:

Theorem 7.2.1 *The sequence $\{(Q^k, \lambda^k)\}$ generated by the SSP2 algorithm (5.6) converges to $(Q_{\lambda^*}, \lambda^*)$, the optimal solution of the average cost problem which satisfies $f(Q_{\lambda^*}) = 0$, w.p.1.*

Proof The ODE limiting argument for the slower process used in Section 2.7 may not be used here because of the projection. We will use a direct argument, which exploits properties of the SSP view of the average cost problem, to characterize the limit point.

Let us define for all k

$$\Delta^k = \lambda^k - \lambda^*,$$

and

$$r^k = f(Q^k) - f(Q_{\lambda^k}).$$

Note that Lemma 7.2.2 implies that

$$\lim_{k \rightarrow \infty} r^k = 0 \tag{7.6}$$

as f is Lipschitz continuous in Q , and Q_λ is continuous in λ . Then the following iteration holds for $\{\Delta^k\}$

$$\begin{aligned} \Delta^{k+1} &= \Gamma_{[-K, K]} \left(\Delta^k + \lambda^* + \beta(k)f(Q^k) \right) - \lambda^*, \\ &= \Gamma_{[-K, K]} \left(\Delta^k + \lambda^* + \beta(k)f(Q_{\lambda^k}) + \beta(k)r^k \right). \end{aligned}$$

However, due to the fact that $f(Q_\lambda)$ is concave and piecewise linear, and that there is finite number of linear pieces (each one corresponds to a policy), it follows that there

exist $0 < \beta_1 \leq \beta_2$ such that

$$-\beta_2(\lambda_1 - \lambda_2) \leq f(Q_{\lambda_1}) - f(Q_{\lambda_2}) \leq -\beta_1(\lambda_1 - \lambda_2), \quad (7.7)$$

for all $\lambda_1, \lambda_2 \in \mathbb{R}$. The above bounds are basically upper and lower bounds on the slope of $f(Q_\lambda)$. Applying equation (7.7) with $\lambda_1 = \lambda^k$ and $\lambda_2 = \lambda^*$, and using the fact that $f(Q_{\lambda^*}) = 0$, we get

$$-\beta_2 \Delta^k \leq f(Q_{\lambda^k}) \leq -\beta_1 \Delta^k.$$

Using the above and the fact that $\Gamma_{[-K, K]}$ is non-decreasing, we get:

$$\begin{aligned} \Gamma_{[-K, K]} \left((1 - \beta_2 \beta(k)) \Delta^k + \beta(k) r^k + \lambda^* \right) - \lambda^* &\leq \Delta^{k+1} \\ &\leq \Gamma_{[-K, K]} \left((1 - \beta_1 \beta(k)) \Delta^k + \beta(k) r^k + \lambda^* \right) - \lambda^*. \end{aligned}$$

As $\beta(k) \rightarrow 0$ as $k \rightarrow \infty$, it follows that for every $\epsilon > 0$, there is a large enough N such that for $k \geq N$,

$$\begin{aligned} (1 - \beta_2 \beta(k)) \Delta^k + \beta(k) r^k + \lambda^* - \epsilon &\leq \Gamma_{[-K, K]} \left((1 - \beta_1 \beta(k)) \Delta^k + \beta(k) r^k + \lambda^* \right) \\ &\leq (1 - \beta_1 \beta(k)) \Delta^k + \beta(k) r^k + \lambda^* + \epsilon. \end{aligned}$$

Therefore,

$$(1 - \beta_2 \beta(k)) \Delta^k + \beta(k) r^k - \epsilon \leq \Delta^{k+1} \leq (1 - \beta_1 \beta(k)) \Delta^k + \beta(k) r^k + \epsilon.$$

Given that $r^k \rightarrow 0$, $\beta(k) \rightarrow 0$, as $k \rightarrow \infty$, and that $\sum_k \beta(k) = \infty$, it follows by expanding out the inequalities above and using identities from Polyak's book [30] (see (A.4)), we get by taking the limits on both sides that $\{\Delta^k\}$ converges to $[-\epsilon, \epsilon]$. The above is true for any $\epsilon > 0$, therefore $\Delta^k \rightarrow 0$. Hence,

$$\lim_{k \rightarrow \infty} \lambda^k = \lambda^*,$$

and

$$\lim_{k \rightarrow \infty} Q^k = Q_{\lambda^*}.$$

Q.E.D.

As a last remark, we will note that equivalent convergence results apply for synchronous versions of the SSP1 and SSP2 algorithms.

Chapter 8

Conclusion

In summary, this thesis has important contributions, which can be classified in three major categories:

- Theoretical framework for establishing boundedness of stochastic approximation iterates for a large and important class of problems.
- Applications to the analysis of Q -learning algorithms.
- Unified presentation of ODE limiting arguments for synchronous and asynchronous stochastic approximation algorithms.

In the theoretical framework, we addressed a difficult issue in the field of stochastic approximation, which is the issue of boundedness. The general convergence results of stochastic approximation algorithms which do not impose the boundedness assumption typically require conditions such as the existence of a smooth Lyapunov function [30], covering basically the important class of stochastic gradient algorithms. However, this does not cover many important algorithms and applications, in particular Q -learning algorithms for dynamic programming problems. Motivated by the Q -learning applications for discounted cost problems and stochastic shortest path problems, Tsitsiklis [40] proved boundedness for synchronous and asynchronous stochastic approximation algorithms which involve contraction mappings with respect to the sup-norm. A major contribution of this thesis is the development of a general framework for

proving boundedness of synchronous and asynchronous stochastic approximation algorithms involving mappings which are non-expansive with respect to the sup-norm. The method of proof involves two major types of lemmas: the boundedness lemmas and the lemmas establishing the limiting ODE of the scaled iterations. The boundedness lemmas were inspired by a similar idea in Jaakola, Jordan, and Singh [17]. The analysis of the scaled iterations is new and involves sophisticated mathematical ideas such as the converse Lyapunov theorem and convergence in the Skorohod topology. The boundedness proofs, along with the stability of ODE's whose mappings are non-expansive and have unique fixed points, result in the general convergence theorems 3.4.1 and 3.4.2. The following table summarizes how these results fit with already existing results. It only considers w.p.1 convergence-type of results, and is by no means a summary of all existing stochastic approximation results or applications.

Algorithm Class	Synchronous results	Asynchronous results
Stochastic Gradient	Polyak and Tsytkin	Tsitsiklis
contractive w.r.t sup-norm	Tsitsiklis	Tsitsiklis
Continuous Monotone Unique fixed pt. DP-type properties	Tsitsiklis	Tsitsiklis
Non-expansive w.r.t some norm Unique fixed point	New	x
Non-expansive w.r.t sup-norm Unique fixed point	New	New

Figure 8-1: Summary of Stochastic Approximation Results

In terms of applications, the entire work in this thesis was motivated by the search for Q -learning algorithms for average cost problems which can be shown to converge. In fact, we propose two types of Q -learning algorithms: The first one (AC Q -learning algorithm) is based on the relative value iteration, and the second type (SSP1 and SSP2 Q -learning algorithms) is based on the stochastic shortest path view of the average cost problem. The general theoretical framework does not apply directly to these algorithms, but it plays a major role in the analysis of the AC algorithm as we were able to easily extend it to cover synchronous variant of the algorithm by exploiting the span semi-norm non-expansiveness and other properties of the average cost mapping. The proof of convergence of the AC algorithm involves characterizing a general class of relative value iteration algorithms, and analyzing the corresponding ODE's. This is a new contribution, which builds on the ODE analysis of non-expansive mappings [13] and exploits properties of Bellman's equation for the average cost problem. Based on the theoretical framework developed in this thesis, we proved the following:

1. Convergence of Q -learning algorithms for discounted cost and stochastic shortest path problems. This provides an ODE analysis for existing proofs [17] and [40], with the exception that it allows for improper policies with infinite costs for stochastic shortest path problems.
2. Convergence of the synchronous variant of the AC algorithm.
3. Convergence of the asynchronous variant of the AC algorithm, under the boundedness assumption.
4. Stability of a general class of ODE's corresponding to different variants of the relative value iteration algorithm.

The analysis of the SSP algorithms is based on the two-time scale stochastic approximation results of Borkar [11]. Although Borkar's paper only presents results for synchronous algorithms, we show how these can be easily extended to asynchronous algorithms. The analysis of the SSP1 algorithm follows directly from this extension.

The SSP2 algorithm however involves a projection, and cannot be handled directly by the results of [11]. We carried out a separate ODE limiting argument for the SSP2 algorithm to prove convergence. The following summarizes the convergence results of the SSP algorithms:

1. Convergence of the SSP1 algorithm, under the boundedness assumption.
2. Convergence of the SSP2 algorithm.

The applications results are summarized in the following table:

Result	Contribution	Technique
Q -learning for discounted cost and SSP (proper policies)	Alternate proof	New convergence theorems
Q -learning for SSP	New	New convergence theorems
Stability of Average Cost ODE	New	Extension of Borkar's stability of non-expansive maps
Synchronous Average Cost Q -learning	New	New framework
Asynchronous Average Cost Q -learning	New	Standard ODE argument with boundedness
Stability of SSP ODE's	New	Known analysis of ODE's with contraction or linear maps
SSP1 Q -learning	New	Two-time scale with boundedness
SSP2 Q -learning	New	Slight adaptation of two-time scale, boundedness follows from contraction property

Figure 8-2: Summary of Q -learning Analysis Results

An important contribution of the thesis is the uniform presentation that has been used throughout. The results presented rely heavily on establishing limiting deterministic ODE's for stochastic discrete-time algorithms. There are basically two

approaches for achieving this: the Kushner and Clark approach and the Borkar approach. Kushner and Clark use left-shifted interpolated processes and basically shift the asymptotic behavior of the discrete iteration to the origin. The characterization of the limiting process is based on the Arzela-Ascoli theorem, and the convergence of the discrete iteration to the unique asymptotically stable limit point of the ODE again exploits these left shifts. Borkar bases his analysis on comparing different interpolated processes and ODE trajectories over a finite time interval out in the future, and showing that the difference between the two goes to 0 using Gronwall's lemma. He shows the convergence of the discrete iteration to the unique asymptotically stable equilibrium point using a perturbation-type of argument. In Chapter 2, we presented alternate limiting ODE arguments for both Borkar's results on asynchronous and two-time scale stochastic approximation [10] in a manner consistent with the Kushner and Clark approach. The two approaches yield the same results, but choosing a uniform approach throughout the thesis provides a simple way of visualizing the results and a uniform framework for the proofs.

Finally, there are different directions and possible extensions of the work in this thesis. From a theoretical point of view, there are some interesting open problems:

1. Convergence proof of the AC asynchronous Q -learning without the boundedness assumption.
2. Convergence proof of the SSP1 algorithm without the boundedness assumption.
3. Generalization of the theoretical framework to include a larger class of problems.
4. Generalization of the general theoretical framework to include stochastic algorithm with two-time scales.
5. Generalization of the results of the thesis to asynchronous algorithms with delays.

From a practical point of view, this thesis does not address the practicality of Q -learning algorithms. In general, there are two major drawbacks for Q -learning algorithms:

- The look-up table representation is prohibitive for large problems.
- The convergence rate is slow due to the high variance of each iteration.

These issues have been addressed extensively in the context of discounted cost and stochastic shortest path problems, and several ideas have been explored such as the use of state aggregation or function approximation [2]. All these ideas can be used to carry out policy evaluation and then some sort of policy iterations using two different approaches:

1. For problems with a known recurrent state under all policies, the average cost of a policy is simply the ratio of two stochastic shortest path problems (see Chapter 5). Therefore, all the computational simulation-based methods of SSP problems such as $TD(\lambda)$, Monte-Carlo simulation with function approximation can be applied. Then policy iteration can be performed, and it will be interesting to explore ideas such as optimistic policy iteration in this context.
2. For general average cost problems, monte-carlo simulation could be used to carry out policy evaluation or a $TD(\lambda)$, based on the relative value iteration algorithm could be used in conjunction with function approximation.

Q -learning could be used as a basis for different computational methods, as issues of function approximation, adequate exploration of the state space, and step size selection to optimize convergence rates are serious implementation issues and should be addressed per specific application. Hence, it is important to try out these different ideas for some important large-scale problems. This would be an interesting direction in the study of the applicability of Q -learning for average cost problems.

Appendix A

Mathematical Background

The following is a list of identities and formulas used in the thesis. The arguments provided here were omitted in the major sections of the thesis to keep the flow uncluttered by the details.

Variation of Constants Formula It is easy to verify that if $x(t)$ is a unique solution of the ODE

$$\dot{x} = h(x) - x, \quad \text{with initial condition} \quad x(0) = x_0,$$

then

$$x(t) = x_0 e^{-t} + \int_0^t e^{-(t-\tau)} h(x(\tau)) d\tau. \quad (\text{A.1})$$

The identity (A.1) can be verified by taking the derivative of both sides with respect to t .

Gronwall's Lemma The following version of Gronwall's Lemma is from Vidyasagar [44]. Let $x : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ be a non-negative function defined for $t \geq 0$. Let b and c be two non-negative constants. If

$$z(t) \leq b + \int_0^t cz(\tau) d\tau, \quad \text{for} \quad t \geq 0,$$

then

$$z(t) \leq be^{ct}, \quad \text{for } t \geq 0. \quad (\text{A.2})$$

Convergence Identity Let $z(t)$ be a continuous function from $\mathbb{R}^+$ to $\mathbb{R}^n$, and let c be a positive constant. If

$$\lim_{t \rightarrow \infty} z(t) = z^*,$$

then

$$\lim_{t \rightarrow \infty} \int_0^t e^{-c(t-\tau)} z(\tau) d\tau = \frac{z^*}{c}.$$

The identity above follows from the following argument: Given any $\epsilon > 0$, there is T_ϵ such that

$$\|z(t) - z^*\| < \frac{\epsilon}{2} \quad \text{for } t \geq T_\epsilon. \quad (\text{A.3})$$

But

$$\begin{aligned} \left\| \int_0^t e^{-c(t-\tau)} z(\tau) d\tau - \frac{z^*}{c} \right\| &= \left\| \int_0^t e^{-c(t-\tau)} (z(\tau) - z^*) d\tau \right\| \\ &\leq \int_0^{T_\epsilon} e^{-c(t-\tau)} \sup_{\tau'} \|z(\tau') - z^*\| d\tau + \int_{T_\epsilon}^t e^{-c(t-\tau)} \frac{\epsilon}{2} d\tau \\ &= \frac{1}{c} (e^{-c(t-T_\epsilon)} - e^{-ct}) \sup_{\tau'} \|z(\tau') - z^*\| + \frac{1}{c} (1 - e^{t-T_\epsilon}) \frac{\epsilon}{2}. \end{aligned}$$

The first term of the last equation can be made smaller than $\frac{\epsilon}{2}$ for $t \geq T'$ for some T' . Then $\left\| \int_0^t e^{-c(t-\tau)} z(\tau) d\tau - \frac{z^*}{c} \right\| < \epsilon$ for $t \geq T$, where $T = \max(T', T_\epsilon)$.

Identities The following identities are at the heart of the convergence of algorithms with diminishing step sizes. These are all stated and proved in Polyak [30]. We will re-state them here in a form that is useful for the proofs in the thesis, in particular the last part of Chapter 7. Let $\{\beta(k)\}$ be a non-negative sequence such that

$$\begin{aligned} \lim_{k \rightarrow \infty} \beta(k) &= 0, \\ \sum_{k=0}^{\infty} \beta(k) &= \infty, \end{aligned}$$

and $\{r^k\}$ be a sequence such that $\lim_{k \rightarrow \infty} r^k = 0$. Then the following identities hold:

$$\begin{aligned} \lim_{k \rightarrow \infty} \Pi_{i=0}^k (1 - \beta(i)) &= 0, \\ \lim_{k \rightarrow \infty} \sum_{i=0}^k \left(\Pi_{j=i+1}^k (1 - \beta(j)) \right) \beta(i) r^i &= 0. \end{aligned} \tag{A.4}$$

Appendix B

Asynchronous Relative Value Iteration

B.1 Introduction

Computing the optimal average cost for Markov Decision Problems (MDP's) is a very important problem for several applications. The value iteration algorithm converges under generally reasonable assumptions [2]. We're interested in understanding the behavior of the value iteration algorithm when implemented in an asynchronous fashion. It is well known that in general it does not converge asynchronously, as shown by the counterexample in [3]. We hope to show that if we introduce some relaxation to the asynchronous algorithms, then we might be able to prove the convergence.

B.2 2-state Markov Chain

B.2.1 Value Iteration: Gauss Seidel Version

We'll first consider single policy Markov Decision Problems, as these correspond to a single Markov chain and the value iteration algorithm can be analyzed as a linear dynamic system. As in [3], consider at first a 2-state problem. Hence the state space is given by $S = \{0, 1\}$. Under the policy at hand, a cost g_0 is incurred at state 0,

and a transition to state 1 occurs with probability p_0 . Similarly, a cost g_1 is incurred at state 1, and a transition to state 1 occurs with probability p_1 . To insure that the Markov chain is unichain and aperiodic, p_0 and p_1 are required to be strictly between 0 and 1. The transition probability matrix is given by:

$$P = \begin{pmatrix} 1 - p_0 & p_0 \\ 1 - p_1 & p_1 \end{pmatrix} \quad (\text{B.1})$$

The value iteration algorithm for computing the average cost of the policy above is given by:

$$\begin{aligned} J^{k+1}(0) &= g_0 + p_0 (J^k(1) - J^k(0)) \\ J^{k+1}(1) &= g_1 + p_1 (J^k(1) - J^k(0)) \end{aligned} \quad (\text{B.2})$$

The algorithm above converges for $p_0, p_1 \in (0, 1)$, and the average cost is given by $\lim_{k \rightarrow \infty} J^k(0)$. The iteration above can be rewritten as:

$$J^{k+1} = g + \tilde{P} J^k \quad (\text{B.3})$$

where

$$J^k = \begin{pmatrix} J^k(0) \\ J^k(1) \end{pmatrix},$$

$$g = \begin{pmatrix} g_0 \\ g_1 \end{pmatrix},$$

and

$$\tilde{P} = P - \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

The eigenvalues of $\tilde{P}$ are 0 and $(p_1 - p_0)$, and hence the iteration above converges. The iteration above will be referred to as the Jacobi iteration.

The Gauss Seidel version of the iteration in (B.3) is given by:

$$J^{k+1} = g_{GS} + \tilde{P}_{GS} J^k$$

where

$$g_{GS} = \begin{pmatrix} g_0 \\ g_0 - p_1 g_0 \end{pmatrix} \quad (\text{B.4})$$

$$\tilde{P}_{GS} = \begin{pmatrix} -p_0 & p_0 \\ p_1 p_0 & p_1(1 - p_0) \end{pmatrix} \quad (\text{B.5})$$

The eigenvalues of $\tilde{P}_{GS}$ are given by:

$$\lambda_1 = \frac{-p_0 + p_1(1 - p_0) - \sqrt{(-p_0 + p_1(1 - p_0))^2 + 4p_0 p_1}}{2} \quad (\text{B.6})$$

$$\lambda_2 = \frac{-p_0 + p_1(1 - p_0) + \sqrt{(-p_0 + p_1(1 - p_0))^2 + 4p_0 p_1}}{2} \quad (\text{B.7})$$

It can be shown that if $p_0, p_1 \in (0, 1)$, then the eigenvalues of $\tilde{P}_{GS}$ are real and strictly less than 1. The algorithm will diverge if $\tilde{P}_{GS}$ has an eigenvalue less or equal than -1 . Simple algebra shows that any choice of p_0, p_1 that satisfies $2p_0 p_1 \geq (p_1 - p_0)$ yields a diverging iteration. An example of such a choice is $p_0 = 0.8, p_1 = 0.5$. This provides a counterexample that shows that for average cost problems, asynchronous value iteration (in particular a Gauss Seidel version) might not converge.

B.2.2 Relaxed Gauss Seidel: GS1

Given an iterative algorithm of the form

$$x^{k+1} = T(x^k), \quad (\text{B.8})$$

we define a relaxed version of the iteration as

$$x^{k+1} = (1 - \gamma)x^k + \gamma T(x^k) \quad \text{where } \gamma \in (0, 1]. \quad (\text{B.9})$$

In the case of linear iterations, the eigenvalues of the relaxed version are simply convex combinations of the eigenvalues of the original iteration and 1. If all the eigenvalues of the original iteration have real parts less than 1, then for a small enough choice of γ , the relaxed version of the iteration will have eigenvalues inside the unit disk and hence will converge.

We'll refer to the relaxed version of the Gauss Seidel iteration defined above as the GS1 algorithm. For the 2-state Markov chain, it can be easily shown that all the eigenvalues of the Gauss Seidel matrix $\tilde{P}_{GS}$ are less than 1. Hence, GS1 converges for α small enough. In particular, if the algorithm is implemented with a diminishing stepsize, then it will converge. Hence, although the Gauss Seidel iteration might diverge (e.g. $p_0 = 0.8, p_1 = 0.5$), the GS1 algorithm, either with a fixed small enough γ or with a diminishing stepsize converges.

B.2.3 Gauss Seidel of Relaxed Jacobi: GS2

Another way of generating a relaxed Gauss Seidel is to implement a Gauss Seidel version of a relaxed Jacobi iteration. For a 2-state Markov chain, the algorithm is as follows:

$$J^{k+1}(0) = (1 - \gamma(k))J^k(0) + \gamma(k) \left(g_0 + p_0 (J^k(1) - J^k(0)) \right) \quad (\text{B.10})$$

$$J^{k+1}(1) = (1 - \gamma(k))J^k(1) + \gamma(k) \left(g_1 + p_1 (J^k(1) - J^{k+1}(0)) \right) \quad (\text{B.11})$$

Rearranging the terms above, we get

$$J^{k+1} = (1 - \gamma)J^k + \gamma(g + \tilde{P}J^k) + o(\gamma^2) \quad (\text{B.12})$$

If the algorithm above is implemented with a diminishing stepsize, then as $\gamma(k) \rightarrow 0$, the algorithm approaches asymptotically the Jacobi iteration and hence converges.

B.3 3-state Markov Chain

For the 2-state example above, we saw that the Jacobi iteration converges. There are examples where the Gauss Seidel iteration does not converge. However, both GS1 and GS2 always converge.

For a 3-state Markov chain, the following gives an example where the Jacobi iteration converges, but both the Gauss Seidel and GS1 do not converge. The GS2 algorithm converges for the same reason as in the 2-state Markov chain example.

Let P be the transition probability matrix of the policy whose average cost is to be evaluated. We will assume that the corresponding Markov chain is unichain and aperiodic. Under this assumption, the value iteration algorithm given by (B.3) converges.

Consider the following example:

$$P = \begin{pmatrix} 0.005 & 0.005 & 0.99 \\ 0.05 & 0.2 & 0.75 \\ 0.05 & 0.2 & 0.75 \end{pmatrix} \quad (\text{B.13})$$

It can be easily verified that P corresponds to a unichain, aperiodic chain and hence value iteration converges.

The corresponding Gauss Seidel matrix is:

$$\tilde{P}_{GS} = \begin{pmatrix} -0.995 & 0.005 & 0.99 \\ 0.9452 & 0.1953 & -0.1905 \\ 1.1343 & 0.0343 & 0.7119 \end{pmatrix} \quad (\text{B.14})$$

The eigenvalues of $\tilde{P}_{GS}$ are $-1.50, 1.22$, and 0.18 . As $\tilde{P}_{GS}$ has an eigenvalue greater than 1, the relaxed version GS1 does not converge.

However, for the relaxed version GS2, the same argument used for the 2-state example applies. The effect of γ is 2nd order, and with a diminishing stepsize, the GS2 algorithm asymptotically converges to the Jacobi iteration and hence converges.

B.4 Conclusion

The conclusion is that although there are counterexamples to the convergence of asynchronous relative value iteration algorithms, the analysis of the 2-state and 3-state Markov chains above suggests that it might converge if a constant small enough step size is introduced. This fact has not been proved here, and is still an open problem. However, it suggests that there is no reason to stay away from the relative value iteration algorithm as a basis for asynchronous stochastic approximation algorithms just because it does not converge asynchronously.

Bibliography

- [1] A. Benveniste, M. Metivier, and P. Priouret, Adaptive Algorithms and Stochastic Approximations, Springer Verlag, N.Y., 1990.
- [2] D. P. Bertsekas, Dynamic Programming and Optimal Control, Belmont: Athena Scientific, 1995.
- [3] D. P. Bertsekas, "Distributed Dynamic Programming," *IEEE Transactions on Automatic Control*, vol. AC-27, No. 3, 1982.
- [4] D. P. Bertsekas, "A New Value Iteration Method for the Average Cost Dynamic Programming Problem," *SIAM Journal on Control and Optimization*, 1996.
- [5] D. P. Bertsekas and J. N. Tsitsiklis, Parallel and Distributed Computation: Numerical Methods. New Jersey: Prentice Hall, 1989.
- [6] D. P. Bertsekas and J. N. Tsitsiklis, "An Analysis of Stochastic Shortest Path Problems," *Mathematics of Operations Research*, Vol. 16, pages 580-595, 1991.
- [7] D. P. Bertsekas and J. N. Tsitsiklis, Neuro-Dynamic Programming, Belmont, MA: Athena Scientific, 1995
- [8] P. Billingsley, Convergence of Probability Measures. New York: John Wiley & Sons, 1968.
- [9] V. S. Borkar, "Distributed Computation of Fixed Points of ∞ -expansive maps," *Proc. of the Indian Academy of Sciences (Mathematical Sciences)*, 106 (3), Aug 1996, 289-300.

- [10] V. S. Borkar, "Asynchronous Stochastic Approximations," March 1995, to appear in SIAM Journal of Control and Optim.
- [11] V. S. Borkar, "Stochastic Approximation with Two-time Scales", System and Control Letters 29 (1997), 291-294.
- [12] V. S. Borkar and V. R. Konda, "The actor-critic algorithm as multi-time scale stochastic approximation," Sadhana (Proc. Indian Academy of Sciences - Eng. Sciences) 27(4), 1-19, Aug 1997.
- [13] V. S. Borkar and Soumyanath: "A New Analog Parallel Scheme for Fixed point Computation Part I:Theory", Submitted to IEEE Transactions on Control Systems.
- [14] R. H. Crites and A.G. Barto, "Improving Elevator Performance using Reinforcement Learning," Advances in Neural Network Processing Systems, Vol.8, pages 1017-1023, Cambridge, MA: MIT Press, 1996.
- [15] W. Hahn, Stability of Motion, Trans. by A. P. Baartz: Springer Verlag New York, 1967.
- [16] Morris W. Hirsh, "Convergent Activation Dynamics in Continuous Time Networks," Neural Networks, Vol. 2, pages 331-349, 1989.
- [17] T. Jaakola, M. I. Jordan, and S. P. Singh, "On the Convergence of Stochastic Iterative Dynamic Programming Algorithms", *Neural Computation*, 6,1185-1201, 1994.
- [18] A. Jalali and M. Ferguson. "Adaptive Control of Markov Chains with Local Updates," Systems and Control Letters, Vol. 14, pp.209-218, 1990.
- [19] J. Kiefer and J. Wolfowitz, "Stochastic Estimation of the Maximum of a Regression Function," Ann. Math. Statist., Vol. 23, pages 462-466, 1952.

- [20] J. Kurzweil, "On the inversion of Lyapunov's second theorem on stability of motion," Czechoslovak Math. J., Vol. 81, 1956; American Math. Soc. Transl., Vol. 24, 1963.
- [21] H. J. Kushner and D. S. Clark, Stochastic Approximation Methods for Constrained and Unconstrained Systems. New York: Springer Verlag, 1978.
- [22] H. J. Kushner and G. Yin, "Stochastic Approximation Algorithms for Parallel and Distributed Processing," Stochastics, Vol. 22, pages 219-250, 1987.
- [23] H. J. Kushner and G. Yin, "Asymptotic Properties of Distributed and communicating Stochastic Approximation Algorithms," SIAM J. Control and Optimization, Vol. 25, pages 1266-1290, 1987.
- [24] H. J. Kushner and G. Yin, Stochastic Approximation Algorithms and Applications, Springer, New York, 1997.
- [25] S. Lefschetz, Differential Equations: Geometry Theory, Interscience, N.Y. 1957.
- [26] Ljung, L. "Analysis of Recursive Stochastic Algorithms", I.E.E.E. Transactions on Automatic Control, Vol. 22, pages 551-575, 1977.
- [27] S. Mahadevan, Average reward reinforcement learning: Foundations, algorithms, and empirical results,' Machine Learning, Vol. 22, pages 1-38, 1996.
- [28] M. Metivier and P. Priouret, "Applications of a Kushner and Clark Lemma to General Classes of Stochastic Algorithms," IEEE Transactions on Information Theory, Vol. IT-30, No. 2, March 1984.
- [29] J. Neveu, Discrete Parameter Martingales, North Holland, Amsterdam, 1975.
- [30] B. T. Polyak, Introduction to Optimization, Optimization Software, New York, 1987.
- [31] B. T. Polyak and Y.Z. Tsytkin, "Pseudogradient Adaptation and Training Algorithms", Automation and Remote Control, Vol. 12, pp. 83-94, 1973.

- [32] M. L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming. New York: John Wiley & Sons, 1994.
- [33] H. Robbins and S. Monro, "A Stochastic Approximation Method," Ann. Math. Statistics, Vol. 22, pp. 400-407, 1951.
- [34] A. Schwartz, "A reinforcement learning method for maximizing undiscounted rewards," Proceedings of the Tenth International Conference on Machine Learning, pp. 298-305. Morgan Kaufmann, 1993.
- [35] G. R. Sell, Topological Dynamics and Ordinary Differential Equations. London: Van Nostrand Reinhold Company, 1971.
- [36] S. P. Singh, Reinforcement learning algorithms for average payoff Markovian Decision Processes, Proceedings of the Twelvth National Conference on Artificial Intelligence, pp. 202-207, 1994.
- [37] R. S. Sutton, "Learning to Predict by the Methods of Temporal Differences," Machine Learning, Vol.3, p. 9-44, 1988.
- [38] Tijms, Stochastic Modeling and Analysis: A Computational Approach. New York, John Wiley and Sons, 1986.
- [39] P. Tseng, D. P. Bertsekas, and J. N. Tsitsiklis, "Partially Asynchronous Parallel Algorithms for Network Flow and Other Problems," Siam J. Control and Optimization, Vol 28, pp. 678-710, May 1990.
- [40] J. N. Tsitsiklis, "Asynchronous Stochastic Approximation and Q -learning," LIDS-P2172, 1993.
- [41] J. N. Tsitsiklis, "Problems in Decentralized Decision Making and Computation," Ph.D. Thesis. Massachusetts Institute of Technology, Cambridge, MA, 1984.
- [42] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed Asynchronous Deterministic and Stochastic Gradient Optimization Algorithms," IEEE Transactions on Automatic Control, Vol. AC-31, P. 803-812, 1986.

- [43] P. Varaiya, "Optimal and Suboptimal Stationary Controls for Markov Chains," IEEE Transactions on Automatic Control, Vol. AC-23, No. 3, Jun 1978.
- [44] M. Vidyasagar, Nonlinear Systems Analysis, Prentice-Hall, New Jersey, 1993.
- [45] C. J. Watkins, "Learning from Delayed Rewards," Ph.D. Thesis. Cambridge University, England, 1989.
- [46] , C. J. Watkins & P. Dayan, "Q-Learning," Machine Learning, Vol. 8, p.279-292, 1992.
- [47] F. W. Wilson, "Smoothing Derivatives of Functions and Applications," Transaction of American Mathematical Society, P. 413-428, May 1969.
- [48] T. Yoshizawa, Stability Theory by Lyapunov second method. Tokyo, Mathematical Society of Japan, 1966.