Actor-Critic Algorithms

by

Vijaymohan Konda

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

June 2002

Author .....
Department of Electrical Engineering and Computer Science
March 15, 2002

Certified by .....
John N. Tsitsiklis
Professor
Thesis Supervisor

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

Actor-Critic Algorithms

by

Vijaymohan Konda

Submitted to the Department of Electrical Engineering and Computer Science
on March 15, 2002, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Many complex decision making problems like scheduling in manufacturing systems, portfolio management in finance, admission control in communication networks etc., with clear and precise objectives, can be formulated as stochastic dynamic programming problems in which the objective of decision making is to maximize a single "overall" reward. In these formulations, finding an optimal decision policy involves computing a certain "value function" which assigns to each state the optimal reward one would obtain if the system was started from that state. This function then naturally prescribes the optimal policy, which is to take decisions that drive the system to states with maximum value.

For many practical problems, the computation of the exact value function is intractable, analytically and numerically, due to the enormous size of the state space. Therefore one has to resort to one of the following approximation methods to find a good sub-optimal policy : (1) Approximate the value function. (2) Restrict the search for a good policy to a smaller family of policies.

In this thesis, we propose and study actor-critic algorithms which combine the above two approaches with simulation to find the best policy among a parameterized class of policies. Actor-critic algorithms have two learning units: an actor and a critic. An actor is a decision maker with a tunable parameter. A critic is a function approximator. The critic tries to approximate the value function of the policy used by the actor, and the actor in turn tries to improve its policy based on the current approximation provided by the critic. Furthermore, the critic evolves on a faster time-scale than the actor.

We propose several variants of actor-critic algorithms. In all the variants, the critic uses Temporal Difference (TD) learning with linear function approximation. Some of the variants are inspired by a new geometric interpretation of the formula for the gradient of the overall reward with respect to the actor parameters. This interpretation suggests a natural set of basis functions for the critic, determined by the family of policies parameterized by the actor's parameters. We concentrate on the average expected reward criterion but we also show how the algorithms can be modified for other objective criteria. We prove convergence of the algorithms for problems with general (finite, countable, or continuous) state and decision spaces.

To compute the rate of convergence (ROC) of our algorithms, we develop a general theory of the ROC of two-time-scale algorithms and we apply it to study our algorithms. In the process, we study the ROC of TD learning and compare it with related methods such as Least Squares TD (LSTD). We study the effect of the basis

functions used for linear function approximation on the ROC of TD. We also show that the ROC of actor-critic algorithms does not depend on the actual basis functions used in the critic but depends only on the subspace spanned by them and study this dependence.

Finally, we compare the performance of our algorithms with other algorithms that optimize over a parameterized family of policies. We show that when only the "natural" basis functions are used for the critic, the rate of convergence of the actor-critic algorithms is the same as that of certain stochastic gradient descent algorithms. However, with appropriate additional basis functions for the critic, we show that our algorithms outperform the existing ones in terms of ROC.

Thesis Supervisor: John N. Tsitsiklis
Title: Professor

Acknowledgments

I gratefully acknowledge the constant support and encouragement of my thesis advisor Professor John Tsitsiklis. His extraordinary patience, and his approach to research, writing, teaching, and mentoring has been a great source of inspiration. My association with him for the past three years has been one great learning experience.

I am also indebted to Professor Sanjoy Mitter for his support and encouragement during the most difficult part of my life. I would like to thank Professor Mitter for his interest in my studies and career, and his advice. Professor Mitter and Professor Bertsekas served as thesis readers. I would like to thank Professor Bertsekas for his encouragement and advice.

My master's thesis advisor Professor Vivek Borkar introduced me to the field of Neuro-Dynamic programming/Reinforcement learning. My understanding of probability theory and stochastic control was very much influenced by our interaction.

My numerous discussions with Anand Ganti, Anant Sahai, Sekhar Tatikonda, Maurice Chu, Constantine Caramanis and Ramesh Johari have improved my understanding of various research areas.

Finally, I thank my parents for their unconditional support and encouragement. They have made great sacrifices to promote my career. I would also like to thank my wife Manjula, without her support and understanding this thesis would not have been completed. The innocent smile of our newborn son Vinay helped me withstand the stress of thesis writing.

This research was partially supported by the National Science Foundation under contract ECS-9873451 and by the AFOSR under contract F49620-99-1-0320.

Contents

Chapter 1

Introduction

Actor-Critic algorithms (Barto *et al.*, 1983) originated in the Artificial Intelligence (AI) literature in the context of Reinforcement Learning (RL). In an RL task, an agent learns to make "correct" decisions in an uncertain environment. Unlike supervised learning in which the agent is given examples of "correct" behavior, an RL agent has to work with much less feedback as it is given only rewards (or penalties) for its decisions. Moreover, the decision making consists of several stages and the reward may be obtained either at the end or at each stage of the decision making process. The main difficulty with such problems is that of "temporal credit assignment" which is to rank the decisions based on the overall reward when only the immediate rewards at each stage are available.

Some of the methods of RL were in part inspired by studies of animal learning and hence were heuristic and ad hoc. Nevertheless, some of them have been systematized by establishing their connections with dynamic programing and stochastic approximation. The original purpose of RL research was two-fold. On one hand, the goal was to understand the learning behavior of animals and on the other, to apply this understanding to solve large and complex decision making (or control) problems. In this thesis, the focus is only on control or decision making in large systems and on the development of learning methods with good mathematical foundations.

The outline of this chapter is as follows. In the next section, various issues that arise in management and control of complex systems and a broad overview of the approaches to tackle them are discussed. Then, a survey of the relevant literature is presented. The third section of this chapter introduces actor-critic algorithms and discusses various open problems. The final two sections of this chapter describe the contributions and the outline of the thesis.

1.1 Control of Complex Systems

Although "complex systems" are difficult to characterize precisely, these systems are typically uncertain, distributed, and asynchronous. Scheduling in manufacturing systems, admission control in communication networks, admission and power control in wireless networks, inventory control in supply chains, etc., are some examples of

control problems related to complex systems. The following are some salient features of such control problems :

1. The total information about the state of the system (which is seldom available) relevant to the decision making process, is high-dimensional. In control theoretic terms, this translates to a large number of states (in the discrete case) or a large dimension of the state space (in the continuous case). Therefore a realistic control policy can only use a small number (compared to the dimension of the state space) of features extracted from the state information. Furthermore, the choice of these features (also called *feature extraction*) is a part of the control design, unlike in the case of traditional partially observed control problems.

2. Models for such large and complex systems are difficult and costly to construct. Even when such models are available, it is not easy to solve corresponding the control problems as they are intractable. Therefore, one has to resort to simulation to understand the dynamics and design a control policy.

3. Such problems cannot be solved by general purpose methods due to their inherent computational complexity and the size of underlying systems. Instead one needs methods tuned to specific systems. This in turn requires engineering insight, intuition, experimentation and analysis.

The types of problems we have mentioned above can often be formulated as optimal control problems in which the objective is to maximize a single "overall" reward over all policies. Many of the simplified versions of such problems have been very well studied under the umbrella of Dynamic Programming (DP) (Bertsekas, 1995b), and many impressive results on the existence and structure of the optimal policies have been obtained.

The key concept in DP is that of a value function. The value function associated with a particular policy assigns to each state the overall reward one would obtain if the system was started from that state and the given policy was followed to make the decisions. Finding an optimal decision policy using DP involves computing the optimal value function (or simply the value function) which satisfies a nonlinear equation called the Bellman or the DP equation. This function then naturally prescribes an optimal policy, which is to take decisions that drive the system to states with maximum value. However, the classical DP computational tools are often inadequate for the following reason.

The amount of computational resources (in particular, space) required for classical dynamic programming methods is at least proportional to (if not polynomial or exponential in) the size of the state space. The number of states (or the dimension of the state space in continuous case) in many practical problems is so high that it prohibits the use of the classical methods. This has been a major drawback of computational dynamic programming and has been named the "Curse of Dimensionality" by Bellman.

Due to the inadequacy of DP tools, the approach to control of complex systems has mainly been heuristic and ad hoc. In an attempt to bridge the gap between

the existing theory of DP and the practice of control design for large systems a new field of research has emerged. In addition to developing new methods based on DP and simulation, this research laid foundations for many existing RL methods and demarcated the science and the art in these methods. The thesis contributes to this field by formalizing actor-critic algorithms and providing an analysis of their convergence and rate of convergence.

There are at least three different approaches to handle the difficulties previously mentioned .

Model approximation One approach is to approximate a complex system by a simpler tractable model and apply an optimal policy, designed using DP, for the simpler model to the complex system. Taking this one step further, one can start with a class of simple tractable models and then determine, through some identification procedure, the best approximation (among this class) for the complex system.

Value function approximation The second approach is to approximate the optimal value function for the control problem. Often, the "form" of this function can be guessed to a certain extent in spite of the complexity of the system. For example, one might guess that the solution is monotonic or concave or polynomial in the state variables. Then, one can either hand code a value function or select the "best" approximation from a class of functions with these properties. Once an approximation to the value function is obtained, it can then be used to generate controls as if this were the exact value function.

Policy approximation Finally, instead of approximating the model or the value function, a good policy can be directly selected from a set of candidate policies, arrived at through various considerations like convenience of implementation and prior insights into the structure of an optimal policy. A straightforward strategy to selection of good policies, that is feasible only with finite and reasonably small set of candidate policies, is to evaluate the performance, in terms of overall reward, of each candidate policy. A more widely applicable approach is possible when the set of policies can be parameterized by a vector of reasonably small dimension. In this case, the selection of a good policy can be thought of as an optimization over the parameter space, where the reward of a parameter vector is the overall reward obtained by using the corresponding policy. Since the reward of a parameter can only be determined by simulation, stochastic optimization methods are used to determine good policy parameters. The candidate policies are often chosen to be randomized to incorporate sufficient exploration of decisions and also to make the overall reward a differentiable function of the policy parameters. In cases where the implementation of a randomized policy is not appropriate, the randomized policy obtained by this approach can be "rounded off" to its "nearest" deterministic policy.

While many permutations and combinations of the above three approaches are possible, in this thesis, we are primarily concerned with methods called actor-critic algorithms which combine policy approximation with value function approximation.

The aim of this thesis is to show that these methods have significant advantages over their counterparts which are based solely on policy approximation.

Though value function approximation methods are adequate for many applications, there are other important applications where the actor-critic or policy approximation methods might be more desirable than value function methods:

- In problems with complicated decision spaces, given a value function, the computation of the "optimal" decisions implied by the value function may be non-trivial because it involves an optimization of the value function over the decision space. In such problems, storing an explicit representation of a policy may be advantageous compared to an implicit representation based on a value function.

- In some problems, the policy implied by a value function approximation may not be implementable, e.g., due to the distributed nature of the state information. In these cases, it is more appropriate to optimize over an "implementable" family of policies than to approximate value function.

The general structure of actor-critic algorithms is illustrated by Figure 1-1. As the name suggests, actor-critic algorithms have two learning units, an actor and a critic, interacting with each other and with the system during the course of the algorithm. The actor has a tunable parameter vector that parameterizes a set of policies and at any given time instant, it generates a control using the policy associated with its current parameter value. The actor updates its parameter vector at each time step using its observations of the system and the information obtained from the critic. Similarly, at each time step, the critic updates its approximation of the value function corresponding to the current policy of the actor. Note the similarity between Actor-Critic methods and policy iteration (Puterman, 1994) in dynamic programming. While the value function approximation methods can be thought of as simulation-based counterpart of value iteration, actor-critic methods can be thought of as counterparts of policy iteration.

1.2 Previous Work

Adaptive control methods similar to actor-critic algorithms were first proposed in (Witten, 1977). The actor-critic architecture as described by Figure 1-1 was introduced and applied to the pole-balancing problem (Michie & Chambers, 1968) in the seminal work of (Barto *et al.*, 1983). Later, these methods were extensively studied in (Sutton, 1984; Anderson, 1986). A key step was taken by (Sutton, 1988) by separating the critic and treating it as a general method for policy evaluation (approximating the value function corresponding to a particular policy). This policy evaluation method was named *temporal difference learning*. Finally, (Watkins, 1989) developed a method called Q-learning for approximating the optimal value function. This separation of policy evaluation methods and the advent of Q-learning led to a shift of focus of RL research from actor-critic schemes to those based on value function approximation. Another reason for this shift of focus was that the convergence

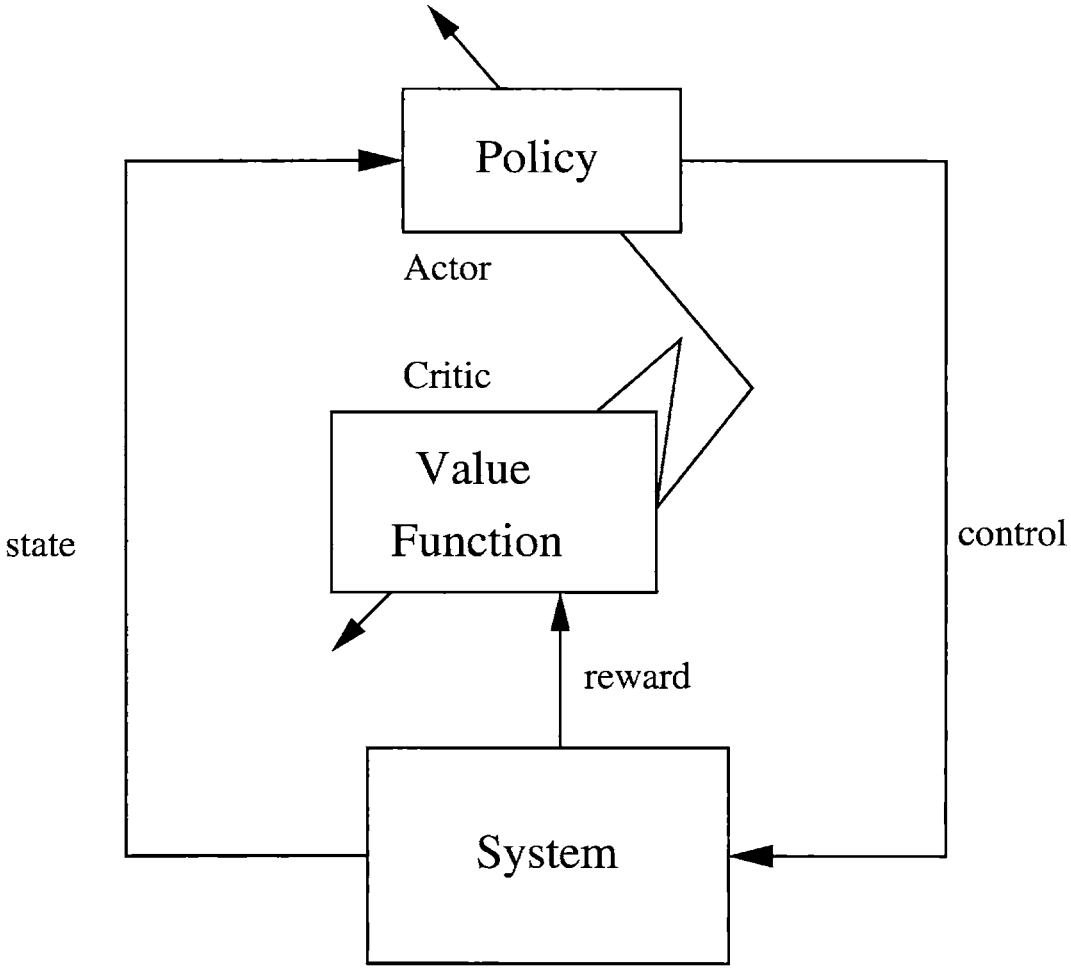

Figure 1-1: Actor-Critic Algorithms

of value function approximation methods was better understood than that of the actor-critic schemes.

The convergence of Q-learning and temporal difference learning with lookup-table representations and state aggregation was established in (Tsitsiklis, 1994; Jaakola *et al.*, 1994; Abounadi *et al.*, 2001; Abounadi *et al.*, 1998). Similarly, the convergence of temporal difference learning with linear function approximation was established in (Tsitsiklis & Van Roy, 1997; Tsitsiklis & Van Roy, 1999a). While the (optimal) value function approximation methods led to some impressive empirical results, they lacked satisfactory convergence guarantees except for some special function approximation schemes (Tsitsiklis & Van Roy, 1996; Ormoneit & Sen, 2000; Ormoneit & Glynn, 2001) and optimal stopping problems (Tsitsiklis & Van Roy, 1999b). For a textbook account of RL and its history see (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998).

Meanwhile, another approach based only on policy approximation was explored by (Glynn, 1986; Glynn, 1987) and later independently rediscovered by (Williams, 1992). The convergence analysis of these methods was carried out in (Marbach, 1998; Marbach & Tsitsiklis, 2001; Baxter & Barlett, 1999). In contrast to value function approximation, policy approximation schemes have good convergence guarantees but suffer from slow convergence.

Since their introduction in (Barto *et al.*, 1983), actor-critic algorithms have eluded satisfactory convergence analysis. Due to this lack of understanding and poor performance of policy approximation methods, value function based methods received much of the attention even though the actor-critic architecture predated value function approximation methods. In (Williams & Baird, 1990), an attempt was made to understand these algorithms through the analysis of asynchronous versions of policy iteration. A heuristic analysis of a special case of actor-critic algorithms was presented in (Kimura & Kobayashi, 1998).

The two main reasons the actor-critic methods are difficult to analyze are the following.

- First, for the critic to provide an accurate evaluation of the actor's policy, it should observe, for an indefinite amount of time, the behavior of the system under the influence of the actor's decisions. However, in actor-critic algorithms, the actor's decision policy changes continuously.

- Second, there can be large approximation errors, due to function approximation, in the critic's evaluation of the policy and it is not clear whether a policy can be improved even with an erroneous approximation of a value function.

In (Konda, 1997; Konda & Borkar, 1999), the first issue was circumvented by using different step-sizes for the actor and the critic: the critic uses infinitely large step-sizes relative to the actor. Therefore, the actor looks stationary to the critic and the critic behaves as if it can evaluate actor's policy instantly. However, the algorithms in (Konda, 1997; Konda & Borkar, 1999) use look-up table representations and therefore do not address the second issue.

The following section discusses these issues in more detail and describes the contribution of the thesis towards the understanding of actor-critic algorithms.

1.3 Problem Description

To make the discussion more concrete and to put the contributions of the thesis in perspective, a semi-formal discussion of some preliminaries and actor-critic algorithms is presented in this section. To keep the discussion simple, consider a finite state, discrete-time stochastic system, with state space $\mathbb{X}$, that evolves under the influence of a decision making agent as follows:

- At each time instant k, the agent chooses a decision U_k, from a finite set of choices $\mathbb{U}$, based on the current state of the system X_k.
- The agent receives a reward $g(X_k, U_k)$ for his decision at time k.
- The system moves to a new state X_{k+1} according to transition probabilities $p(X_{k+1}|X_k, U_k)$ where for each state x and decision u, $p(\cdot|x, u)$ is a probability mass function on the state space $\mathbb{X}$.

A policy is a mapping μ that assigns to each state x, a probability mass function $\mu(\cdot|x)$ on the decisions according to which the decisions are generated when the system is in state x. A special class of policies is the class of deterministic policies which can be thought of as functions $\mu : \mathbb{X} \to \mathbb{U}$. For each policy μ, the associated value function $V_\mu : \mathbb{X} \to \mathbb{R}$ (corresponding to the total reward problem) is defined as

$$V_\mu(x) = \sum_{k=0}^{\infty} \mathbf{E}_\mu[g(X_k, U_k)|X_0 = x],$$

where $\mathbf{E}_\mu$ denotes the expectation with respect to the probability distribution of the process $\{(X_k, U_k)\}$ when the agent uses policy μ to generate decisions. The optimal value function is defined by

$$V(x) = \max_\mu V_\mu(x).$$

For the value function to be finite for all policies, assume that there is an absorbing, reward-free, terminal state t which is hit with probability one from all starting states. A standard result in dynamic programming states that the optimal value function is the unique solution of the DP equation (or the Bellman equation):

$$V(x) = \max_u \left[g(x, u) + \sum_y p(y|x, u)V(y) \right].$$

Furthermore, deterministic policies μ which take decisions that maximize the r.h.s in the Bellman equation are optimal.

Note that both the value function V and the probabilities $p(y|x,u)$ are needed to compute the optimal policy. However, for some systems, the transition probabilities may be unavailable. For this reason, the concept of a Q-value function (also called state-decision value function) was introduced in (Watkins, 1989). The state-decision value function $Q_\mu : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ of a policy μ is defined as

$$Q_\mu(x,u) = \sum_{k=0}^{\infty} \mathbf{E}_\mu[g(X_k,U_k)|X_0 = x, U_0 = u].$$

It is now easy to see that the optimal state-decision value function

$$Q(x,u) = \max_\mu Q_\mu(x,u),$$

satisfies a modified Bellman equation

$$Q(x,u) = g(x,u) + \sum_y p(y|x,u) \max_{\bar{u}} Q(y,\bar{u})$$

and a policy which takes decisions that maximize $Q(x,u)$ is optimal.

Some value function based methods learn an approximation $\hat{Q}$ of the optimal state-decision value function using simulation. This learned approximation $\hat{Q}$ is used to obtain an approximation to an optimal policy by setting

$$\mu(x) = \arg\max_u \hat{Q}(x,u).$$

There are two problems with this approach. There are counterexamples showing that these methods may fail to converge. Furthermore, when they converge, there are no guarantees on the quality of the policies obtained using these methods (Bertsekas, 1995a).

In contrast, methods called temporal difference (TD) learning which approximate state or state-decision value function for a *particular policy* μ are well understood. These methods often use linear function approximation schemes. That is, they approximate the value function V_μ by a linear combination of basis functions:

$$\hat{V}(x) = \sum_i r^i \phi^i(x),$$

where the r^i are tunable parameters and the ϕ^i are predetermined basis functions, often called features. Such methods are guaranteed to converge and are widely applicable (Tsitsiklis & Van Roy, 1997). For example, TD methods can be used in the policy evaluation step of policy iteration. Policy iteration is a classical DP algorithm used to compute an optimal policy. It starts with a deterministic policy μ_0 and improves it iteratively as follows:

Policy Evaluation Compute the value function V_{μ_k} of the current policy μ_k.

Policy Improvement A new policy μ_{k+1} is obtained by

$$\mu_{k+1}(x) = \arg\max_u \left[g(x,u) + \sum_y p(y|x,u) V_{\mu_k}(y) \right].$$

It is well known, that if the policy evaluation is exact, the policy is strictly improved in each iteration, unless the current policy is optimal. Therefore, this method converges to an optimal policy after a finite number of iterations. When approximation methods such as TD are used in the policy evaluation step, the policy need not improve in each iteration. However, it can be shown that in the limit, the algorithm performs a random walk on a set of policies whose distance from optimality is at most linear in the approximation error during the policy evaluation phase (Bertsekas & Tsitsiklis, 1996).

The oscillatory behavior of approximate policy iteration is due to the fact that in the policy improvement step, the algorithm takes a large step in policy space, based only on an approximation. This, in turn, is due to the fact that the search for an optimal policy during policy iteration is restricted to deterministic policies and the fact that large steps (i.e., jumps) are needed to move from one deterministic policy to another. The oscillatory behavior can be potentially reduced if the search is performed over a set of randomized policies (which is continuous). However, the set of randomized policies can be huge in real world problems.

Therefore, we are led to the problem of optimizing the expected total reward over a family of policies $\{\mu_\theta; \theta \in \mathbb{R}^n\}$ parameterized by a vector θ of small dimension. The choice of this family of policies may be due to prior intuition, analysis, experimentation or simply a belief that it contains a good approximation to an optimal policy. Whatever the reasons behind this choice, once it is made, the optimization over the parameterized family of policies $\{\mu_\theta; \theta \in \mathbb{R}^n\}$ is a well defined problem and is central to our thesis. We assume that the system transition probabilities $p(y|x,u)$ are unknown but only a simulator of the system is available, and that we have an "actor" with a tunable parameter θ that generates decisions using the policy corresponding to the value of its parameter.

This problem has already been well studied in (Glynn, 1986; Glynn, 1987; Marbach, 1998; Baxter & Barlett, 1999). However, the algorithms proposed in these references do not involve value function approximation and can be viewed as "actor-only" algorithms. These "actor-only" methods suffer from large variance and therefore can be unsuitable for certain problems.

The aim of this thesis is to explore the role of value function approximation in optimizing over a parameterized family of policies and to understand actor-critic algorithms. In particular, we seek answers to the following questions:

- How is value function approximation relevant to the optimization over a family of policies?
- Value function approximation is crucial for policy improvement in policy iteration. How crucial is value function approximation for actor-critic algorithms?

- What are the advantages of actor-critic schemes over actor-only schemes?

While attempting to answer the above questions in Chapters 5 and 8, the thesis makes contributions on three fronts described separately in the following subsections.

1.3.1 Actor-Critic Algorithms

The thesis proposes in Chapter 5, two variants of actor-critic algorithms which use TD learning with linear function approximation for the critic. Chapter 5 also discusses various issues that arise in the design of the proposed algorithms. The critic part of the actor-critic algorithms is described and analyzed in Chapter 4. The gradient formulas on which the actor updates are based, are established in Chapter 2, for various reward criteria, and for systems with general state and action spaces. Under certain conditions presented in Chapters 2, 4 and 5 on

- the smoothness of dependence of the transition probabilities μ_θ on the policy parameters
- the ergodicity of the system
- the bounds on the growth of feature vectors used by the actor and the critic
- the relation between the features used by the critic and the family of policies used by the actor
- the relation between step-sizes used by the critic and the step-sizes used by the actor

we prove that the proposed algorithms converge, in a certain sense, with probability one. Chapter 2 considers some examples and verifies some of the assumptions.

1.3.2 Rate of Convergence (ROC)

ROC of episodic variants of a special case of the algorithms proposed in this thesis is studied in Chapter 8. The ROC of these algorithms is compared with that of their actor-only counterparts. We also study the rate of convergence of TD and related algorithms in Chapter 6. In particular, this chapter studies the effect of the choice of feature vectors and the eligibility traces on the ROC of TD algorithms

1.3.3 Stochastic Approximation

This thesis proves two new results on stochastic approximation that are applicable to a wider context. Chapter 3 contains a result on the tracking ability of linear iterations driven by Markov chains, which is useful in designing certain two-time-scale algorithms. Chapter 7 contains the first results on the ROC of two-time-scale algorithms.

1.4 Contributions and Outline of the Thesis

The rest of the thesis is divided into eight chapters. The second chapter formulates the central problem of the thesis, *i.e.*, the optimization Markov decision processes over a parameterized family of policies. The next three chapters are devoted to actor-critic algorithms and their convergence analysis. In the remaining chapters we study the rate of convergence of the algorithms proposed in the thesis. In the process, we establish a result on the rate of convergence of two-time scale stochastic approximation (Chapter 7) and also study the rate of convergence of temporal difference learning algorithms (Chapter 6). In Chapter 8, we use the results of the previous two chapters to understand the rate of convergence of actor-critic algorithms. The concluding chapter summarizes the thesis and discusses future research directions. The detailed contributions of each of the chapters are as follows.

Chapter 2

In Chapter 2, we start with a formal definition of Markov decision processes and randomized stationary policies with state and decision spaces that are not necessarily discrete. We formally describe the problem of optimization over a parametric family of policies. We present conditions under which the average reward is a well-behaved function of the policy parameters and we derive a formula for its gradient. We present an example and verify these conditions for that particular example. Throughout the chapter, we comment on the specialization of the assumptions and results to Markov decision processes with finite state space. We also extend these results to other criteria such as discounted and total rewards. Finally, we present the intuition behind the formulas for the gradient of overall reward.

Chapter 3

In this chapter, we prove a general result that will be used in the next chapter. This result concerns stochastic approximation driven by Markov noise whose transition probabilities change "slowly" with time. The proof is quite technical as we consider Markov chains with more general state spaces than usually encountered. It is the first available result on the tracking ability of stochastic approximation with decreasing step-sizes.

Chapter 4

This chapter describes several variants of TD algorithms used in the critic part of our algorithms. We describe TD algorithms for different reward criteria. We analyze the convergence of TD algorithms only for the case of average reward (the analysis of the algorithms for other criteria is similar). In particular, the central result of this chapter is the following:

In any variant, the difference between the critic's approximate value function and the value function to which the critic would converge if the actor parameters were to

be frozen at their current values, becomes arbitrarily small with time. This result is the first available on "controlled TD" albeit with "slowly" varying control policy.

Chapter 5

This chapter describes several variants of actor-critic algorithms for optimization of the average reward. We explain why these variants are expected to work, in view of the gradient formulas and the results on controlled TD of the previous chapters. The variants use either a TD(1) or a TD(λ), $\lambda < 1$, critic with linear function approximation. We also discuss the choice of basis functions for each of these critics. We prove a convergence result for the algorithms in this chapter, which is the first available on the convergence of actor-critic algorithms with function approximation. This result clarifies various ingredients needed for the convergence of actor-critic algorithms.

Chapter 6

This chapter studies the rate of convergence of temporal difference and related methods. We propose a common figure of merit for TD and related policy evaluation methods. We calculate this figure of merit for the case of TD and compare it with that of a related method called Least Squares TD (LSTD). The results obtained in this chapter are as follows.

1. We show that the sequence of value function approximations obtained by LSTD is dependent only on the subspace spanned by the basis functions.

2. The rate of convergence of TD is worse than that of LSTD.

3. We derive a bound on the rate of convergence of LSTD that captures the dependence on the factor λ and the mixing time of the Markov chain.

These results are the first on the rate of convergence of TD with function approximation.

Chapter 7

In order to analyze the rate of convergence of actor-critic algorithms, we need a theory on rate of convergence for two-time-scale stochastic approximation. In this chapter, we start with two-time-scale linear iterations driven by i.i.d. noise and present results on their rate of convergence. We then extrapolate these results to the more general case of non-linear iterations with Markov noise. We derive, as a consequence of our results, the well known result on optimality of Polyak's averaging. We also discuss informally the effect of separation of time-scales on the rate of convergence. The results of this chapter are the first on rate of convergence of two-time-scale algorithms.

Chapter 8

In this chapter, we study the rate of convergence of a class of actor-critic algorithms by combining various results from previous chapters. In particular, the following are the contributions of this chapter:

1. We show that, as in the case of LSTD, the rate of convergence of actor-critic algorithms depends only on the subspace spanned by the basis functions used by the critic.

2. If the critic uses TD(1), we show that the rate of convergence of actor-critic methods is same as that of certain actor-only methods.

3. We illustrate, with a numerical example, that the performance (both rate of convergence and quality of solutions) of actor-critic methods can be substantially better than that of corresponding actor-only methods.

Chapter 2

Optimization over a Family of Policies

In this chapter, Markov Decision Processes (MDPs) are defined formally and the problem of optimizing them over a parameterized family of policies is formulated precisely. The state and decision spaces of MDPs considered in this thesis are assumed to be either discrete or real Euclidean spaces, or a combination thereof. When the state and decision spaces are finite, most of the technical difficulties encountered in this chapter vanish.

In the next section, MDPs, randomized stationary policies (RSPs), and various objective criteria for optimizing MDPs are defined formally. Later, the formulas for the gradients of various objective criteria with respect to the parameters of RSPs are derived in separate sections.

The MDP framework is quite broad and includes a great many optimization models as special cases. For a comprehensive treatment of MDPs and their applications see (Puterman, 1994). For recent advances, see (Feinberg & Schwartz, 2001).

2.1 Markov Decision Processes

Markov decision processes are models of discrete time systems which evolve randomly under the influence of a decision maker or a controller. The influence of decisions on the evolution of the system is often described by an equation of the form

$$X_{k+1} = f(X_k, U_k, W_k),$$

where $\{W_k\}$ is an i.i.d. sequence of random variables that represents uncertainty in the system, and X_k, U_k denote the system state and the decision at time k. The decision maker obtains a reward for his decisions at each time step and the objective is to find a decision policy that maximizes an "overall reward".

An MDP is formally defined as follows.

Definition 2.1. A Markov decision process (MDP) is a discrete-time stochastic dynamical system described by the following ingredients:

1. State space $\mathbb{X}$, the elements of which are called *states*;
2. Decision or control space $\mathbb{U}$;
3. Transition kernel p, which is a stochastic kernel on the state space $\mathbb{X}$ given $\mathbb{X} \times \mathbb{U}$;
4. One-stage reward function $g : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$.

The system evolves as follows. Let X_k denote the state of the system at time k. If the decision maker takes decision U_k at time k, then:

1. The system moves to the next state X_{k+1} according to the probability law $p(\cdot|X_k, U_k)$;
2. The controller or decision maker receives a reward of $g(X_k, U_k)$.

The state space $\mathbb{X}$ and the decision space $\mathbb{U}$ are assumed to be of the form $\mathbb{R}^d \times \mathbb{E}$ where d is a nonnegative integer and $\mathbb{E}$ is a countable set.[1] The collections of Borel subsets of $\mathbb{X}$ and $\mathbb{U}$ are denoted by $\mathcal{B}(\mathbb{X})$ and $\mathcal{B}(\mathbb{U})$ respectively.

Informally, the "rule" with which the decision maker computes his decision based on his observations, is called the *decision policy.* A formal description of the space of all decision policies is quite technical and tedious. However, for all practical purposes, one can restrict attention to a special class of decision policies called Markov randomized polices. A Markov randomized policy (MRP) is one by which the decision maker randomly chooses a decision based only on the current state and time. An MRP is described by a sequence $\mu = \{\mu_k\}$ of stochastic kernels on $\mathbb{U}$ given $\mathbb{X}$. An MRP in which all the stochastic kernels μ_k are the same is called a randomized stationary policy (RSP). That is, the decision chosen using an RSP depends only on the current state but not the current time.

Note that the transition kernel alone does not completely describe the evolution of the system. The probability distribution of the initial state X_0 and the decision policy are also needed to describe completely the probability law of the state-decision process $\{(X_k, U_k)\}$. Furthermore, the state-decision process $\{(X_k, U_k)\}$ is a (time-inhomogenous) Markov chain when the decision policy is an MRP. Let $\mathbf{P}_{\mu,x}$ denote the probability law of the state-decision process when the starting state of the system is x and the decision policy is the MRP $\mu = \{\mu_k\}$. Let $\mathbf{E}_{\mu,x}$ denote expectation with respect to $\mathbf{P}_{\mu,x}$. The objective of optimizing an MDP is to maximize a "performance" or an "overall reward" criterion which can be one of the following:

Average Reward The average reward associated with policy μ and starting state x is defined as

$$\limsup_k \frac{1}{k} \sum_{l=0}^{k-1} \mathbf{E}_{\mu,x}[g(X_l, U_l)].$$

[1] However, the results of this thesis extend easily to problems in which the state and action spaces are Polish spaces (complete separable metric spaces).

Discounted Reward For this criterion, ρ is a fixed discount factor. The discounted reward associated with a policy μ and a starting state x is defined as

$$\sum_{k=0}^{\infty} \rho^k \mathbf{E}_{\mu,x}[g(X_k, U_k)].$$

Total Reward The total reward associated with a decision policy μ and a starting state x is defined as

$$\sum_{k=0}^{\infty} \mathbf{E}_{\mu,x}[g(X_k, U_k)].$$

The optimization of MDPs, in the classical sense, means finding an MRP that yields maximum overall reward. Under reasonable conditions, an optimal policy that is deterministic and stationary exists, and is optimal for all starting states. However, finding an optimal policy in most real world problems is computationally unfeasible. Therefore a different optimization problem is studied in this thesis. This involves finding a policy that is optimal over a family of RSPs parameterized by a finite number of parameters. The premise is that a good family of policies is known a priori and the optimization over this small family of policies is easier than optimization over all policies. Indeed, the new optimization problem can be viewed as a non-linear program on the parameter space of the family of policies. For this non-linear program to be manageable, we require that the parametric family of policies be such that the overall reward is differentiable with respect to the parameters of the policy. The policies are chosen to be randomized instead of deterministic, because of the following reasons:

1. In the case of discrete state and decision spaces, the set of all deterministic policies is also discrete. Therefore a "smooth" parameterization of the set of deterministic policies is not possible.

2. In the case of continuous decision spaces, even when a smooth parameterization of deterministic policies is possible, the map from a deterministic policy to the overall reward corresponding to that policy may not be differentiable. Therefore, the map from the policy parameters to the overall reward associated with the corresponding policy may be nonsmooth.

The new optimization problem is precisely formulated in the next section. The next section also discusses various issues involved in the choice of the family of policies to optimize over.

2.2 Problem Formulation

A parametric family of RSPs for MDPs with discrete decision spaces can be represented by a parameterized family of probability mass functions. Similarly, for an MDP

with continuous decision spaces, a parametric family of RSPs can be represented by a family of probability density functions. To describe a parametric family of RSPs for more general decision spaces, we use a "reference" or a "base" measure. Let ν be a fixed measure on the decision space $\mathbb{U}$. Then a parameterized family of RSPs can be represented by a parametric family of probability density functions with respect to the measure ν. In particular, we consider a parametric family of RSPs parameterized by $\theta \in \mathbb{R}^n$, and specified by a parametric family $\{\mu_\theta; \theta \in \mathbb{R}^n\}$ of positive measurable functions on $\mathbb{X} \times \mathbb{U}$ such that for each $x \in \mathbb{X}$, $\mu_\theta(\cdot|x)$ is a probability density function with respect to ν, i.e.,

$$\int_{\mathbb{U}} \mu_\theta(u|x)\nu(du) = 1, \quad \forall \theta, x.$$

The semantics of the family of functions $\mu_\theta(u|x)$ depend on the base measure ν. For example, if ν assigns positive mass to a decision u then, for a state x and parameter θ, $\mu_\theta(u|x)\nu(\{u\})$ equals the probability that decision u is taken when the policy used corresponds to θ and the system state is x. For discrete decision spaces, the most natural base measure is one that assigns unit mass to each decision. In this case, $\mu_\theta(u|x)$ denotes the probability that decision u is taken given that the current state is x, under the policy corresponding to θ. However, when the decision space is a combination of discrete and continuous spaces, the base measure ν and the semantics of the functions $\mu_\theta(u|x)$ might be more complicated. The following are some examples of a parameterized family of RSPs.

Example 2.2. In many dynamic programming problems, one can either guess or prove rigorously that the solution to the Bellman equation has certain structural properties, e.g., the solution is a quadratic in x, u, when $\mathbb{X}, \mathbb{U}$ are subsets of $\mathbb{R}$. In this case, one usually starts with an approximation architecture $\{Q_\theta : \theta \in \mathbb{R}^n\}$ for Watkin's Q-value function (Watkins, 1989; Watkins & Dayan, 1992; Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998) with the required structural properties and finds, through some learning algorithm, the best fit from this family for the true Q-value function. The hope is that the performance of the greedy policy,

$$\mu(x) = \arg\max_u Q^*(x, u),$$

with respect to the "best fit" Q^*, will be close to that of the optimal policy.

The knowledge of the structure of the solution to the Bellman equation can be used in a different way in the context of actor-critic algorithms. Assuming that the decision space is discrete, we first approximate the set of greedy policies with respect to the family $\{Q_\theta : \theta \in \mathbb{R}^n\}$ by the set of RSP's $\{\mu_\theta : \theta \in \mathbb{R}^n\}$, where

$$\mu_\theta(u|x) = \frac{\exp\left(\frac{Q_\theta(x,u)}{T}\right)}{\sum_u \exp\left(\frac{Q_\theta(x,u)}{T}\right)}, \quad \forall x \in \mathbb{X}, u \in \mathbb{U}.$$

We then apply our actor-critic algorithms to find the optimal RSP in this family of

RSPs. Note that the above approximation of greedy policies by RSP's depends on a parameter T which we call the temperature. As the temperature T goes to zero, the above policy becomes more and more deterministic and greedy with respect to the Q-function Q_θ. Therefore, if deterministic policies are more desirable than RSPs for an application, the result of our algorithm (say the policy associated with θ^*) can be rounded off to its nearest deterministic policy by taking the greedy policy with respect to the state-decision value function Q_{θ^*}.

The following is a more concrete example of a finite state MDP and a family of RSP's.

Example 2.3. Another concrete example of a family of RSP's is the class of threshold policies for multiple service, multiple resource (MSMR) systems (Jordan & Varaiya, 1994). In these systems the state space is a subset of ${\mathbb{Z}_+}^s$, where $\mathbb{Z}_+$ is the set of non-negative integers. The i-th component of the state vector $x = (x_1, \ldots, x_s)$ represents the number of customers of type i in the system who need resources $(a_{i1}, \ldots, a_{ip})$, and where j-th component a_{ij} denotes the amount of resources of type j. It is easy to see that, for a resource constrained system the state space is of the form

$$\mathbb{X} = \{x \in \mathbb{Z}_+ : Ax \leq r\},$$

where the vector r denotes available resources, and where it is the matrix with elements a_{ij}. The problem, in these systems, is to decide whether to admit or reject a customer when he arrives. Note that, when we model the above decision problem as an MDP the state consists of the vector x and the type i of the customer requesting service. A natural class of control policies is that of threshold policies in which we have a matrix B and thresholds b. Whenever a customer of type i arrives, Bx is calculated with x being the current state. The i-th component of Bx is compared with the i-th component of b and the customer is admitted only if the former is lesser than the later. This class of policies can be approximated by the following class of RSPs. If we denote the decision to admit by 1 and the decision to reject by 0 the threshold RSP's can be described as

$$\mu_{B,b}(1|x,i) = \frac{1}{2}\left(1 + \tanh\left(\frac{b_i - B_i x}{T}\right)\right),$$

where T is the temperature parameter.

Note that, in all of the above examples and in general, the family of RSPs chosen to optimize over approximates a deterministic family of policies. The accuracy of the approximation depends on the parameter T which we call the temperature of the family of policies. As the temperature T goes to zero, the family of RSPs "freezes" to a family of deterministic policies. While it is the optimal policy in the family of deterministic policies that one is usually interested in, the optimization is performed over the family of RSPs. Therefore, it is important to understand how randomization affects the "quality" of the policies and how the optimal policy in the family of deterministic policies is related to the optimal policy in the family of RSPs. The

answers to these questions in general are problem specific, and the extent to which our formulation is applicable for real world problems where the use of a randomized policy does not make much sense (e.g., inventory control) depends on these answers. However, there are some general comments that we can make about the appropriate choice of the temperature parameter T. If the temperature is chosen to be too low, the resulting non-linear program might involve optimizing a function that is almost discontinuous whereas if the temperature is chosen to be too high, the optimal policy in the family of RSPs might bear no relation to the optimal policy in the family of deterministic policies. In the rest of the thesis, it is assumed that these issues have been taken care of, and that the user has decided on a parameterized family of RSPs to optimize over. We denote this family of RSPs by $\{\mu_\theta; \theta \in \mathbb{R}^n\}$.

As we have noted earlier, a policy that is optimal over all MRPs is optimal for all starting states. However, the optimal policy in a parameterized family of RSPs might depend on the starting state unless the overall reward depends only on the policy but not the starting state. Therefore, a precise statement of our problem requires the probability distribution of the initial state X_0 also. We assume throughout the thesis that the probability distribution of the initial state X_0 of the system is ξ. Let $\mathbf{P}_{\theta,x}$ denote the probability law for the Markov chain $\{(X_k, U_k)\}$ controlled by the RSP associated with θ, when started from state x and let $\mathbf{E}_{\theta,x}$ denote the corresponding expectation. Similarly, for a probability distribution ϑ on $\mathbb{X}$ let $\mathbf{P}_{\theta,\vartheta}$ denote the law of the Markov chain $\{(X_k, U_k)\}$ whose starting state X_0 has distribution ϑ.

The central problem of this thesis is the simulation-based optimization of the average reward over the family of policies $\{\mu_\theta; \theta \in \mathbb{R}^n\}$. More precisely, consider an MDP with transition kernel p. Suppose the following are given

- A simulator of the transition kernel p that takes a state-decision pair (x, u) as input and generates a state according to the probability distribution $p(\cdot|x, u)$.
- A parametric family of RSPs $\{\mu_\theta; \theta \in \mathbb{R}^n\}$.

Assume that the simulator of p is memoryless in the sense that its current output is conditionally independent of its past inputs given the current input. For each parameter value θ, let $\bar{\alpha}(\theta)$ denote the average reward (assuming that the MDP is ergodic under all policies within the given family of policies) associated with this policy:

$$\bar{\alpha}(\theta) = \lim_k \frac{1}{k} \sum_{l=0}^{k-1} \mathbf{E}_{\theta,x} \left[g(X_k, U_k) \right].$$

The problem is to find a parameter θ that maximizes the function $\bar{\alpha}(\theta)$ using simulation.

The solution methodology adopted in the thesis is that of recursive gradient algorithms. However, since the function $\bar{\alpha}(\theta)$ is not directly accessible, simulation is used to estimate the gradient of average reward and update parameters in this estimated gradient direction. To arrive at an estimate of the gradient, a formula for

the gradient of $\bar{\alpha}(\theta)$ is needed. In the following sections, several objective criteria for optimization over the family of policies $\{\mu_\theta; \theta \in \mathbb{R}^n\}$ are studied. For each criterion, assumptions on the family of RSPs which ensure that the overall reward is well behaved as a function of the policy parameter vector θ are described. In particular, the given conditions ensure that the overall reward is differentiable. Furthermore, the parameterized family of functions $\psi_\theta : \mathbb{X} \times \mathbb{U} \to \mathbb{R}^n$ defined by

$$\psi_\theta(x, u) = \nabla \ln \mu_\theta(u|x), \quad \forall x, u, \tag{2.1}$$

where ∇ denotes the gradient with respect to θ, plays a central role in the formula for the gradient of the overall reward.

2.3 The Gradient of the Average Reward

The average expected reward or simply the average reward of an RSP μ is defined as:

$$\lim_k \frac{1}{k} \sum_{l=0}^{k-1} \mathbf{E}_{\mu,x} \left[g(X_k, U_k) \right].$$

The average reward is well defined for each policy in the family, under the following assumptions on the parameterized family of RSP's:

Assumption 2.4. *(Irreducibility and aperiodicity) For each $\theta \in \mathbb{R}^n$, the process $\{X_k\}$ controlled by the RSP associated with θ is irreducible and aperiodic.*

The notion of irreducibility is well known for discrete state spaces (there are no transient states and every state is accessible from every other state). For processes with more general state spaces, the usual notion of irreducibility is not applicable. More generally, a Markov chain can only be irreducible relative to a notion of "mass" or "size" which can be formalized by a measure χ on $\mathbb{X}$. Formally, the Markov chain $\{X_k\}$ is said to be χ-irreducible if for all $S \in \mathcal{B}(\mathbb{X})$ the following holds:

$$\chi(S) > 0 \qquad \Rightarrow \qquad \sum_k \mathbf{P}_{\theta,x}(X_k \in S) > 0 \qquad \forall x \in \mathbb{X}.$$

In other words, the Markov chain $\{X_k\}$ is irreducible if all sets that have positive "mass" are reachable with positive probability from any starting state. Note that for discrete state spaces (assume χ to be the counting measure) χ-irreducibility is equivalent to the usual notion of irreducibility of countable state space Markov chains. Meyn and Tweedie (Meyn & Tweedie, 1993) show that this notion of irreducibility is essentially independent of the measure χ relative to which it is defined. That is, if $\{X_k\}$ is χ -irreducible for some measure χ then there exists a maximal irreducibility measure $\bar{\chi}$ such that the Markov chain $\{X_k\}$ is χ'-irreducible if and only if χ' is absolutely continuous with respect to $\bar{\chi}$.

For finite MDPs, the above assumption would have been sufficient for the average reward to be well defined. However, for infinite MDPs, more assumptions are needed.

In particular, a certain part of the state space should be visited sufficiently often by the MDP. Furthermore, the MDP should "regenerate" itself every time it hits this part of the state space. The assumptions required to ensure that this happens contain two parts. The first part assumes that probabilities of the transitions out of the states in a certain set, are lower bounded by probabilities that are independent of the policy and the starting state. The second part is a Foster's Lyapunov criterion which guarantees that this set is reached fast enough from the states outside this set.

Assumption 2.5. (Uniform Geometric Ergodicity)

1. *There exists a positive integer N, a set $\mathbb{X}_0 \in \mathcal{B}(\mathbb{X})$, a constant $\delta > 0$ and a probability measure ϑ on $\mathbb{X}$ such that*

$$\mathbf{P}_{\theta,x}(X_N \in S') \geq \delta\vartheta(S') \qquad \forall\theta \in \mathbb{R}^n, \quad x \in \mathbb{X}_0, \quad S' \in \mathcal{B}(\mathbb{X}). \tag{2.2}$$

2. *There exists a function $\tilde{L} : \mathbb{X} \to [1, \infty)$ and constants $0 \leq \rho < 1$, $b > 0$ such that for each $\theta \in \mathbb{R}^n$,*

$$\mathbf{E}_{\theta,x}[\tilde{L}(X_1)] \leq \rho\tilde{L}(x) + bI_{\mathbb{X}_0}(x), \qquad \forall x \in \mathbb{X}, \tag{2.3}$$

where $I_{\mathbb{X}_0}(\cdot)$ is the indicator function of the set $\mathbb{X}_0$. We call a function satisfying the above inequality a stochastic Lyapunov function.

The above assumption is one of the most restrictive of all the assumptions we make. The second part of this assumption can be stated equivalently as follows. For a deterministic sequence $\{\theta_k\}$ of policy parameters, consider the time varying Markov chain obtained by using the MRP associated with the sequence $\{\theta_k\}$. For $s > 1$, consider the function

$$\tilde{L}(x) = \sup_{\{\theta_k\}} \mathbf{E}_{\{\theta_k\},x}\left[s^\tau\right], \tag{2.4}$$

where $\tau = \min\{k > 0 : X_k \in \mathbb{X}_0\}$ is the first time after time 0 that the Markov chain $\{X_k\}$ hits the set $\mathbb{X}_0$. If $\tilde{L}(\cdot)$ is finite for some $s > 1$, it is a matter of simple algebraic calculations to see that $\tilde{L}(x)$ satisfies (2.3). Conversely, it is also easy to see that if (2.3) is satisfied for some $\tilde{L}(\cdot)$, then for s sufficiently close to 1, the r.h.s. in Eq. (2.4) is finite for all x. The following theorem gives sufficient conditions on the family of RSPs $\{\mu_\theta; \theta \in \mathbb{R}^n\}$ for finite MDPs to satisfy Assumption 2.5.

Theorem 2.6. *If $\mathbb{X}$, $\mathbb{U}$ are finite and if for any $x \in \mathbb{X}$ and $u \in \mathbb{U}$, either*

$$\inf_\theta \mu_\theta(u|x) > 0,$$

or

$$\mu_\theta(u|x) = 0, \quad \forall\theta,$$

then Assumption 2.4 implies Assumption 2.5.

Proof. Assuming that $\mathbb{X}$ and $\mathbb{U}$ are finite, for each θ, the policy probabilities

$$(\mu_\theta(u|x); x \in \mathbb{X}, u \in \mathbb{U})$$

can be thought of as a vector of finite dimension. It is easy to see that all the vectors in the closure $\mathcal{P}$ of the set $\{\mu_\theta, \theta \in \mathbb{R}^n\}$ correspond to probabilities of some RSP. For a sequence of RSPs $\hat{\mu} = \{\mu_k\}$ in $\mathcal{P}$, let $\mathbf{P}_{\hat{\mu},x}$ denote the probability law of the MDP $\{(X_k, U_k)\}$ started from state x and controlled by the MRP $\hat{\mu}$. Since $\mu_\theta(u|x)$ is either zero for all θ or uniformly bounded below by a positive quantity, for two sequences $\hat{\mu}_1$ and $\hat{\mu}_2$, the probability measures $\mathbf{P}_{\hat{\mu}_1,x}(X_N \in \cdot)$ and $\mathbf{P}_{\hat{\mu}_2,x}(X_N \in \cdot)$ are absolutely continuous with respect to each other. Therefore, if

$$\mathbf{P}_{\hat{\mu},x}(X_N = y) > 0, \quad \forall x, y \in \mathbb{X}, \tag{2.5}$$

for some sequence $\hat{\mu}$ and integer $N > 0$, then the above strict inequality holds for all sequences $\hat{\mu}$. Furthermore, since the set of N-tuples $\{(\mu_1, \ldots, \mu_N); \mu_i \in \mathcal{P}\}$ is compact, and the map from this N-tuple to the distribution of X_N under $\mathbf{P}_{\hat{\mu},x}$ is continuous, there exists $\epsilon > 0$, such that

$$\mathbf{P}_{\hat{\mu},x}(X_N = y) > \epsilon, \quad \forall x, y \in \mathbb{X},$$

which in turn implies Assumption 2.5 with $\mathbb{X}_0$ being any subset of $\mathbb{X}$. Assumption 2.5(1) is satisfied as (2.5) holds for some N and a sequence $\mu_k = \mu_\theta$ for all k, for some θ. Assumption 2.5(2) is easily verified by checking that the function $\tilde{L}$ defined by Eq. (2.4) is finite. □

Assumption 2.5 and the geometric ergodicity results of (Meyn & Tweedie, 1993) imply that for each $\theta \in \mathbb{R}^n$, the Markov chains $\{X_k\}$ and $\{(X_k, U_k)\}$ have steady state distributions which we denote by $\pi_\theta(dx)$ and

$$\eta_\theta(dx, du) = \pi_\theta(dx)\mu_\theta(u|x)\nu(du),$$

respectively. Moreover, the steady state is reached at a geometric rate. In other words, if $\mathbf{E}_\theta$ denotes the expectation with respect to the stationary distribution of the process $\{(X_k, U_k)\}$, we have

$$\mathbf{E}_\theta[\tilde{L}(X_0)] < \infty, \quad \forall \theta,$$

and there exists $C > 0$ such that for any real valued measurable function f on $\mathbb{X}$ that satisfies $|f| \leq \tilde{L}$, we have

$$|\mathbf{E}_{\theta,x}[f(X_k)] - \mathbf{E}_\theta[f(X_0)]| \leq \rho^k C L(x) \qquad \forall x \in \mathbb{X}, \theta \in \mathbb{R}^n. \tag{2.6}$$

The previous assumptions ensure that the steady state distributions of $\{X_k\}$ and $\{(X_k, U_k)\}$ are well defined. However, this is not enough for the average reward function $\bar{\alpha}(\theta)$ to be well defined and differentiable. In particular, the steady state expectation $\mathbf{E}_\theta[|g(X_0, U_0)|]$ must be finite for all θ. Furthermore, the steady state

distributions π_θ and η_θ must be "differentiable" with respect to the policy parameters. The finiteness of the steady state expectations is automatically true for finite MDPs. For infinite MDPs, the expectations are finite if the one-stage reward function $g(x,u)$ is upper bounded by another function already known to have finite expectation. If g was just a function of the state then $\tilde{L}$ could serve as the upper bounding function. However, since g is a function of both state and decision, a new bounding function is required. Let $L : \mathbb{X} \times \mathbb{U} \to [1, \infty)$ be a function that satisfies the following condition:

Assumption 2.7. *For each $d > 0$ there is $K_d > 0$ such that*

$$\mathbf{E}_{\theta,x}[L(x,U_0)^d] \leq K_d \tilde{L}(x) \quad \forall x \in \mathbb{X}, \theta \in \mathbb{R}^n. \tag{2.7}$$

The function L will be used to bound various functions of state and decision encountered in the thesis. Since

$$\mathbf{E}_\theta[\tilde{L}(X_0)] < \infty, \quad \forall \theta,$$

it is easy to see that

$$\mathbf{E}_\theta[L(X_0,U_0)^d] < \infty, \quad \forall \theta \in \mathbb{R}^n, d > 0.$$

Note that if a function is upper bounded by L, then all its steady state moments are finite. This is a stronger conclusion than what is needed for such functions in this thesis. The above assumption can be weakened by restricting the values of d. However, this path is not pursued as we believe that the resulting conditions are quite artificial and will only be artifacts of the proof techniques employed here.

The function L is used in the following assumption that ensures the steady state distribution of the Markov chain under policy θ is "smooth" in the policy parameter θ. Throughout, ∇ denotes the gradient with respect to the policy parameter vector θ.

Assumption 2.8. (Differentiability)

1. $\mu_\theta(u|x) > 0, \qquad \forall x \in \mathbb{X}, u \in \mathbb{U}, \theta \in \mathbb{R}^n.$
2. *For each $x \in \mathbb{X}$ and $u \in \mathbb{U}$ the map $\theta \to \mu_\theta(u|x)$ is continuously differentiable.*
3. *There exists $K > 0$, such that for each $\theta \in \mathbb{R}^n$,*

$$\begin{aligned} \sup_{|\theta - \bar{\theta}| < \epsilon} \frac{\mu_\theta(u|x)}{\mu_{\bar{\theta}}(u|x)} &\leq 1 + \epsilon K L(x,u) \quad \forall x,u, \\ \sup_{|\theta - \bar{\theta}| < \epsilon} \left| \frac{\nabla \mu_{\bar{\theta}}(u|x)}{\mu_\theta(u|x)} \right| &\leq K L(x,u), \quad \forall x,u, \end{aligned}$$

for some $\epsilon > 0$ possibly depending on θ.

The first part of the above assumption can be weakened to require only that for any x, u, $\mu_\theta(u|x)$ either be positive for all θ or be zero for all θ. This makes our

theory and algorithms applicable to a larger class of RSPs. A further expansion of this class of policies is possible if the reference measure ν is taken to be dependent on the state. In other words, the RSP corresponding to θ is $\mu_\theta(u|x)\nu_x(du)$ where $\nu_x(du)$ is the reference measure corresponding to state x. While all these variations are perfectly compatible with the theory and algorithms of this thesis, we do not present our results in their full generality so that we do not obscure with technicalities the simple intuition behind our algorithms.

Recall the function ψ_θ defined by

$$\psi_\theta = \nabla \ln \mu_\theta(u|x).$$

An immediate consequence of the above differentiability assumption is that for each $\theta \in \mathbb{R}^n, d > 0$, we have

$$\mathbf{E}_\theta[|\psi_\theta(X_0, U_0)|^d] < \infty.$$

When all the assumptions described until now are satisfied, the proofs in (Glynn & L'Ecuyer, 1995) (which we will outline in the proof of the next theorem) can be imitated to show that for any measurable function $f : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$, such that $|f| \leq L$, the map $\theta \mapsto \mathbf{E}_\theta[f(X_0, U_0)]$ is bounded with bounded derivatives. We would like the average reward function $\bar{\alpha}(\cdot)$ to be bounded with bounded derivatives, and for this reason, we assume the following about the one-stage reward function $g : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$.

Assumption 2.9. *There exists $K > 0$ such that*

$$|g(x,u)| \leq KL(x,u), \qquad \forall \theta \in \mathbb{R}^n.$$

For each $\theta \in \mathbb{R}^n$, let $\mathcal{L}^2_\theta$ be the set of all functions f of state and decision such that

$$\mathbf{E}_\theta[|f(X_0, U_0)|^2] < \infty.$$

For two functions f_1, f_2 in $\mathcal{L}^2_\theta$, let

$$\langle f_1, f_2 \rangle_\theta = \mathbf{E}_\theta[f_1(X_0, U_0) f_2(X_0, U_0)].$$

Similarly, for two matrix-valued functions $A(\cdot)$ and $B(\cdot)$ on $\mathbb{X} \times \mathbb{U}$ such that

$$A_{ij}, B_{ij} \in \mathcal{L}^2_\theta, \quad \forall i, j,$$

let $\langle A, B \rangle_\theta$ denote

$$\mathbf{E}_\theta[A(X_0, U_0) B(X_0, U_0)].$$

For each θ, $\langle \cdot, \cdot \rangle_\theta$ defines an inner product on $\mathcal{L}^2_\theta$. Let $\| \cdot \|_\theta$ denote the corresponding norm. Let $\underline{1}$ denote the function in $\mathcal{L}^2_\theta$ that assigns the value 1 to all state-decision pairs. Since $L \in \mathcal{L}^2_\theta$, Assumption 2.9 implies that $g \in \mathcal{L}^2_\theta$ and therefore the average

reward function can be written as

$$\bar{\alpha}(\theta) = \mathbf{E}_\theta[g(X_0, U_0)] = \langle g, \underline{1} \rangle_\theta.$$

For each $\theta \in \mathbb{R}^n$, let P_θ denote the operator on $\mathcal{L}^2_\theta$ defined as

$$(P_\theta f)(x, u) = \mathbf{E}_{\theta,x}[f(X_1, U_1)|U_0 = u].$$

for all $(x, u) \in \mathbb{X} \times \mathbb{U}$ and $f \in \mathcal{L}^2_\theta$. We say that $Q \in \mathcal{L}^2_\theta$ is a solution of the Poisson equation with parameter θ if Q satisfies

$$Q = c - \bar{\alpha}(\theta)\underline{1} + P_\theta Q. \tag{2.8}$$

It is well known (see Proposition 17.4.1 from (Meyn & Tweedie, 1993)) that a solution to the Poisson equation with parameter θ exists and is unique up to a constant. That is, if Q_1, Q_2 are two solutions, then $Q_1 - Q_2$ and $\underline{1}$ are collinear. One family of solutions to the Poisson equation is the following:

$$Q_\theta(x, u) = \sum_{k=0}^{\infty} \mathbf{E}_{\theta,x} \left[(g(X_k, U_k) - \bar{\alpha}(\theta)) | \, U_0 = u \right].$$

(The convergence of the above series is a consequence of (2.6).) There are other (e.g., regenerative) representations of solutions to the Poisson equation which are useful for both purposes of analysis and derivation of algorithms.

The following theorem gives formula for the gradient of $\bar{\alpha}(\theta)$ in terms of any solution $Q_\theta : \mathbb{X} \times \mathbb{U} \to \mathbb{R}$ of the Poisson equation with parameter θ.

Theorem 2.10. *Under Assumptions 2.4, 2.5, 2.7, 2.8 and 2.9,*

$$\nabla \bar{\alpha}(\theta) = \langle \psi_\theta, Q_\theta \rangle_\theta. \tag{2.9}$$

Proof. (Outline) Fix some $\theta_0 \in \mathbb{R}^n$. Assume that there is an $\epsilon > 0$ and a family of functions $\{\hat{Q}_\theta : \mathbb{X} \times \mathbb{U} \to \mathbb{R}, |\theta - \theta_0| < \epsilon\}$ such that $\hat{Q}_\theta$ is a solution of the Poisson equation with parameter θ. Moreover, assume that the map $\theta \mapsto \hat{Q}_\theta(x, u)$ is differentiable for each state decision pair (x, u), and for each $x \in \mathbb{X}$, the family of functions

$$\left\{ \frac{|\hat{Q}_\theta(x, \cdot) - \hat{Q}_{\theta_0}(x, \cdot)|^2}{|\theta - \theta_0|^2} : |\theta - \theta_0| < \epsilon \right\} \tag{2.10}$$

has bounded expectation with respect to $\mu_\theta(u|x)\nu(du)$. Then, one can differentiate both sides of equation (2.8) with respect to θ at θ_0, to obtain

$$\nabla \bar{\alpha}(\theta_0)\underline{1} + \nabla \hat{Q}_{\theta_0} = P_{\theta_0}(\psi_{\theta_0} \hat{Q}_{\theta_0}) + P_{\theta_0}(\nabla \hat{Q}_{\theta_0}).$$

The interchange of differentiation and integration is justified by uniform integrability (2.10). Taking inner product with $\underline{1}$ on both sides of the above equation and using the facts that $|\nabla Q_\theta| \in \mathcal{L}^2_\theta$ and

$$\langle \underline{1}, P_{\theta_0} f \rangle_{\theta_0} = \langle \underline{1}, f \rangle_{\theta_0}, \qquad \forall f \in \mathcal{L}^2_{\theta_0},$$

we obtain $\nabla \bar{\alpha}(\theta_0) = \langle \hat{Q}_{\theta_0}, \psi_{\theta_0} \rangle_{\theta_0} = \langle Q_{\theta_0}, \psi_{\theta_0} \rangle_{\theta_0}$ where the second equality follows from the fact that $Q_{\theta_0} - \hat{Q}_{\theta_0}$ and $\underline{1}$ are necessarily collinear, and the easily verified fact $\langle \underline{1}, \psi_\theta \rangle_\theta = 0$.

To complete the proof, we need to show the existence of the family of functions $\hat{Q}_\theta$. This can be shown by imitating the proofs of Glynn and L'Ecuyer (Glynn & L'Ecuyer, 1995) which we will only outline here. Using Assumptions 2.4 and 2.5, one can construct (on a slightly enlarged probability space) a regeneration time $\hat{\tau}$ for the sampled Markov chain $\{X_{kN}\}$ controlled by any policy μ_θ using the splitting technique of Athreya, Ney (Athreya & Ney, 1978) and Nummelin (Nummelin, 1978) (see (Glynn & L'Ecuyer, 1995) for details of this construction). This regeneration time can be used to obtain the following representation for the average reward function $\bar{\alpha}(\cdot)$ and solutions $\hat{Q}_\theta$ to Poisson equations:

$$\begin{aligned} \bar{\alpha}(\theta) &= \frac{\mathbf{E}_{\theta,\vartheta}\left[\sum_{k=0}^{\hat{\tau}-1} g(X_k, U_k)\right]}{\mathbf{E}_{\theta,\vartheta}[\hat{\tau}]}, \\ \hat{Q}_\theta(x,u) &= \mathbf{E}_{\theta,x}\left[\sum_{k=0}^{\hat{\tau}-1} (g(X_k,U_k) - \bar{\alpha}(\theta)) \,|U_0 = u\right]. \end{aligned}$$

Furthermore, the positivity of $\mu_\theta(u|x)$ implies that the restriction of measures $\mathbf{P}_{\theta,\vartheta}$ and $\mathbf{P}_{\theta_0,\vartheta}$ to the σ-algebra $\mathcal{F}_{\hat{\tau}}$ corresponding to stopping time $\hat{\tau}$ are equivalent for every θ and θ_0. Therefore the above equations can be rewritten as

$$\begin{aligned} \bar{\alpha}(\theta) &= \frac{\mathbf{E}_{\theta_0,\vartheta}\left[\sum_{k=0}^{\hat{\tau}-1} g(X_k, U_k) l(\theta, \theta_0, \omega)\right]}{\mathbf{E}_{\theta_0,\vartheta}[\hat{\tau} l(\theta, \theta_0, \omega]}, \\ \hat{Q}_\theta(x,u) &= \mathbf{E}_{\theta_0,x}\left[\sum_{k=0}^{\hat{\tau}-1} (g(X_k,U_k) - \bar{\alpha}(\theta)) l(\theta,\theta_0,\omega) \,|U_0 = u\right], \end{aligned}$$

where the "likelihood ratio"

$$l(\theta, \theta_0, \omega) = \mathbf{E}_{\theta_0}\left[\frac{dP_{\theta,\vartheta}|_{\mathcal{F}_{\hat{\tau}}}}{dP_{\theta_0,\vartheta}|_{\mathcal{F}_{\hat{\tau}}}}\right]$$

is the Radon-Nikodym derivative of restriction of $\mathbf{P}_{\theta,\vartheta}(d\omega)$ to $\mathcal{F}_{\hat{\tau}}$ with respect to that of $\mathbf{P}_{\theta_0,\vartheta}(d\omega)$. Assumptions 2.8(1-3) imply that the map $\theta \mapsto l(\theta, \theta_0, \omega)$ is differentiable

at θ_0 and the family of functions

$$\left\{ \frac{|l(\theta, \theta_0, \omega) - 1|}{|\theta - \theta_0|} : |\theta - \theta_0| < \epsilon \right\}$$

is $\mathbf{P}_{\theta_0,\vartheta}(d\omega)$-uniformly integrable. This implies that the average reward function $\bar{\alpha}(\cdot)$ and the solutions $\hat{Q}_\theta$ of the Poisson equation are differentiable in θ, since differentiation with respect to θ and $\mathbf{E}_{\theta_0,\vartheta}[\cdot]$ can be interchanged. □

Recall that the assumptions of the previous theorem are quite strong. The following example illustrates how these can be verified in the context of an inventory control problem. The purpose of the example is to show that they are not vacuous. The verification of these assumptions for several other problems is of similar flavor.

Example 2.11. Consider a facility with $X_k \in \mathbb{R}$ amount of stock at the beginning of the k-th period, with negative stock representing the unsatisfied (or backlogged) demand. Let $D_k \geq 0$ denote the random demand during the k-th period. The problem is to decide the amount of stock to be ordered at the beginning of the k-th period based on the current stock and the previous demands. If $U_k \geq 0$ represents the amount of stock ordered at the beginning of the k-th period, then the cost incurred is assumed to be

$$c(X_k, U_k) = h \max(0, X_k) + b \max(0, -X_k) + pU_k,$$

where p is the price of the material per unit, b is the cost incurred per unit of backlogged demand, and h is the holding cost per unit of stock in the inventory. Moreover, the evolution of the stock X_k is given by

$$X_{k+1} = X_k + U_k - D_k, \quad k = 0, 1, \ldots .$$

If the demands $D_k, \quad k = 0, 1 \ldots$ are assumed to be nonnegative and i.i.d. with finite mean, then it is well known (e.g. see (Bertsekas, 1995b)) that there is an optimal policy μ^* of the form

$$\mu^*(x) = \max(S - x, 0)$$

for some $S > 0$ depending on the distribution of D_k. A good approximation for policies having the above form is the family of randomized policies in which S is chosen at random from the density

$$p_\theta(s) = \frac{1}{2T} \operatorname{sech}^2 \left(\frac{s - \bar{s}(\theta)}{T} \right)$$

where $\bar{s}(\theta) = e^\theta C/(1 + e^\theta)$. The constant C is picked based on our prior knowledge of an upper bound on the parameter S in an optimal policy. To define the family of

density functions $\{\mu_\theta\}$ for the above family of policies, let $\nu(du)$ be the sum of the Dirac measure at 0 and the Lebesgue measure on $[0,\infty)$. Then, the density functions are given by

$$\mu_\theta(0|x) = \frac{1}{2}\left(1 + \tanh\left(\frac{x - \bar{s}(\theta)}{T}\right)\right),$$

$$\mu_\theta(u|x) = \frac{1}{2T}\mathrm{sech}^2\left(\frac{x + u - \bar{s}(\theta)}{T}\right), \qquad u > 0.$$

The dynamics of the stock in the inventory, when controlled by policy μ_θ, are described by

$$X_{k+1} = \max(X_k, S_k) - D_k, \qquad k = 0, 1 \ldots,$$

where the $\{S_k\}$ are i.i.d. with density p_θ and independent of the demands D_k and the stock X_k. The Markov chain $\{X_k\}$ is easily seen to be χ-irreducible with χ being the Lebesgue measure on $\mathbb{R}$. To prove that the Markov chain is aperiodic it suffices to show that (2.2) holds with $N = 1$. Indeed, for $\mathbb{X}_0 = [-a, a], x \in \mathbb{X}_0$, and a Borel set B consider

$$\begin{aligned}
\mathbf{P}_{\theta,x}(X_1 \in B) &= \mathbf{P}_{\theta,x}(\max(x, S_0) - D_0 \in B), \\
&\geq \mathbf{P}_{\theta,x}(S_0 - D_0 \in B, S_0 \geq a), \\
&\geq \int_B \int_{a-t}^{\infty} \left(\inf_\theta p_\theta(t + y)\right) D(dy)dt,
\end{aligned}$$

where $D(dy)$ is the probability distribution of D_0.

To prove the Lyapunov condition (2.3), assume that D_k has exponentially decreasing tails. In other words, assume that there exists $\gamma > 0$ such that

$$\mathbf{E}[\exp(\gamma D_0)] < \infty.$$

Intuitively, the function $\tilde{L}(x) = \exp(\bar{\gamma}|x|)$, for some $\bar{\gamma}$ with $\min(\gamma, \frac{1}{T}) > \bar{\gamma} > 0$, should be a good candidate Lyapunov function. To see this, note that the Lyapunov inequality says that the Lyapunov function should decrease by a common factor outside some set $\mathbb{X}_0$. Let us try the set $\mathbb{X}_0 = [-a, a]$ for a sufficiently larger than C. If the inventory starts with a stock larger than a, then no stock is ordered with very high probability (since S_0 is most likely less than C) and therefore the stock decreases by D_0, decreasing the Lyapunov function by a factor of $\mathbf{E}[\exp(-\bar{\gamma}D_0)] < 1$. If the inventory starts with a large backlogged demand then most likely new stock will be ordered to satisfy all the backlogged demand decreasing the Lyapunov function to almost 1. This can be made precise as follows:

$$\begin{aligned}
\mathbf{E}_{\theta,x}[\tilde{L}(X_1)] &= \mathbf{E}_{\theta,x}[\exp(\bar{\gamma}|\max(x, S_0) - D_0|)] \\
&= \exp(\bar{\gamma}x)\mathbf{P}_{\theta,x}(S_0 \leq x)\mathbf{E}_{\theta,x}[\exp(-\bar{\gamma}D_0); D_0 \leq x]
\end{aligned}$$

$$+\exp(-\bar{\gamma}x)\mathbf{P}_{\theta,x}(S_0 \leq x)\mathbf{E}_{\theta,x}[\exp(\bar{\gamma}D_0); D_0 > x]$$
$$+\mathbf{E}_{\theta,x}[\exp(\bar{\gamma}|S_0 - D_0|); S_0 > x].$$

Note that the third term is bounded uniformly in θ, x since $\bar{\gamma} < \min(\frac{1}{T}, \gamma)$. The first term is bounded when x is negative and the second term is bounded when x is positive. Therefore the Lyapunov function decreases by a factor of $\mathbf{E}[\exp(-\bar{\gamma}D_0)] < 1$ when $x > a$ and decreases by a factor of $\mathbf{P}(S_0 \leq -a)\mathbf{E}[\exp(\bar{\gamma}D_0)] < 1$ for a sufficiently large. The rest of the assumptions are easy to verify with $L(x, u) = |x| + u$.

2.4 The Gradient of the Discounted Reward

In this section, a formula for the gradient of the discounted reward is derived using the results of the previous section. However, it is important to note that the same formula can be derived using methods more direct than the one presented here.

Unlike average reward, the discounted reward depends on the probability distribution of the initial state. Therefore, consider MDPs with a fixed initial distribution $\xi(dx)$. The discounted reward with discount factor $0 \leq \rho < 1$ is given by

$$\bar{\alpha}(\theta) = \sum_k \rho^k \mathbf{E}_{\theta,\xi}[g(X_k, U_k)].$$

For $\bar{\alpha}(\theta)$ to be finite, we need $g(x, u)$ to bounded in some sense. As in the previous section, the function $L : \mathbb{X} \times \mathbb{U} \to [1, \infty)$ will be used to bound various functions of state and decisions.

Assumption 2.12.

1. *There exists a function $\tilde{L}$ on $\mathbb{X}$ and a constant $\rho_1 > \rho$ such that $\tilde{L} \geq 1$ and for all θ, x we have*

$$\begin{aligned} \mathbf{E}_{\theta,\xi}[\tilde{L}(X_0)] &< \infty, \\ \mathbf{E}_{\theta,x}[\tilde{L}(X_1)] &\leq \left(\frac{1}{\rho_1}\right)\tilde{L}(x). \end{aligned}$$

2. *There exists a function $L : \mathbb{X} \times \mathbb{U} \to [1, \infty)$ such that the following holds: for every $d > 0$, there exists $K_d > 0$ such that*

$$\mathbf{E}_{\theta,x}[L(x, U_0)^d] \leq K_d \tilde{L}(x), \quad \forall \theta, x.$$

The boundedness assumption on $g(x, u)$ is the same as for the average reward case with L being the function described above. This assumption ensures that $\bar{\alpha}(\theta)$ is finite for each θ. If the differentiability Assumption 2.8 is also satisfied with the function L satisfying Assumption 2.12 instead of Assumption 2.7, a formula for the gradient of $\bar{\alpha}(\theta)$ can be derived by showing that $\bar{\alpha}(\theta)$ is the average reward of a certain artificial MDP controlled by a parametric family of policies. Intuitively, the discounted reward

is the expected total reward up to time τ where τ is a geometric random variable with parameter $(1-\rho)$ independent of MDP $\{(X_k, U_k)\}$. If the time τ is thought of as a hitting time for a reward-free artificial state t from which the artificial MDP jumps to any of the state in $\mathbb{X}$ according to the probability distribution ξ, it is easy to see that the average reward of such an MDP is $(1-\rho)\bar{\alpha}(\theta)$. Formally, consider an artificial MDP with transition kernel:

$$\tilde{p}(S|x,u) = \rho p(S|x,u) + (1-\rho)\xi(S), \quad \forall x, u.$$

It is easy to see that part 1 of Assumption 2.5 is satisfied with $\mathbb{X}_0$ being any subset of $\mathbb{X}$, $N = 1$ and $\vartheta = \xi$. Part 2 of Assumption 2.5, can also be verified with $\mathbb{X}_0$ being a subset of $\mathbb{X}$ of the form

$$\mathbb{X}_0 = \{x | \tilde{L}(x) < C\}$$

for a suitable constant C. Furthermore, the steady state probability measure $\tilde{\pi}_\theta$ for the artificial MDP controlled by RSP θ is given by

$$\tilde{\pi}_\theta(S) = (1-\rho)\pi_\theta(S),$$

where π_θ is the finite measure

$$\sum_k \rho^k \mathbf{P}_{\theta,\xi}(X_k \in \cdot).$$

Similarly, the average reward $\tilde{\alpha}$ and a solution $\tilde{Q}_\theta$ to Poisson equation associated with the artificial MDP are given by

$$\begin{aligned} \tilde{\alpha}(\theta) &= (1-\rho)\bar{\alpha}(\theta), \\ \tilde{Q}_\theta(t,u) &= 0, \quad \forall u, \\ \tilde{Q}_\theta(x,u) &= Q_\theta(x,u) - \bar{\alpha}(\theta), \quad \forall x, u, \end{aligned}$$

where for each θ, Q_θ is given by

$$Q_\theta(x,u) = \sum_k \rho^k \mathbf{E}_{\theta,x}\left[g(X_k, U_k) | U_0 = u\right].$$

Using these relations and the result on the gradient of average reward, we have the following result.

Theorem 2.13. *Under Assumptions 2.12, 2.9 and 2.8,*

$$\nabla\bar{\alpha}(\theta) = \langle Q_\theta, \psi_\theta \rangle_\theta$$

where for each θ, and any two functions $f_1(x,u)$ and $f_2(x,u)$,

$$\langle f_1, f_2 \rangle_\theta = \sum_k \rho^k \mathbf{E}_{\theta,\xi}[f_1(X_k, U_k) f_2(X_k, U_k)].$$

2.5 The Gradient of the Total Reward

In this section, a formula for the gradient of the total reward is derived using the results on average reward. As in the case of discounted reward, the formula for the gradient of the total reward can also be derived using methods more direct than the one presented here.

As in the case of discounted reward, the total reward also depends on the probability distribution of the initial state X_0. Therefore, the probability distribution of X_0 is assumed to be a fixed ξ. In this case, the total reward $\bar{\alpha}(\theta)$ of the policy associated with θ is given by

$$\sum_k \mathbf{E}_{\theta,\xi}[g(X_k, U_k)].$$

Our assumptions for the total reward problem are quite restrictive because the problem is difficult to handle for more general cases. In particular, we assume that there exists a reward free absorbing state t that is reachable from any state. As in the previous section, a function $L : \mathbb{X} \times \mathbb{U} \to [1, \infty)$ will be used to bound various functions of state and decisions. Let L satisfy the following assumption.

Assumption 2.14.

1. *There exists a function $\tilde{L}$ on $\mathbb{X}$ and a constant $0 \leq \rho < 1$ such that*

$$\begin{aligned} \tilde{L}(x) &= 0 \quad \textit{if} \quad x = t, \\ &\geq 1 \quad \textit{otherwise}, \end{aligned}$$

and

$$\begin{aligned} \mathbf{E}_{\theta,\xi}[\tilde{L}(X_0)] &< \infty, \\ \mathbf{E}_{\theta,x}[\tilde{L}(X_1)] &\leq \rho \tilde{L}(x), \quad \forall \theta, x. \end{aligned}$$

2. *For each $d > 0$, there exists $K_d > 0$ such that*

$$\mathbf{E}_{\theta,x}[L(x, U_0)^d] \leq K_d \tilde{L}(x), \quad \forall \theta, x.$$

As in the previous section, assuming that Assumptions 2.9 and 2.8 are satisfied with L defined above, the formula for the gradient of total reward can be derived using the result for the gradient of the average reward.

Consider the artificial MDP in which every time the state t is hit, the next state is chosen randomly with probability distribution ξ independent of the control. That is, consider an MDP with the following transition kernel:

$$\begin{aligned} \tilde{p}(S|t,u) &= \xi(S), \quad , \forall x, u, \\ \tilde{p}(S|x,u) &= p(S|x,u), \quad \forall x \neq t, u. \end{aligned}$$

Every time the artificial MDP hits t, it regenerates. Assuming, without loss of generality, that the support of ξ is $\mathbb{X}$ (including t), it is easy to see that the artificial MDP is irreducible and aperiodic. Therefore, the artificial MDP satisfies Assumption 2.4. It also satisfies Assumption 2.5 with $\mathbb{X}_0 = \{t\}$ and with the Lyapunov function

$$\begin{aligned} \tilde{L}_1(x) &= 1, \quad \text{if} \quad x = t, \\ &= \tilde{L}(x), \quad \text{otherwise.} \end{aligned}$$

Furthermore, if τ represents the first time the (original or artificial) MDP hits the terminal state t and $\bar{\tau}(\theta)$ is its expectation:

$$\bar{\tau}(\theta) = \mathbf{E}_{\theta,\xi}[\tau],$$

then the steady state distribution of the artificial MDP under policy θ is given by

$$\tilde{\pi}_\theta(\cdot) = \frac{1}{\bar{\tau}(\theta)}\pi_\theta(\cdot) + \frac{1}{\bar{\tau}(\theta)}\delta_t(\cdot),$$

where π_θ is the finite measure defined by

$$\sum_k \mathbf{P}_{\theta,\xi}(X_k \in \cdot, \tau > k).$$

Therefore, the average reward $\tilde{\alpha}(\theta)$ corresponding to the artificial MDP under policy θ is given by

$$\tilde{\alpha}(\theta) = \frac{\bar{\alpha}(\theta)}{\bar{\tau}(\theta)}.$$

In other words, the total reward associated with policy θ is given by

$$\bar{\alpha}(\theta) = \frac{\tilde{\alpha}(\theta)}{(1/\bar{\tau}(\theta))}. \tag{2.11}$$

Note that $(1/\bar{\tau}(\theta))$ is the average reward for the artificial MDP under policy θ when the one-stage reward function is $I_{\{t\}}(x)$. Similarly, the solution $\tilde{Q}_\theta$ to the Poisson equation corresponding to the policy θ and one-stage reward function g is given by

$$\begin{aligned} \tilde{Q}_\theta(x,u) &= 0 \quad \text{if} \quad x = t, \\ &= Q_\theta(x,u) - \bar{\alpha}(\theta)T_\theta(x,u), \end{aligned}$$

where for each θ, Q_θ and T_θ are given by

$$\begin{aligned} Q_\theta(x,u) &= \sum_k \mathbf{E}_{\theta,x}\left[g(X_k,U_k)\middle|\, U_0 = u\right], \\ T_\theta(x,u) &= \mathbf{E}_{\theta,x}[\tau | U_0 = u]. \end{aligned}$$

Similarly, the solution to the Poisson equation corresponding to the policy θ and one-

stage reward function $I_{\{t\}}(x)$ is $-\bar{\tau}(\theta)T_\theta(x,u)$. Using these solutions and Theorem 2.10 to calculate the gradients of $1/\bar{\tau}$ and $\bar{\alpha}$, and using Equation (2.11) we can obtain the following formula for the gradient of total reward.

Theorem 2.15. *Under Assumptions 2.14, 2.9 and 2.8, we have*

$$\nabla\bar{\alpha}(\theta) = \langle Q_\theta, \psi_\theta\rangle_\theta,$$

where for any two functions f_1 and f_2

$$\langle f_1, f_2\rangle_\theta = \mathbf{E}_{\theta,\xi}\left[\sum_{k=0}^{\tau-1} f_1(X_k,U_k)f_2(X_k,U_k)\right].$$

2.6 Closing Remarks

We are not the first to derive formula for the gradient of the overall reward with respect to policy parameters. The formula for the gradient of total reward over a deterministic finite horizon was derived in (Williams, 1992). In (Glasserman, 1991; Cao & Chen, 1997), an approach to deriving gradient formulas for objective functions on generalized semi-Markov processes was presented. The likelihood ratio approach to gradient estimation was introduced in (Glynn, 1987; Glynn & L'Ecuyer, 1995). The gradient of the average reward for finite MDPs was also derived in (Marbach & Tsitsiklis, 2001; Baxter & Barlett, 1999). These works also propose several algorithms based on this formula. While the gradient formulas were derived previously by various authors, their interpretation as an inner product is new. This interpretation was independently arrived at in (Sutton *et al.*, 2000). It implies that actor-critic algorithms are robust to approximation errors (in value function) that are orthogonal to the functions ψ_θ^i's which depend only on the policy parameterization.

The formula for the gradient of discounted reward was first derived in (Sutton *et al.*, 2000). However, our approach to deriving gradient formulas by reducing the overall reward to an average reward is new. Furthermore, (Sutton *et al.*, 2000) considers finite MDPs whereas we deal with MDPs with more general state and decision spaces. Also, the derivation of the gradient formula for average reward by differentiating both sides of the Poisson equation is new, much more direct and simpler than previous methods.

Technically, our assumptions bear a lot of resemblance with those of (Glynn & L'Ecuyer, 1995). In fact, the proofs of differentiability of the reward function and the solutions to the Poisson equation are inspired by this work. The difference in assumptions is due to the fact that they consider Markov chains evolving according to recursions of the form

$$X_k = f(X_k, W_k),$$

where W_k are $\mathbb{R}^d$ valued i.i.d. random variables whose distribution depends on the parameter θ. The global assumptions concerning the behavior of the Markov chain

$\{X_k\}$ are the same in both cases. However, the local assumptions like differentiability are stated in terms of the parametric family of densities of W_k, whereas our local assumptions are stated in terms of the parametric family of RSPs. Fundamental to both the works in particular, and the theory of Markov chains on general state spaces in general, is the splitting technique of Nummelin (Nummelin, 1978), and Athreya and Ney (Athreya & Ney, 1978) which allows us to extend results for the countable or finite state case to results on Markov chains with general state spaces with appropriate modifications.

Chapter 3

Linear Stochastic Approximation Driven by Slowly Varying Markov Chains

In this chapter, we state and prove a new theorem regarding the tracking ability of linear stochastic iterations driven by a slowly varying Markov chain. This result will be used in the next chapter to prove the convergence of the critic part of our actor-critic algorithms.

Convergence of stochastic approximation driven by stationary or asymptotically stationary ergodic noise has been extensively studied in the stochastic approximation literature (Kushner, 1984; Benveniste *et al.*, 1990; Duflo, 1997; Kushner & Yin, 1997). The gist of these results is that the iterate converges to a point that depends on the update direction and the statistics of the driving noise. In some of the applications, the statistics of the driving noise may change with time. In such cases, the point to which the algorithm would converge, if the noise was held stationary with current statistics, also changes with time. The objective of stochastic approximation is to track this changing point closely after an initial transient period. Such algorithms were named adaptive algorithms as they adapt themselves to the changing environment. For a textbook account of adaptive algorithms see (Benveniste *et al.*, 1990).

The tracking ability of adaptive algorithms has been analysed in several contexts (Widrow *et al.*, 1976; Eweda & Machi, 1984). The consensus is that the usual stochastic approximation with constant step-sizes can adapt to changes in statistics of the driving noise that are "slow" relative to the step-size of the algorithm. However, algorithms with decreasing step-sizes have never been touched upon because one would have to assume that the environment changes slowly relative to the step-sizes employed by the user. This assumption is too restrictive to be satisfied in applications. However, if the statistics of the driving noise is deliberately changed by the user at a rate slower than that of stochastic approximation, it would be meaningful to study the tracking ability of stochastic approximation with decreasing step-sizes. It is this scenario we are interested in and the following result is the first on "adaptive" algorithms with decreasing step-sizes.

Consider a stochastic process $\{Y_k\}$ taking values in a Polish (complete, separable, metric) space $\mathbb{Y}$ with Borel σ-field denoted by $\mathcal{B}(\mathbb{Y})$. Let $\{P_\theta(y, d\bar{y}); \theta \in \mathbb{R}^n\}$ be a parameterized family of transition kernels on $\mathbb{Y}$. Consider the following iteration to update a vector $r \in \mathbb{R}^m$:

$$r_{k+1} = r_k + \gamma_k(h_{\theta_k}(Y_k) - G_{\theta_k}(Y_k)r_k) + \gamma_k\xi_{k+1}r_k. \tag{3.1}$$

In the above iteration, $\{h_\theta(\cdot), G_\theta(\cdot) : \theta \in \mathbb{R}^n\}$ is a parameterized family of m-vector valued and $m \times m$-matrix valued measurable functions on $\mathbb{Y}$. For any measurable function f on $\mathbb{Y}$, let $P_\theta f$ denote the measurable function $y \mapsto \int P_\theta(y, d\bar{y})f(\bar{y})$. Let $\mathcal{F}_k$ be the σ-field generated by $Y_l, r_l, \theta_l, l \leq k$. We make the following assumptions.

Assumption 3.1. *1. For a measurable set $A \subset \mathbb{Y}$,*

$$\mathbf{P}(Y_{k+1} \in A|\mathcal{F}_k) = \mathbf{P}(Y_{k+1} \in A|Y_k, \theta_k) = P_{\theta_k}(Y_k, A).$$

2. The step-size sequence $\{\gamma_k\}$ is deterministic, non-increasing, and satisfies

$$\sum_k \gamma_k = \infty, \qquad \sum_k \gamma_k^2 < \infty.$$

3. The (random) sequence of parameters $\{\theta_k\}$ satisfies:

$$|\theta_{k+1} - \theta_k| \leq \beta_k H_k,$$

for some nonnegative process $\{H_k\}$ with bounded moments and deterministic sequence $\{\beta_k\}$ such that

$$\sum_k \left(\frac{\beta_k}{\gamma_k}\right)^d < \infty$$

for some $d > 0$.

4. ξ_k is a $m \times m$-matrix valued $\mathcal{F}_k$-martingale difference with bounded moments i.e.,

$$\mathbf{E}[\xi_{k+1}|\mathcal{F}_k] = 0, \qquad \sup_k \mathbf{E}[|\xi_k|^d] < \infty, \quad \forall d > 0.$$

5. (Existence of solutions to the Poisson Equation) For each θ, there exist $\bar{h}(\theta) \in \mathbb{R}^m$, $\bar{G}(\theta) \in \mathbb{R}^{m\times m}$, $\hat{h}_\theta : \mathbb{Y} \to \mathbb{R}^m$, and $\hat{G}_\theta : \mathbb{Y} \to \mathbb{R}^{m\times m}$ that satisfy the Poisson equation. That is, for each $y \in \mathbb{Y}$,

$$\begin{aligned} \hat{h}_\theta(y) &= h_\theta(y) - \bar{h}(\theta) + (P_\theta\hat{h}_\theta)(y), \\ \hat{G}_\theta(y) &= G_\theta(y) - \bar{G}(\theta) + (P_\theta\hat{G}_\theta)(y). \end{aligned}$$

6. *(Boundedness) For some constant C and for all θ, we have*

$$\max(|\bar{h}(\theta)|, |\bar{G}(\theta)|) \leq C.$$

7. *(Boundedness in expectation) For any $d > 0$, there exists $C_d > 0$ such that*

$$\sup_k \mathbf{E}[|f_{\theta_k}(Y_k)|^d] \leq C_d,$$

where $f_\theta(\cdot)$ represents any of the functions $\hat{h}_\theta(\cdot), h_\theta(\cdot), \hat{G}_\theta(\cdot), G_\theta(\cdot)$.

8. *(Lipschitz continuity) For some constant $C > 0$, and for all $\theta, \bar{\theta} \in \mathbb{R}^n$,*

$$\max(|\bar{h}(\theta) - \bar{h}(\bar{\theta})|, |\bar{G}(\theta) - \bar{G}(\bar{\theta})|) \leq C|\theta - \bar{\theta}|.$$

9. *(Lipschitz continuity in expectation) There exists a positive measurable function $C(\cdot)$ on $\mathbb{Y}$ such that for each $d > 0$,*

$$\sup_k \mathbf{E}[C(Y_k)^d] < \infty,$$

and

$$|f_\theta(y) - f_{\bar{\theta}}(y)| \leq C(y)|\theta - \bar{\theta}|,$$

where $f_\theta(\cdot)$ is any of the functions $\hat{h}_\theta(\cdot), h_\theta(\cdot), \hat{G}_\theta(\cdot), G_\theta(\cdot)$.

10. *There exists $a > 0$ such that for all $r \in \mathbb{R}^m$ and $\theta \in \mathbb{R}^n$:*

$$r'\bar{G}(\theta)r \geq a|r|^2.$$

Theorem 3.2. *If Assumptions 3.1(1-10) are satisfied then*

$$\lim_k |\bar{G}(\theta_k)r_k - \bar{h}(\theta_k)| = 0.$$

The above theorem, when $\theta_k = \theta^*$ for all k, is a special case of Theorem 17 on page 239 of (Benveniste *et al.*, 1990). However, since θ_k is changing, albeit slowly, we need to use different techniques to prove the above result. Our proofs use a combination of techniques used in (Benveniste *et al.*, 1990; Borkar & Meyn, 2000; Borkar, 1996). In the next subsection, we present an overview of the proof and the intuition behind it.

3.1 Overview of the Proof

The techniques to prove convergence of stochastic approximation can be broadly classified into two categories: martingale methods (probabilistic) and ODE methods (deterministic). In martingale methods, one constructs a super-martingale (or

almost super-martingale) and uses martingale convergence theorems to infer the convergence of the super-martingale, which in turn implies convergence of the stochastic approximation. In ODE methods, one views the stochastic approximation update as a random perturbation of a deterministic iteration and proves that the perturbation noise is asymptotically negligible. This implies that the asymptotic behavior of the deterministic iteration and the asymptotic behavior of stochastic approximation are the same. Although one uses martingale convergence theorems to prove that the perturbation noise is asymptotically negligible, the rest of the proof is essentially deterministic.

It is the second (ODE) method that we use in our proofs. In particular, note that the sequence $\hat{\rho}_k = \bar{G}(\theta_k)r_k - \bar{h}(\theta_k)$ satisfies the iteration:

$$\hat{\rho}_{k+1} = \hat{\rho}_k - \gamma_k \bar{G}(\theta_{k+1})\hat{\rho}_k + \gamma_k \epsilon_{k+1}^{(1)} + \gamma_k \epsilon_{k+1}^{(2)}.$$

where

$$\begin{aligned}
\epsilon_{k+1}^{(1)} &= \bar{G}(\theta_{k+1})(h_{\theta_k}(Y_k) - \bar{h}(\theta_k)) - \bar{G}(\theta_{k+1})(G_{\theta_k}(Y_k) - \bar{G}(\theta_k))r_k \\
&\quad + \bar{G}(\theta_{k+1})\xi_{k+1}r_k, \\
\epsilon_{k+1}^{(2)} &= \frac{1}{\gamma_k}((\bar{G}(\theta_{k+1}) - \bar{G}(\theta_k))r_k - (\bar{h}(\theta_{k+1}) - \bar{h}(\theta_k))).
\end{aligned}$$

Assumption 3.1(5) implies that for the Markov chain Y_k with transition kernel P_θ, the vector $h_\theta(Y_k)$ and the matrix $G_\theta(Y_k)$ have steady state expected values $\bar{h}(\theta)$ and $\bar{G}(\theta)$ respectively. Therefore, we argue that the effect of noise $\epsilon_{k+1}^{(1)}$ should be "averaged out" in the long term. Similarly, since θ_k is changing very slowly with respect to the step size γ_k, we expect that $\epsilon_{k+1}^{(2)}$ goes to zero. The proof is then completed by showing that the noise components $\epsilon_{k+1}^{(i)}, i = 1, 2$, can be taken out of the picture and observing that the sequence $\hat{\rho}_k$ converges to zero if the perturbation noise is zero.

We formalize this intuition in the next two subsections. Note that the noise components are affine in r_k and therefore can be very large if r_k is large. A major part of this proof involves the proof of boundedness of the iterates r_k which is presented separately in the next subsection. The claimed convergence is then proved in the subsequent subsection. The approach and techniques used here are essentially specialization of general techniques developed in (Benveniste *et al.*, 1990; Borkar & Meyn, 2000). The following are some facts useful in proving both boundedness and convergence.

Lemma 3.3. *Let $\{a_k\}$ be a non-negative sequence satisfying*

$$a_{k+1} \leq \lambda a_k + \delta_k,$$

for some $0 \leq \lambda < 1$ and non-negative sequence $\{\delta_k\}$.

1. *If $\sup_k \delta_k < \infty$ then $\sup_k a_k < \infty$.*
2. *If $\delta_k \to 0$ then $a_k \to 0$.*

Lemma 3.4. *If an $m \times m$ matrix G is such that*

$$r'Gr \geq \delta|r|^2, \qquad \forall r \in \mathbb{R}^m,$$

then for sufficiently small $\gamma > 0$,

$$|(I - \gamma G)r| \leq (1 - \frac{1}{2}\gamma\delta)|r|.$$

Proof. $|(I - \gamma G)r|^2 \leq |r|^2 - 2\gamma\delta|r|^2 + \gamma^2|r|^2C^2 \leq (1 - \gamma\delta)|r|^2$, for sufficiently small $\gamma > 0$. The result follows from the inequality $\sqrt{(1-x)} \leq (1 - \frac{x}{2})$ for $0 \leq x \leq 1$. □

3.2 Proof of Boundedness

Note that the difference between two successive iterates at time k is of the order γ_k which goes to zero as k goes to infinity. Therefore, to study the asymptotic behavior of the sequence r_k we focus on a subsequence r_{k_j} where the sequence of nonnegative integers $\{k_j\}$ is defined by

$$k_0 = 0, \qquad k_{j+1} = \min\left\{ k > k_j \,\middle|\, \sum_{l=k_j}^{k-1} \gamma_k > T \right\},$$

for some $T > 0$. The sequence $\{k_j\}$ is chosen so that two successive elements in it are sufficiently apart for the difference in r_{k_j} and r_{k_j+1} to be non-trivial and informative. To obtain a relationship between r_{k_j+1} and r_{k_j}, for each j, define a sequence $\hat{r}_k^j$ by

$$\hat{r}_k^j = r_k / \max(1, |r_{k_j}|), \qquad \text{for} \quad k \geq k_j.$$

Note that $\hat{r}_k^j$ is $\mathcal{F}_k$-adapted and satisfies the iteration

$$\hat{r}_{k+1}^j = \hat{r}_k^j + \gamma_k \left(\frac{\bar{h}(\theta_k)}{\max(1, |r_{k_j}|)} - \bar{G}(\theta_k)\hat{r}_k^j \right) + \gamma_k \hat{\epsilon}_{k+1}^j, \qquad k \geq k_j,$$

where for $k \geq k_j$,

$$\hat{\epsilon}_{k+1}^j = \left(\frac{h_{\theta_k}(Y_k) - \bar{h}(\theta_k)}{\max(1, |r_{k_j}|)} - (G_{\theta_k}(Y_k) - \bar{G}(\theta_k))\hat{r}_k^j \right) + \xi_{k+1}\hat{r}_k^j,$$

can be viewed as perturbation noise. Similarly, for each j, define a sequence r_k^j by the iteration

$$\begin{aligned} r_{k_j}^j &= \hat{r}_{k_j}, \\ r_{k+1}^j &= r_k^j + \gamma_k \left(\frac{\bar{h}(\theta_k)}{\max(1, |r_{k_j}|)} - \bar{G}(\theta_k) r_k^j \right), \quad k \geq k_j, \end{aligned}$$

which is the same as that of $\hat{r}_k^j$ but without the perturbation noise. The relationship between r_{k_j+1} and r_{k_j} is obtained by showing that the perturbation noise is negligible and that $\hat{r}_k^j$ tracks r_k^j in the sense that

$$\lim_j \max_{k_j \le k \le k_{j+1}} |\hat{r}_k^j - r_k^j| = 0, \quad \text{w.p.1.}$$

To show this, we use the stopping times $\tau_j^{(1)}(C)$ and $\tau_j^{(2)}(\delta)$ defined as follows: for each $C > 0$ and $\delta > 0$, let

$$\begin{aligned} \tau_j^{(1)}(C) &= \min\{k \ge k_j : |\hat{r}_k^j| \ge C\}, \\ \tau_j^{(2)}(\delta) &= \min\{k \ge k_j : |\hat{r}_k^j - r_k^j| \ge \delta\}. \end{aligned}$$

Since $\bar{h}(\cdot), \bar{G}(\cdot)$ are bounded, using Assumption 3.1(10) and Lemma 3.4 it is easy to see that

$$\sup_j \max_{k_j \le k} |r_k^j| \le C$$

for some constant C. Therefore

$$\tau_j^{(1)}(C+\delta) \ge \tau_j^{(2)}(\delta), \quad \forall j, \quad \text{w.p.1.}$$

That is, by the time $\hat{r}_k^j$ gets out of the ball (around the origin) with radius $C+\delta$, $\hat{r}_k^j$ must have deviated from r_k^j by at least δ, since r_k^j lies completely inside the ball with radius C. Fix these constants and j, and, for convenience, drop these constants from the notation in the following subsection.

The following subsection derives bounds on the "effect" of the perturbation noise $\hat{\epsilon}_k^j$.

3.2.1 Bounds on the Perturbation Noise

By definition, we have

$$|\hat{r}_l| I\{l < \tau_j^{(1)}\} \le C, \ \forall j, k.$$

The following lemma shows that all the moments of

$$|\hat{r}_l| I\{l \le \tau_j^{(1)}\}$$

are bounded uniformly in j and l.

Lemma 3.5. *For each $d > 0$, there exists $C_d > 0$ such that*

$$\mathbf{E}\left[|\hat{r}_l|^d I\{l \le \tau_j^{(1)}\}\right] \le C_d, \quad \forall l > k_j.$$

Proof. It suffices to prove that $\mathbf{E}\left[|\hat{r}_l|^d I\{l = \tau_j^{(1)}\}\right]$ is bounded uniformly in l, j. For

each l, $\hat{r}_l$ is affine in $\hat{r}_{l-1}$ and if $\tau_j^{(1)} = l$ then $|\hat{r}_{l-1}| < C$. In view of Assumptions 3.1(4,7), this implies that $\hat{r}_l I\{\tau_j^{(1)} = l\}$ has bounded moments. More precisely, we have

$$\begin{aligned} I\{l = \tau_j^{(1)}\}|\hat{r}_l| &\leq I\{l = \tau_j^{(1)}\}|\hat{r}_{l-1}| + \gamma_{l-1} a_j \left|h_{\theta_{l-1}}(Y_{l-1})\right| \\ &\quad + \gamma_{l-1}\left(\left|G_{\theta_{l-1}}(Y_{l-1})\right| + |\xi_l|\right) I\{l = \tau_j^{(1)}\}|\hat{r}_{l-1}| \\ &\leq C\left(1 + \gamma_{l-1}|\xi_l| + \gamma_{l-1}\left|G_{\theta_{l-1}}(Y_{l-1})\right|\right) + \gamma_{l-1}\left|h_{\theta_{l-1}}(Y_{l-1})\right|. \end{aligned}$$

The rest follows from Assumptions 3.1(4,7) and Holder's inequality. □

We wish to show that if $\hat{r}_k$ is bounded, then the noise $\hat{\epsilon}_{k+1}$ is negligible in the sense that there exists a constant $C_1 > 0$ such that

$$\mathbf{E}\left[\max_{k_j \leq k \leq \tau_j^{(1)} \wedge k_{j+1}} \left|\sum_{l=k_j}^{k} \gamma_l \hat{\epsilon}_{l+1}\right|^2\right] \leq C_1 \sum_{l=k_j}^{k_{j+1}-1} \gamma_l^2, \tag{3.2}$$

Since $\left|\hat{r}_{k_j}\right| \leq 1$, Assumption 3.1(7) and Holder's inequality imply that

$$\mathbf{E}[|\gamma_{k_j}\hat{\epsilon}_{k_j+1}|^2] \leq C_1 \gamma_{k_j}^2,$$

for some $C_1 > 0$. Since the l.h.s. of (3.2) is less than

$$\mathbf{E}\left[\max_{k_j < k \leq \tau_j^{(1)} \wedge k_{j+1}} \left|\sum_{l=k_j}^{k} \gamma_l \hat{\epsilon}_{l+1}\right|^2\right] + \mathbf{E}[|\gamma_{k_j}\hat{\epsilon}_{k_j+1}|^2]$$

we can restrict the max operator to $k_j < k \leq \tau_j^{(1)} \wedge k_{j+1}$. Furthermore, since

$$\left|\sum_{l=k_j}^{k} \gamma_l \hat{\epsilon}_{l+1}\right|^2 \leq 2|\gamma_{k_j}\hat{\epsilon}_{k_j+1}|^2 + 2\left|\sum_{l=k_j+1}^{k} \gamma_l \hat{\epsilon}_{l+1}\right|,^2$$

we can concentrate only on deriving an upper bound for

$$\mathbf{E}\left[\max_{k_j < k \leq \tau_j^{(1)} \wedge k_{j+1}} \left|\sum_{l=k_j+1}^{k} \gamma_l \hat{\epsilon}_{l+1}\right|^2\right].$$

The next step is to decompose the noise $\hat{\epsilon}_k$ into parts that are easy to handle. For $r \in \mathbb{R}^n$, let

$$\begin{aligned} F_\theta(r; y) &= a_j h_\theta(y) - G_\theta(y) r, \\ \hat{F}_\theta(r; y) &= a_j \hat{h}_\theta(y) - \hat{G}_\theta(y) r, \\ \bar{F}_\theta(r) &= a_j \bar{h}(\theta) - \bar{G}(\theta) r. \end{aligned}$$

It follows from Assumption 3.1(5) that $\hat{F}_\theta(r;\cdot)$ satisfies the Poisson equation

$$\hat{F}_\theta(r;y) = F_\theta(r;y) - \bar{F}_\theta(r) + (P_\theta \hat{F}_\theta)(r;y).$$

Therefore, for $k > k_j$ we have

$$\begin{aligned}
\hat{\epsilon}_{k+1} &= \xi_{k+1}\hat{r}_k + F_{\theta_k}(\hat{r}_k;Y_k) - \bar{F}_{\theta_k}(\hat{r}_k) \\
&= \xi_{k+1}\hat{r}_k + (\hat{F}_{\theta_k}(\hat{r}_k;Y_k) - (P_{\theta_k}\hat{F}_{\theta_k})(\hat{r}_k;Y_k)) \\
&= \xi_{k+1}\hat{r}_k + (\hat{F}_{\theta_k}(\hat{r}_k;Y_{k+1}) - (P_{\theta_k}\hat{F}_{\theta_k})(\hat{r}_k;Y_k)) \\
&\quad +(\hat{F}_{\theta_{k-1}}(\hat{r}_{k-1};Y_k) - \hat{F}_{\theta_k}(\hat{r}_k;Y_{k+1})) \\
&\quad +(\hat{F}_{\theta_k}(\hat{r}_k;Y_k) - \hat{F}_{\theta_k}(\hat{r}_{k-1};Y_k)) \\
&\quad +(\hat{F}_{\theta_k}(\hat{r}_{k-1};Y_k) - \hat{F}_{\theta_{k-1}}(\hat{r}_{k-1};Y_k)).
\end{aligned}$$

Let

$$\begin{aligned}
\hat{\epsilon}^{(1)}_{k+1} &= \xi_{k+1}\hat{r}_k + \hat{F}_{\theta_k}(\hat{r}_k;Y_{k+1}) - (P_{\theta_k}\hat{F}_{\theta_k})(\hat{r}_k;Y_k), \\
\hat{\epsilon}^{(2)}_{k+1} &= \frac{\gamma_{k-1}\hat{F}_{\theta_{k-1}}(\hat{r}_{k-1};Y_k) - \gamma_k\hat{F}_{\theta_k}(\hat{r}_k;Y_{k+1})}{\gamma_k}, \\
\hat{\epsilon}^{(3)}_{k+1} &= \frac{(\gamma_k - \gamma_{k-1})}{\gamma_k}\hat{F}_{\theta_{k-1}}(\hat{r}_{k-1};Y_k), \\
\hat{\epsilon}^{(4)}_{k+1} &= \hat{F}_{\theta_k}(\hat{r}_k;Y_k) - \hat{F}_{\theta_k}(\hat{r}_{k-1};Y_k), \\
\hat{\epsilon}^{(5)}_{k+1} &= \hat{F}_{\theta_k}(\hat{r}_{k-1};Y_k) - \hat{F}_{\theta_{k-1}}(\hat{r}_{k-1};Y_k).
\end{aligned}$$

Then

$$\hat{\epsilon}_{k+1} = \hat{\epsilon}^{(1)}_{k+1} + \hat{\epsilon}^{(2)}_{k+1} + \hat{\epsilon}^{(3)}_{k+1} + \hat{\epsilon}^{(4)}_{k+1} + \hat{\epsilon}^{(5)}_{k+1}.$$

The following lemma derives several bounds that will be used later.

Lemma 3.6. *For each $d > 0$, there exists a constant C_d such that for all $l \geq k > k_j$ the following inequalities hold.*

1. $\mathbf{E}\left[I\{l \leq \tau_j^{(1)}\}\,|F_{\theta_l}(\hat{r}_k, Y_l)|^d\right] \leq C_d.$
2. $\mathbf{E}\left[I\{l \leq \tau_j^{(1)}\}\left|\hat{F}_{\theta_l}(\hat{r}_k, Y_l)\right|^d\right] \leq C_d.$
3. $\mathbf{E}\left[I\{l \leq \tau_j^{(1)}\}\left|\hat{F}_{\theta_l}(\hat{r}_k, Y_{l+1})\right|^d\right] \leq C_d.$
4. $\mathbf{E}\left[I\{l \leq \tau_j^{(1)}\}\left|P_{\theta_l}\hat{F}_{\theta_l}(\hat{r}_l, Y_l)\right|^d\right] \leq C_d.$

Proof. Consider the first inequality. Since $a_j \leq 1$, we have for $d > 1$,

$$\begin{aligned} |F_{\theta_l}(\hat{r}_k, Y_l)|^d &= |a_j h_{\theta_l}(Y_l) + G_{\theta_l}(Y_l)\hat{r}_k|^d \\ &\leq 2^{d-1} |h_{\theta_l}(Y_l)|^d + 2^{d-1} |G_{\theta_l}(Y_l)|^d |\hat{r}_k|^d. \end{aligned}$$

The first inequality follows from Assumption 3.1(4,7), Lemma 3.5 and the fact that $k \leq l$. The proof of the second inequality is similar. To prove the third inequality, note that Assumption 3.1(8) implies that for $l > k_j$,

$$\begin{aligned} \left|\hat{F}_{\theta_l}(\hat{r}_k; Y_{l+1})\right| &\leq \left|\hat{F}_{\theta_{l+1}}(\hat{r}_k; Y_{l+1})\right| + \left|\hat{F}_{\theta_{l+1}}(\hat{r}_k; Y_l) - \hat{F}_{\theta_l}(\hat{r}_k; Y_l)\right| \\ &\leq \left|\hat{F}_{\theta_l}(\hat{r}_k; Y_l)\right| + (1 + |\hat{r}_k|)C(Y_l) \\ &\leq \left|\hat{h}_{\theta_l}(Y_l)\right| + (1 + |\hat{r}_k|)\left(|C(Y_l)| + |G_{\theta_l}(Y_l)|\right). \end{aligned}$$

Therefore, the third inequality follows from the fact that

$$I\{l \leq \tau_j^{(1)}\}|\hat{r}_k| \leq I\{k \leq \tau_j^{(1)}\}|\hat{r}_k|,$$

Holder's inequality, Lemma 3.5 and Assumption 3.1(4,7). To prove the fourth inequality, note that

$$P_{\theta_l}\hat{F}_{\theta_l}(\hat{r}_l, Y_l) = \hat{F}_{\theta_l}(\hat{r}_l, Y_l) - F_{\theta_l}(\hat{r}_l, Y_l) - \bar{F}_{\theta_l}(\hat{r}_l).$$

The inequality follows from the first two inequalities and the fact that (cf. Assumption 3.1(6))

$$\left|\bar{F}_{\theta_l}(\hat{r}_l)\right| \leq C_2(1 + |\hat{r}_l|)$$

for some $C_2 > 0$. □

In the following lemmas, the inequality (3.2) is established by showing that it holds when $\hat{\epsilon}_{k+1}$ is replaced by any of the components $\hat{\epsilon}_{k+1}^{(i)}, i = 1, \ldots, 5$. For two integers j, k, let $j \wedge k$ denote the minimum of j and k.

Lemma 3.7. *There exists $C_1 > 0$ such that for $i = 1, \ldots, 5$,*

$$\mathbf{E}\left[\max_{k_j < k \leq \tau_j^{(1)} \wedge k_{j+1}} \left|\sum_{l=k_j+1}^{k} \gamma_l \hat{\epsilon}_{l+1}^{(i)}\right|^2\right] \leq C_1 \sum_{k=k_j+1}^{k_{j+1}} \gamma_k^2.$$

Proof. The following are the proofs corresponding to each of the components $\hat{\epsilon}_{k+1}^{(i)}, i = 1, \ldots, 5$, presented separately.

1. Note that

$$\sum_{l=k_j+1}^{k\wedge\tau_j^{(1)}} \gamma_l \hat{\epsilon}_{l+1}^{(1)}$$

is a martingale. Therefore, Doob's inequality[1] can be used to see that

$$\begin{aligned}\mathbf{E}\left[\max_{k_j<k\le k_{j+1}}\left|\sum_{l=k_j+1}^{k\wedge\tau_j^{(1)}}\gamma_l\hat{\epsilon}_{l+1}^{(1)}\right|^2\right] &\le 4\mathbf{E}\left[\left|\sum_{l=k_j+1}^{k_{j+1}\wedge\tau_j^{(1)}}\gamma_l\hat{\epsilon}_{l+1}^{(1)}\right|^2\right] \\ &= 4\sum_{l=k_j+1}^{k_{j+1}}\gamma_l^2\mathbf{E}\left[I\{l\le\tau_j^{(1)}\}\left|\hat{\epsilon}_{l+1}^{(1)}\right|^2\right].\end{aligned}$$

The rest follows from Holder's inequality, Assumptions 3.1 (4) and Lemma 3.6 (1) (2) and (4).

2. By definition, we have

$$\begin{aligned}&\mathbf{E}\left[\max_{k_j<k\le\tau_j^{(1)}\wedge k_{j+1}}\left|\sum_{l=k_j+1}^{k}\gamma_l\hat{\epsilon}_{l+1}^{(2)}\right|^2\right] \\ &\le \mathbf{E}\left[\max_{k_j<k\le\tau_j^{(1)}\wedge k_{j+1}}\left|\sum_{l=k_j+1}^{k}\left(\gamma_{k-1}\hat{F}_{\theta_{k-1}}(\hat{r}_{k-1};Y_k)-\gamma_k\hat{F}_{\theta_k}(\hat{r}_k;Y_{k+1})\right)\right|^2\right] \\ &\le \mathbf{E}\left[\max_{k_j<k\le\tau_j^{(1)}\wedge k_{j+1}}\left|\gamma_k\hat{F}_{\theta_k}(\hat{r}_k;Y_{k+1})\right|^2\right]+\gamma_{k_j}^2\mathbf{E}\left[\left|\hat{F}_{\theta_{k_j}}(\hat{r}_{k_j};Y_{k_j+1})\right|^2\right] \\ &\le \mathbf{E}\left[\sum_{k=k_j+1}^{\tau_j^{(1)}\wedge k_{j+1}}\gamma_k^2|\hat{F}_{\theta_k}(\hat{r}_k;Y_{k+1})|^2\right]+C\gamma_{k_j}^2 \\ &\le \sum_{k=k_j+1}^{k_{j+1}}\gamma_k^2\mathbf{E}\left[I\{k\le\tau_j^{(1)}\}|\hat{F}_{\theta_k}(\hat{r}_k;Y_{k+1})|^2\right]+C\gamma_{k_j}^2 \\ &\le C\sum_{k=k_j}^{k_{j+1}}\gamma_k^2.\end{aligned}$$

[1]Doob's inequality states that for any nonnegative submartingale $\{S_k\}$ and $p>1$,

$$\mathbf{E}\left[\max_{l\le k}S_l^p\right]\le\left(\frac{p}{p-1}\right)^p\mathbf{E}\left[S_k^p\right].$$

Note that if $\{X_k\}$ is a martingale, then $|X_k|$ is a nonnegative submartingale.

The third of the above inequalities follows from the fact that

$$\mathbf{E}\left[\left|\hat{F}_{\theta_{k_j}}(\hat{r}_{k_j}; Y_{k_j+1})\right|^2\right]$$

is uniformly bounded in j (cf. Assumption 3.1 (7)). The fourth follows from Lemma 3.6 (3).

3. Similarly, for $i = 3$, we have

$$\begin{aligned}
&\mathbf{E}\left[\max_{k_j<k\le\tau_j^{(1)}\wedge k_{j+1}}\left|\sum_{l=k_j+1}^{k}\gamma_l\hat{\epsilon}_{l+1}^{(3)}\right|^2\right] \\
&\le \mathbf{E}\left[\left(\sum_{l=k_j+1}^{k_{j+1}\wedge\tau_j^{(1)}}\gamma_l\left|\hat{\epsilon}_{l+1}^{(3)}\right|\right)^2\right] \\
&= \mathbf{E}\left[\left(\sum_{l=k_j+1}^{k_{j+1}\wedge\tau_j^{(1)}}(\gamma_{l-1}-\gamma_l)\left|\hat{F}_{\theta_{l-1}}(\hat{r}_{l-1}; Y_l)\right|\right)^2\right] \\
&\le \left(\gamma_{k_j}-\gamma_{k_{j+1}}\right)\mathbf{E}\left[\sum_{l=k_j+1}^{k_{j+1}\wedge\tau_j^{(1)}}(\gamma_{l-1}-\gamma_l)\left|\hat{F}_{\theta_{l-1}}(\hat{r}_{l-1}; Y_l)\right|^2\right] \\
&\le \left(\gamma_{k_j}-\gamma_{k_{j+1}}\right)^2\sup_{k>k_j}\mathbf{E}\left[\left|\hat{F}_{\theta_{k-1}}(\hat{r}_{k-1}; Y_k)\right|^2 I\{k\le\tau_j^{(1)}\}\right] \\
&\le C_1\left(\gamma_{k_j}-\gamma_{k_{j+1}}\right)^2,
\end{aligned}$$

where the inequalities follow from the fact that $\{\gamma_k\}$ is non-increasing, the Schwartz inequality and Lemma 3.6 (3).

4. By definition, we have for $k > k_j$,

$$\begin{aligned}
\left|\hat{\epsilon}_{k+1}^{(4)}I\{k\le\tau_j^{(1)}\}\right| &= I\{k\le\tau_j^{(1)}\}\left|(\hat{r}_k-\hat{r}_{k-1})\hat{G}_{\theta_k}(Y_k)\right| \\
&\le I\{k\le\tau_j^{(1)}\}\gamma_{k-1}\left|\hat{r}_{k-1}\right|\left|\xi_k-G_{\theta_{k-1}}(Y_{k-1})\right|\left|\hat{G}_{\theta_k}(Y_k)\right| \\
&\quad +I\{k\le\tau_j^{(1)}\}\gamma_{k-1}\left|h_{\theta_{k-1}}(Y_{k-1})\right|\left|\hat{G}_{\theta_k}(Y_k)\right| \\
&\le \gamma_{k-1}C\left(\left|h_{\theta_{k-1}}(Y_{k-1})\right|+\left|G_{\theta_{k-1}}(Y_{k-1})\right|+|\xi_k|\right)\left|\hat{G}_{\theta_k}(Y_k)\right|.
\end{aligned}$$

From Assumption 3.1 (4,7) and Holder's inequality, it follows that

$$\mathbf{E}[|\hat{\epsilon}_{k+1}^{(4)}|^2 I\{k\le\tau_j^{(1)}\}]\le C_1\gamma_{k-1}^2$$

for some $C_1 > 0$. Therefore, using Schwartz inequality, we have

$$\begin{aligned}\mathbf{E}\left[\max_{k_j<k\le\tau_j^{(1)}\wedge k_{j+1}}\left|\sum_{l=k_j+1}^{k}\gamma_l\hat{\epsilon}_{l+1}^{(4)}\right|^2\right] &\le T\mathbf{E}\left[\left(\sum_{l=k_j+1}^{k_{j+1}}\left|\sqrt{\gamma_l}\hat{\epsilon}_{l+1}^{(4)}I\{k\le\tau_j^{(1)}\}\right|^2\right)\right]\\ &\le C_2\sum_{l=k_j+1}^{k_{j+1}}\gamma_l^2,\end{aligned}$$

for some $C_2 > 0$.

5. For the case $i = 5$, note that Assumption 3.1(9) implies that

$$\left|\hat{\epsilon}_{k+1}^{(5)}\right| \le |\theta_k - \theta_{k-1}|(1 + |\hat{r}_k|)C(Y_k).$$

Therefore, from Assumption 3.1(3) and Holder's inequality we have

$$\mathbf{E}[|\hat{\epsilon}_{k+1}^{(5)}|^2 I\{k \le \tau_j^{(1)}\}] \le C_1\beta_k^2.$$

From Assumption 3.1(3) we have $\beta_k \le \gamma_k$ for large enough k. The rest is similar to the previous case.

□

3.2.2 Proof of Boundedness

The previous lemma says that as long as $\hat{r}_k$ is bounded, the perturbation noise remains negligible. In the next lemma, we prove that the sequence $\{\hat{r}_k^j\}$ closely approximates r_k^j.

Lemma 3.8. $\lim_j \max_{k_j\le k\le k_{j+1}} |\hat{r}_k^j - r_k^j| = 0, \quad w.p.1.$

Proof. Note that, since $\bar{G}(\cdot)$ is bounded, for $k \ge k_j$,

$$|\hat{r}_{k+1}^j - r_{k+1}^j| \le C\sum_{l=k_j}^{k}\gamma_l|\hat{r}_l^j - r_l^j| + \left|\sum_{l=k_j}^{k}\gamma_l\hat{\epsilon}_{l+1}\right|.$$

Using the discrete Gronwall inequality[2], it is easy to see that

$$\max_{k_j\le k\le k_{j+1}\wedge\tau_j^{(1)}} |\hat{r}_{k+1}^j - r_{k+1}^j| \le e^{CT}\max_{k_j\le k\le k_{j+1}\wedge\tau_j^{(1)}}\left|\sum_{l=k_j}^{k}\gamma_l\hat{\epsilon}_{l+1}\right|.$$

[2]For a non-negative sequence $\{\gamma_k\}$ and a constant B, let $\{b_k\}$ be a sequence satisfying:

$$b_{k+1} \le \sum_{l=0}^{k} b_k\gamma_k + B, \quad \forall k.$$

Therefore, the above lemma along with the Chebyshev inequality implies that

$$\mathbf{P}\left(\max_{k_j\le k\le k_{j+1}\wedge\tau_j^{(1)}}|\hat{r}_{k+1}^j - r_{k+1}^j|\ge\delta\right)\le\frac{C_1}{\delta^2}\sum_{l=k_j}^{k_{j+1}-1}\gamma_l^2$$

for some $C_1 > 0$. In the above expression, the probability on the left hand side is exactly $\mathbf{P}(\tau_j^{(2)} \le k_{j+1}\wedge\tau_j^{(1)})$ which is the same as $\mathbf{P}(\tau_j^{(2)}\le k_{j+1})$, since $\tau_j^{(1)}\ge\tau_j^{(2)}$. Therefore

$$\mathbf{P}\left(\max_{k_j\le k\le k_{j+1}}|\hat{r}_{k+1}^j - r_{k+1}^j|\ge\delta\right)\le\frac{C_1}{\delta^2}\sum_{l=k_j}^{k_{j+1}-1}\gamma_l^2.$$

The rest follows from the summability of the series $\sum_k\gamma_k^2$ and the Borel-Cantelli Lemma. □

Lemma 3.9. $\sup_k|r_k| < \infty$, *w.p.1.*

Proof. Since $\bar{h}(\cdot)$ is bounded, Assumption 3.1(10) and Lemma 3.4 imply the following: for $k \ge k_j$ and j sufficiently large,

$$|r_{k+1}^j|\le\left(1-\frac{1}{2}\gamma_k a\right)|r_k^j| + \gamma_k\frac{C}{\max(1,|r_{k_j}|)}.$$

Using the inequality $1 - x \le e^{-x}$ we have

$$|r_{k+1}^j|\le e^{-\left(\frac{1}{2}a\sum_{l=k_j}^k\gamma_l\right)}|r_{k_j}^j| + \left(\sum_{l=k_j}^k\gamma_l\right)\frac{C}{\max(1,|r_{k_j}|)}.$$

This, along with the previous Lemma 3.8, implies

$$\frac{|r_{k_{j+1}}|}{\max(1,|r_{k_j}|)}\le e^{-\frac{1}{2}aT}\frac{|r_{k_j}|}{\max(1,|r_{k_j}|)} + T\frac{C}{\max(1,|r_{k_j}|)} + \delta_j$$

where $\delta_j \to 0$ w.p.1. Multiplying both sides by $\max(1,|r_{k_j}|)$ and using the fact that this is less than $(1+|r_{k_j}|)$, we have

$$|r_{k_{j+1}}|\le(e^{-\frac{1}{2}aT}+\delta_j)|r_{k_j}| + CT + \delta_j.$$

Since $e^{-aT} < 1$ and $\delta_j \to 0$ w.p.1., it follows from Lemma 3.3 that $\sup_j|r_{k_j}| < \infty$,

---

Then, for every k, we have

$$b_k \le Be^{\sum_{l=0}^k\gamma_l}.$$

w.p.1. The rest follows from the observation that

$$\begin{aligned}\sup_k |r_k| &= \sup_j \max(1, |r_{k_j}|) \max_{k_j \le k \le k_{j+1}} |\hat{r}_k^j| \\ &\le \sup_j \left\{ (1 + |r_{k_j}|) \left(\max_{k_j \le k \le k_{j+1}} |r_k^j| + \max_{k_j \le k \le k_{j+1}} |r_k^j - \hat{r}_k^j| \right) \right\},\end{aligned}$$

the boundedness of $\{r_k^j\}$, and the previous lemma. □

3.3 Proof of Theorem 3.2

To prove Theorem 3.2, consider the sequence $\hat{\rho}_k = \bar{G}(\theta_k) r_k - \bar{h}(\theta_k)$. This sequence evolves according to the iteration:

$$\hat{\rho}_{k+1} = \hat{\rho}_k - \gamma_k \bar{G}(\theta_{k+1}) \hat{\rho}_k + \gamma_k \epsilon_{k+1}^{(1)} + \gamma_k \epsilon_{k+1}^{(2)}.$$

where

$$\begin{aligned}\epsilon_{k+1}^{(1)} &= \bar{G}(\theta_{k+1})(h_{\theta_k}(Y_k) - \bar{h}(\theta_k)) - \bar{G}(\theta_{k+1})(G_{\theta_k}(Y_k) - \bar{G}(\theta_k)) r_k \\ &\quad + \bar{G}(\theta_{k+1}) \xi_{k+1} r_k, \\ \epsilon_{k+1}^{(2)} &= \tfrac{1}{\gamma_k}((\bar{G}(\theta_{k+1}) - \bar{G}(\theta_k)) r_k - (\bar{h}(\theta_{k+1}) - \bar{h}(\theta_k))).\end{aligned}$$

Lemma 3.10. $\sum_k \gamma_k \epsilon_{k+1}^{(1)}$ *converges w.p.1.*

Proof. A part of this proof is similar to arguments of Section 3.2 and therefore this part will only be outlined. Define a sequence of stopping times $\{\tau_j\}$:

$$\tau_j = \min\{k : |r_k| \ge j\}.$$

For each j, the stopped process $\sum_{l=0}^{k \wedge \tau_j} \gamma_l \epsilon_{l+1}^{(1)}$ can be decomposed, as in Section 3.2, into several components (say $\sum_{l=0}^{k \wedge \tau_j} \gamma_l \epsilon_{l+1}^{(i)}, i > 2$). Some of these components are martingales with bounded second moments and therefore converge w.p.1. By calculating the expectation of $\sum_{l=0}^{k \wedge \tau_j} \gamma_l |\epsilon_{l+1}^{(i)}$— for the remaining components, one can easily see that they are absolutely convergent w.p.1. Therefore, the stopped process converges w.p.1. This implies that $\sum_k \gamma_k \epsilon_{k+1}^{(1)}$ converges on the set of outcomes for which $\tau_j = \infty$. The boundedness of $\{r_k\}$ implies that w.p.1., $\tau_j = \infty$ for some j, and thus the result follows. □

Lemma 3.11. $\lim_k \epsilon_k^{(2)} = 0$, *w.p.1.*

Proof. Assumptions 3.1(3,8) imply that

$$|\epsilon_{k+1}^{(2)}| \le \frac{\beta_k}{\gamma_k} C(1 + |r_k|) H_k.$$

Since $\{H_k\}$ has bounded moments, we have

$$E\left[\sum_k \left(\frac{\beta_k}{\gamma_k}\right)^d H_k^d\right] < \infty,$$

for some $d > 0$. Therefore, $(\beta_k/\gamma_k)H_k$ converges to zero, w.p.1. The rest follows from the boundedness of $\{r_k\}$. □

Recall the notation k_j from the previous section. For each j, define ρ_k^j, for $k \geq k_j$ by:

$$\rho_{k+1}^j = (I - \gamma_k \bar{G}(\theta_k))\rho_k^j \qquad \rho_{k_j}^j = \rho_{k_j}.$$

Lemma 3.12. $\lim_j \max_{k_j \leq k \leq k_{j+1}} |\hat{\rho}_k - \rho_k^j| = 0$, *w.p.1.*

Proof. For each j, $k \geq k_j$,

$$|\hat{\rho}_k - \rho_k^j| \leq C \sum_{l=k_j}^{k-1} \gamma_l |\hat{\rho}_l - \rho_l^j| + \left| \sum_{l=k_j}^{k-1} \gamma_l (\epsilon_{l+1}^{(1)} + \epsilon_{l+1}^{(2)}) \right|$$

Using the discrete Gronwall inequality, it is easy to see that

$$\begin{aligned} \max_{k_j \leq k \leq k_{j+1}} |\hat{\rho}_k - \rho_k^j| &\leq e^{CT} \max_{k_j \leq k \leq k_{j+1}} \left| \sum_{l=k_j}^{k-1} \gamma_l (\epsilon_{l+1}^{(1)} + \epsilon_{l+1}^{(2)}) \right|, \\ &\leq e^{CT} \sup_{k \geq k_j} \left| \sum_{l=k_j}^{k-1} \gamma_l \epsilon_{l+1}^{(1)} \right| + e^{CT} T \sup_{k \geq k_j} |\epsilon_{k+1}^{(2)}|. \end{aligned}$$

The rest follows from the previous two lemmas. □

Theorem 3.13. $\lim_k |\bar{G}(\theta_k) r_k - \bar{h}(\theta_k)| = 0, \qquad$ *w.p.1.*

Proof. Using Lemma 3.4 and Assumption 3.1(10) we have

$$|\rho_{k_{j+1}}^j| \leq e^{-\frac{1}{2}aT} |\rho_{k_j}^j|.$$

Therefore the above lemma implies

$$|\hat{\rho}_{k_{j+1}}| \leq e^{-\frac{a}{2}T} |\hat{\rho}_{k_j}| + \delta_j,$$

where $\delta_j \to 0$ w.p.1. The rest follows from Lemma 3.3(2) and arguments similar to the closing arguments of the proof of Lemma 3.9. □

3.4 Closing Remarks

The general result presented in this chapter is new. Tracking results are not new in stochastic approximation literature (Benveniste *et al.*, 1990). However, they are limited to the tracking ability of constant step-size algorithms and place restrictive

assumptions on the dynamics of the changing parameter. Since the classical literature on adaptive algorithms concerns adaptation to a changing environment, the tracking ability of algorithms with decreasing step-sizes has not received much attention. Our result is not intended to show that linear algorithms with decreasing step-sizes can be used as adaptive algorithms. In fact, our assumption that the parameter θ changes slowly is quite strong and is not satisfied for most environments. However, the results such as the one in this chapter are useful for designing algorithms as we will see in the next two chapters. In the next chapter, we use this to prove that critic tracks the actor's policy when the critic is updated faster than the actor. Later, we use these results to design several variants of actor-critic algorithms.

Chapter 4

The Critic

In this chapter, we study the critic part of actor-critic algorithms. The actor has a tunable parameter and at each time instant, takes a state (say x) as input and generates a decision using RSP corresponding to its parameter. The role of the critic in these algorithms is to evaluate the policies of the actor, which means estimating the information that the actor can use to improve its current policy. In particular, when the actor parameter θ is fixed, the critic acts as a function approximation scheme to approximate a solution Q_θ of the Poisson equation introduced in Chapter 2 or a corresponding function V_θ defined as

$$V_\theta(x) = \mathbf{E}_{\theta,x}[Q_\theta(x, U_0)].$$

The approximation of such a Q_θ or V_θ, in view of the gradient formulas of Chapter 2, can be used to update the actor parameter in an approximate gradient direction of the overall reward. A function Q_θ that satisfies the Poisson equation or the corresponding V_θ can be thought of as the evaluation of the policy θ. Therefore, such functions are often called either the evaluation or the value functions. Since Q_θ depends on both the state and decision, it is called a state-decision value function or a Q-value function corresponding to the policy θ whereas V_θ is called either a state value function or simply a value function of the policy θ.

The critic in our algorithms uses TD learning (Sutton, 1988; Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998) with linear function approximation. We consider several variants of TD learning for different criteria and study their convergence behavior. The existing results on convergence of TD (Tsitsiklis & Van Roy, 1997) do not apply to the way TD is used in the critic, as the policy under evaluation changes continuously. An additional difference with the common usage of TD is that the critic in some of the algorithms estimate the state-decision value functions rather than state value functions, and therefore use functions of both state and decision as basis for linear function approximation[1]. Furthermore, the basis functions depend on the policy under evaluation.

[1]TD algorithms that use functions of both state and decisions as basis are known as SARSA in the AI literature (Sutton & Barto, 1998).

A function depending only on the state can be extended trivially to be a function of both state and decision. Therefore, we describe only the critics with basis functions depending both on state and action because the critic with basis functions depending only on state can be seen as a special case of the former. The critics for different objective criteria are presented separately in the following sections. The assumptions and the convergence results are also stated in these sections and proved later in the section on convergence analysis. The main focus is on the average reward problem and the convergence analysis is presented only for this problem. The convergence analysis of the algorithms for other problems is very similar.

To describe our actor-critic algorithms, we use the following notation and conventions throughout the thesis. The parameter vector of the critic is denoted by r and the corresponding approximation to the state-decision value function under policy θ is given by

$$r'\phi_\theta(x,u) = \sum_{j=1}^{m} r^j \phi_\theta^j(x,u),$$

where $r = (r^1, \ldots, r^m) \in \mathbb{R}^m$. The functions ϕ_θ^j, $j = 1, \ldots, m$, are the basis functions of the critic. For each state-decision pair (x,u) and policy parameter θ, $\phi_\theta(x,u) = (\phi_\theta^1(x,u), \ldots, \phi_\theta^m(x,u))$ is called the feature vector of the state-decision pair (x,u) corresponding to policy θ.

Recall that the input to our actor-critic algorithms is a simulator of the system. The actor parameter is updated from time to time and its value at time k is denoted by θ_k. Similarly, the critic parameter is also updated from time to time and its value at time k is denoted by r_k. These parameters are updated in a direction depending on the simulated state transitions and simulated decisions. Let $\hat{X}_k, \hat{U}_k$ denote the simulated state and decision at time k. Suppose, at time k, $\hat{X}_k, \hat{U}_k, r_k, \theta_k$ are all known. The general scheme of our actor-critic algorithms is as follows:

1. The next state $\hat{X}_{k+1}$ is generated using the current state $\hat{X}_k$ and current decision $\hat{U}_k$.

2. The next decision $\hat{U}_{k+1}$ is generated by using policy θ_k on state $\hat{X}_{k+1}$.

3. The critic parameter is updated to r_{k+1} based on the observed transition from $(\hat{X}_k, \hat{U}_k)$ to $(\hat{X}_{k+1}, \hat{U}_{k+1})$. The exact form of this update, which is the subject of this chapter, depends on the objective criterion .

4. The actor parameter θ_k is also updated.

Step 2 is common to all our algorithms. The details of steps 1 and 3 vary from criterion to criterion, and are discussed in the following sections. The following sections also describe convergence results for the sequence of critic parameter values r_k, and the assumptions on the critic update, basis functions ϕ_θ^j's and the sequence of actor parameter values $\{\theta_k\}$.

Finally note that the sequence of state-decision pairs $\{(\hat{X}_k, \hat{U}_k)\}$ generated during the course of the algorithm are random variables. Since the updates of all the parameters are based on these states and decisions, these parameters are also random variables. Let $\mathbf{P}$ denote the probability law of the stochastic process consisting of all the random variables generated during the course of the algorithm. Let $\mathbf{E}$ denote the corresponding expectation.

4.1 Average Reward

In this section, we describe the critic part of our actor-critic algorithms for the average reward criterion. In Step 1 of the scheme described earlier, the state $\hat{X}_{k+1}$ is the output generated by the simulator of the transition of the given MDP, when the state-decision pair $(\hat{X}_k, \hat{U}_k)$ is given as input. The critic updates its parameter using the following auxiliary parameters:

1. a scalar estimate α of the average cost,
2. an m-vector $\hat{Z}$ which represents Sutton's eligibility trace (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998).

Let α_k and $\hat{Z}_k$ denote the estimate of average reward and eligibility traces at time k. At each time step k, the critic carries out an update similar to the average cost temporal-difference method of (Tsitsiklis & Van Roy, 1999a):

$$\begin{aligned} \alpha_{k+1} &= \alpha_k + \gamma_k(g(\hat{X}_{k+1}, \hat{U}_{k+1}) - \alpha_k), \\ r_{k+1} &= r_k + \gamma_k d_k \hat{Z}_k, \end{aligned} \tag{4.1}$$

where

$$d_k = g(\hat{X}_k, \hat{U}_k) - \alpha_k + r_k' \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) - r_k' \phi_{\theta_k}(\hat{X}_k, \hat{U}_k) \tag{4.2}$$

and γ_k is the positive step-size parameter. The two variants of the critic for the average reward criterion update their eligibility traces $\hat{Z}_k$ in different ways:

TD(1) Critic

$$\hat{Z}_{k+1} = \begin{cases} \hat{Z}_k + \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}), & \text{if } \hat{X}_{k+1} \neq x^*, \\ \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}), & \text{otherwise,} \end{cases} \tag{4.3}$$

where x^* is a fixed state.

TD(λ) Critic, $0 < \lambda < 1$

$$\hat{Z}_{k+1} = \lambda \hat{Z}_k + \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}).$$

The following subsection describes the convergence results on the above variants of the critic.

4.1.1 Convergence Results

Before presenting the convergence results, we state the assumptions on the feature vectors ϕ_θ, the step sizes γ_k, and the actor parameters θ_k. The assumptions common to both the variants are described first.

The first assumption ensures that the feature vector $\phi_\theta = (\phi_\theta^1, \ldots, \phi_\theta^m)$ as a function of the policy parameter θ is "well behaved". This is assumed to hold for all variants of algorithms and for all overall reward criteria. However, the bounding functions L are different for different reward criteria and are as defined in Chapter 2. This assumption is not the only one that we make on the basis functions of the critic. Whenever needed, the choice of the basis functions ϕ_θ^j will be further restricted.

Assumption 4.1. *1. There exists $K > 0$ such that*

$$|\phi_\theta(x,u)| \leq KL(x,u), \qquad \forall \theta \in \mathbb{R}^n, x \in \mathbb{X}, u \in \mathbb{U}.$$

2. There exists $K > 0$ such that

$$|\phi_\theta(x,u) - \phi_{\bar{\theta}}(x,u)| \leq K|\theta - \bar{\theta}|L(x,u). \qquad \forall \bar{\theta}, \theta \in \mathbb{R}^n, x \in \mathbb{X}, u \in \mathbb{U}.$$

One of the crucial ingredients of our algorithms is that the parameter of the actor "changes slowly" compared to that of the critic. In other words, the step-sizes γ_k and the sequence of actor parameters θ_k are assumed to satisfy the following:

Assumption 4.2.

1. γ_k is deterministic, non-increasing and satisfies

$$\sum_k \gamma_k = \infty, \qquad \sum_k \gamma_k^2 < \infty.$$

2. The (random) sequence of parameters $\{\theta_k\}$ satisfies:

$$|\theta_{k+1} - \theta_k| \leq \beta_k H_k,$$

for some nonnegative process $\{H_k\}$ with bounded moments and deterministic sequence $\{\beta_k\}$ such that

$$\sum_k \left(\frac{\beta_k}{\gamma_k}\right)^d < \infty$$

for some $d > 0$.

Since the actor's policy in our algorithms changes continuously, the notion of convergence of the critic in our setting needs to be clarified. Suppose the actor's parameter vector is fixed at θ. It is well known that the critic's parameter in the variants we present here converges to the solution of a linear equation:

$$\bar{h}(\theta) = \bar{G}(\theta)r,$$

where $\bar{h}(\cdot)$ and $\bar{G}(\cdot)$ depend on the particular variant. In our context, the convergence of the critic means that the following holds:

$$\lim_k |r_k - \bar{r}(\theta_k)| = 0,$$

where

$$\bar{r}(\theta) = \bar{G}(\theta)^{-1}\bar{h}(\theta).$$

Before we move on, we need a new concept to state our additional assumptions on the features for each variant. For $\bar{G}(\cdot)$ to be non-singular for each θ, we require that the basis functions $\phi_\theta^i, i = 1, \ldots, m$, be "non-redundant". To describe the corresponding property of basis functions in our context, we need the following definition. Recall the notation for the inner product and the norm on the space $\mathcal{L}_\theta^2$ defined in the Chapter 2. Note that Assumptions 4.1 and 2.7 imply that $\phi_\theta^i \in \mathcal{L}_\theta^2$ for each $\theta \in \mathbb{R}^n$ and $i = 1, \ldots, m$.

Definition 4.3. A parameterized family of basis functions $\{\phi_\theta^i, i = 1, \ldots, m\}$ is said to be **uniformly linearly independent** if there exists $a > 0$ such that for all $r \in \mathbb{R}^m$,

$$\| r'\phi \|_\theta^2 \geq a|r|^2.$$

The notion of uniform linear independence is stronger than linear independence for each θ. To see this, note that the functions $\phi_\theta^i, i = 1, \ldots, n$, are linearly independent for each θ if and only the function $a(\theta)$ defined by

$$a(\theta) = \inf_{|r|=1} \| r'\phi \|_\theta^2,$$

is strictly positive for all θ whereas these functions are uniformly linearly independent if only if $\inf_\theta a(\theta)$ is strictly positive. Such uniform linear independence assumptions are required to ensure that the function $\bar{r}(\cdot)$ and the sequence of critic parameter values obtained during the course of the algorithm are bounded.

With these preliminaries, the additional assumptions and the convergence results for the two variants can be stated separately as follows.

TD(1) Critic

For the TD(1) critic, the MDP and the given family of RSPs are assumed to satisfy a stronger version of Assumption 2.5.

Assumption 4.4. *The set $\mathbb{X}_0$ of Assumption 2.5 consists of only the state x^* appearing in (4.3).*

Note that the requirement that there is a single state that is hit with positive probability is quite strong but is satisfied in many practical situations involving queueing systems as well as for systems that have been made regenerative using the splitting techniques of (Athreya & Ney, 1978) and (Nummelin, 1978).

Finally, the following assumption places further restrictions on the choice of the basis functions for this variant.

Assumption 4.5. *The basis functions $\phi_\theta^i, i = 1, \ldots, m$, are such that*

1. $\mathbf{E}_{\theta,x^*}[\phi_\theta(x^*, U_0)] = 0, \quad \forall\theta.$
2. *The functions $\phi_\theta^i, i = 1, \ldots, m$, are uniformly linearly independent.*

The assumption that the expected value of the features at x^* is zero, is not that restrictive as for any feature vectors ϕ_θ, the new feature vectors given by

$$\phi_\theta - \underline{1}\mathbf{E}_{\theta,x^*}[\phi_\theta(x^*, U_0)]$$

satisfy the assumption. In particular, if we put $\phi_\theta = \psi_\theta$, where ψ_θ is defined by Eq. (2.1) in Section 2.2, then the assumption is satisfied.

Before we describe the convergence result for the TD(1) critic, we need the following notation. For each θ, consider the matrix $\bar{G}(\theta)$ and the vector $\bar{h}(\theta)$ defined as

$$\begin{aligned} \bar{G}(\theta) &= \langle \phi_\theta, \phi'_\theta \rangle_\theta, \\ \bar{h}(\theta) &= \langle \phi_\theta, Q_\theta \rangle_\theta, \end{aligned}$$

where

$$Q_\theta(x,u) = \mathbf{E}_{\theta,x}\left[\sum_{k=0}^{\tau-1}(g(X_k,U_k) - \bar{\alpha}(\theta))\,\middle|\, U_0 = u\right].$$

Theorem 4.6. *If Assumptions 2.4, 2.5, 2.7, 2.9, 4.1, 4.2, 4.4 and 4.5 hold, then for each θ, the linear equation*

$$\bar{G}(\theta)r = \bar{h}(\theta),$$

has a unique solution $\bar{r}(\theta)$ and

$$\lim_k [|r_k - \bar{r}(\theta_k)| + |\alpha_k - \bar{\alpha}(\theta_k)|] = 0, \quad w.p.1.$$

TD(λ) Critic, $0 < \lambda < 1$

When the policy parameter is fixed at θ, this variant of TD learning is known to converge if the functions $\{\underline{1}, \phi_\theta^1, \ldots, \phi_\theta^n\}$ are linearly independent (see (Van Roy, 1998;

Tsitsiklis & Van Roy, 1999a)) . Furthermore, the functions $\{\underline{1}, \phi_\theta^1, \ldots, \phi_\theta^n\}$ are linearly independent if and only if the modified basis functions

$$\hat{\phi}_\theta^i = \phi_\theta^i - \langle \phi_\theta^i, \underline{1} \rangle_\theta \underline{1}, \quad i = 1, \ldots, m,$$

are linearly independent. But when the policy parameters are changing, we need that these functions be linearly independent "uniformly in θ".

Assumption 4.7. *The modified basis functions $\hat{\phi}_\theta^i = \phi_\theta^i - \langle \phi_\theta^i, \underline{1} \rangle_\theta \underline{1}, i = 1, \ldots, m$, are uniformly linearly independent.*

The statement of the convergence result for this variant is similar to Theorem 4.6 with $\bar{G}(\cdot)$ and $\bar{h}(\cdot)$ redefined as

$$\begin{aligned} \bar{G}(\theta) &= \langle \hat{\phi}_\theta, \hat{\phi}'_\theta \rangle_\theta - (1-\lambda) \sum_{k=0}^{\infty} \lambda^k \langle \hat{\phi}_\theta, P_\theta^{k+1} \hat{\phi}'_\theta \rangle_\theta, \\ \bar{h}(\theta) &= \sum_{k=0}^{\infty} \lambda^k \langle \hat{\phi}_\theta, P_\theta^k (g - \bar{\alpha}(\theta)\underline{1}) \rangle_\theta. \end{aligned}$$

For intuition behind these expressions, we refer the reader to earlier works on TD (Tsitsiklis & Van Roy, 1999a; Bertsekas & Tsitsiklis, 1996).

Theorem 4.8. *If Assumptions 2.4, 2.5, 2.7, 2.9, 4.1, 4.2 and 4.7 hold, then for each θ, the linear equation*

$$\bar{G}(\theta) r = \bar{h}(\theta),$$

has a unique solution $\bar{r}(\theta)$ and

$$\lim_k [|r_k - \bar{r}(\theta_k)| + |\alpha_k - \bar{\alpha}(\theta_k)|] = 0, \quad w.p.1.$$

4.2 Discounted Reward

The relation between the discounted reward problem and the average reward problem (cf. Section 2.4) can be used to obtain algorithms for this problem. In this section, we consider a different kind of algorithms that involve additional coin tosses independent of the MDP. The simulation for these algorithms involves a sequence of $\{0, 1\}$-valued random variables (coin tosses) $\{\hat{V}_k\}$ which may contribute to additional variance compared to other approaches (Tsitsiklis & Van Roy, 1997). However, the steady state distribution of the other approaches may be different from what is needed to calculate the gradient of the discounted cost using Theorem (2.13). When ρ is close to 1, the difference in the steady state distributions as well as the variance in the two approaches is small.

At each time instant k, the random variable $\hat{V}_k$ is generated independent of the past $\hat{X}_l, \hat{U}_l, l \leq k, \hat{V}_l, l < k$, with

$$\mathbf{P}(\hat{V}_k = 1) = 1 - \rho.$$

The state transition at time k is dependent on the coin toss $\hat{V}_k$. If $\hat{V}_k$ is 1, the next state is generated independently of the past and of the current decision, according to the distribution ξ. Otherwise, the next state is generated using the simulator of the original MDP. More formally, the simulation can be described as

$$\begin{aligned}
\mathbf{P}(\hat{V}_l = 1 | \hat{X}_l, \hat{U}_l, l \leq k, \hat{V}_l, l < k) &= 1 - \rho, \\
\mathbf{P}(\hat{X}_{k+1} \in S | \hat{V}_l = 1, \hat{X}_l, \hat{U}_l, l \leq k, \hat{V}_l, l < k) &= \xi(S), \\
\mathbf{P}(\hat{X}_{k+1} \in S | \hat{V}_l = 0, \hat{X}_l, \hat{U}_l, l \leq k, \hat{V}_l, l < k) &= p(S | \hat{X}_k, \hat{U}_k).
\end{aligned}$$

Unlike the average reward case, the critic for the discounted reward problem stores and updates only two parameters r_k and $\hat{Z}_k$. The update of r_k takes place only when the transitions correspond to the original MDP. The update is given by:

$$r_{k+1} = \begin{cases} r_k + \gamma_k d_k \hat{Z}_k, & \text{if} \quad \hat{V}_k = 0, \\ r_k & \text{otherwise,} \end{cases}$$

where d_k's are the temporal differences for discounted reward:

$$d_k = g(\hat{X}_k, \hat{U}_k) + \rho \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) - \phi_{\theta_k}(\hat{X}_k, \hat{U}_k).$$

The eligibility trace vector is updated as follows:

$$\hat{Z}_{k+1} = \begin{cases} \lambda \hat{Z}_k + \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}), & \text{if} \quad \hat{V}_k = 0, \\ \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}), & \text{otherwise.} \end{cases}$$

The basis functions $\phi_\theta^i, i = 1, \ldots m$, are assumed to satisfy the following:

Assumption 4.9. *The basis functions $\phi_\theta^i, i = 1, \ldots m$, are uniformly linearly independent.*

Define the functions $\bar{G}(\cdot)$ and $\bar{h}(\cdot)$ as

$$\begin{aligned}
\bar{G}(\theta) &= \langle \phi_\theta, \phi'_\theta \rangle_\theta - \rho(1 - \lambda) \sum_{k=0}^{\infty} (\rho\lambda)^k \langle \phi_\theta, P_\theta^{k+1} \phi'_\theta \rangle_\theta, \\
\bar{h}(\theta) &= \sum_{k=0}^{\infty} (\lambda\rho)^k \langle \phi_\theta, P_\theta^k g \rangle_\theta.
\end{aligned}$$

The convergence theorem for these algorithms is the following:

Theorem 4.10. *If Assumptions 2.12, 2.9, 4.1, 4.2 and 4.7 hold, then for each θ, the linear equation*

$$\bar{G}(\theta) r = \bar{h}(\theta),$$

has a unique solution $\bar{r}(\theta)$ and

$$\lim_k |r_k - \bar{r}(\theta_k)| = 0, \quad w.p.1.$$

4.3 Total Reward

The algorithms for the case of total reward are similar to those for the case of discounted reward. The artificial MDP for total reward problems has the transition kernel:

$$\tilde{p}(x,u,S) = I_{\{t\}}(x)\xi(S) + I_{\mathbb{X}\setminus\{t\}}(x)p(S|x,u).$$

That is, the next state $\hat{X}_{k+1}$ is generated from the current state $\hat{X}_k$ and decision $\hat{U}_k$ as follows:

- If $\hat{X}_k$ is the terminal state, the next state is generated according to the distribution ξ.
- If $\hat{X}_k$ is a non terminal state, the next state is generated using the simulator of the given MDP.

The critic updates r_k and $\hat{Z}_k$ as follows:

$$\begin{aligned} r_{k+1} &= \begin{cases} r_k + \gamma_k d_k \hat{Z}_k, & \text{if} \quad \hat{X}_k \neq t, \\ r_k & \text{otherwise,} \end{cases} \\ \hat{Z}_{k+1} &= \begin{cases} \lambda \hat{Z}_k + \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}), & \text{if} \quad \hat{X}_k \neq t, \\ \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}), & \text{otherwise.} \end{cases} \end{aligned}$$

where

$$d_k = g(\hat{X}_k, \hat{U}_k) + r_k' \phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) - r_k' \phi_{\theta_k}(\hat{X}_k, \hat{U}_k).$$

In the total reward case, we assume the basis functions satisfy the following.

Assumption 4.11.

1. *For all* $u \in \mathbb{U}$, $\phi_\theta(t,u) = 0$.
2. *The basis functions* $\phi_\theta^i, i = 1, \ldots, m$, *are uniformly linearly independent.*

Define the functions $\bar{h}(\cdot)$ and $\bar{G}(\cdot)$ as

$$\begin{aligned} \bar{h}(\theta) &= \sum_{k=0}^{\infty} \lambda^k \langle \phi_\theta, P_\theta^k g \rangle_\theta, \\ \bar{G}(\theta) &= \langle \phi_\theta, \phi_\theta' \rangle_\theta - (1-\lambda) \sum_{k=0}^{\infty} \lambda^k \langle \phi_\theta, P_\theta^{k+1} \phi_\theta' \rangle_\theta. \end{aligned}$$

The convergence result for the critic for the total reward problem is the following:

Theorem 4.12. *If Assumptions 2.14, 2.9, 4.1, 4.7, 4.11 and 4.2 hold, then for each* θ, *the linear equation*

$$\bar{G}(\theta) r = \bar{h}(\theta),$$

has a unique solution $\bar{r}(\theta)$ *and*

$$\lim_k |r_k - \bar{r}(\theta_k)| = 0, \quad w.p.1.$$

Before we move on, let us describe how the expressions for $\bar{G}$ and $\bar{h}$ are obtained in the above sections. Note that if the actor parameter were fixed at θ, the triplet $\hat{Y}_k = (\hat{X}_k, \hat{U}_k, \hat{Z}_k)$ would be a time-homogeneous Markov chain. Furthermore, the update direction for the critic parameter is of the form $(h(Y_k) - G(\hat{Y}_k)r_k)$ and, $\bar{G}$ and $\bar{h}$ are steady state expectations of $G(\hat{Y}_k)$ and $h(\hat{Y}_k)$ respectively. These steady state expectations are calculated either using a regenerative representation or using the fact that the limit of finite-time expectation is the steady state expectation.

4.4 Convergence Analysis of the Critic

In this section, we analyze the convergence of the the critic for the average reward criterion. The analysis of the other variants is similar and therefore will not be described here. If the actor parameter was held constant at some value θ, it would follow from the results of (Tsitsiklis & Van Roy, 1997; Tsitsiklis & Van Roy, 1999a) that the critic parameters would converge to some $\bar{r}(\theta)$ depending on the variant. In our case, θ_k changes with k, but slowly, and this will allow us to show that $r_k - \bar{r}(\theta_k)$ converges to zero. To establish this fact, we will cast the update of the critic as a linear stochastic approximation driven by Markov noise, specifically in the form of Equation (3.1) of the previous chapter. We will show that the critic updates satisfy all the hypotheses of Theorem 3.2, and the desired results (Theorem 4.6, 4.8) will follow. We start with some notation.

For each time k, let

$$\begin{aligned} Y_k &= (\hat{X}_k, \hat{U}_k, \hat{Z}_k), \\ R_k &= \begin{pmatrix} M\alpha_k \\ r_k \end{pmatrix}, \end{aligned}$$

for some deterministic constant $M > 0$, whose purpose will be clear later. Let $\mathcal{F}_k$ be the σ-field generated by $Y_l, R_l, \theta_l, l \leq k$. For $y = (x, u, z)$, define

$$\begin{aligned} \tilde{h}_\theta(y) &= \begin{pmatrix} Mg(x,u) \\ zg(x,u) \end{pmatrix}, \\ \tilde{G}_\theta(y) &= \begin{pmatrix} 1 & 0 \\ z/M & G_\theta(y) \end{pmatrix}, \end{aligned}$$

where

$$G_\theta(y) = z\left(\phi'_\theta(x,u) - \mathbf{E}_{\theta,x}\left[\phi'_\theta(X_1, U_1) \middle| U_0 = u\right]\right).$$

To understand the role of the scale factor M, consider the steady state expectation $\bar{G}_1(\theta)$ of the matrix $\tilde{G}_\theta(y)$. It is required that this matrix be positive definite, so that certain linear equations have unique solutions. However, $\bar{G}_1(\theta)$ need not be positive definite even though it is almost block diagonal and the matrices on the diagonal of $\bar{G}_1(\theta)$ can be shown to be positive definite. The role of the scale factor M is to reduce the non-diagonal terms and thus make the matrix $\bar{G}_1(\theta)$ positive definite. The value of M will be chosen sufficiently large, depending on the parameter λ. The dependence of M on λ is stated in the proofs of Lemmas 4.18 and 4.22.

Let

$$\xi_{k+1} = \begin{pmatrix} 0 & 0 \\ 0 & \tilde{\xi}_{k+1} \end{pmatrix},$$

where

$$\tilde{\xi}_{k+1} = \hat{Z}_k \phi'_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) - \mathbf{E}[\hat{Z}_k \phi'_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}))|\mathcal{F}_k]$$

is an $m \times m$-matrix valued $\mathcal{F}_k$-martingale difference. Then the update (4.1) for the critic can be written as

$$R_{k+1} = R_k + \gamma_k(\tilde{h}_{\theta_k}(Y_k) - \tilde{G}_{\theta_k}(Y_k)R_k) + \gamma_k \xi_{k+1} R_k.$$

To apply Theorem 3.2 to this update equation, we need to prove that it satisfies Assumptions 3.1 (1)-(10). We will verify these assumptions for the two cases $\lambda = 1$ and $\lambda < 1$ separately. Some of the common lemmas will be proved here and the rest will be presented in the following two subsections. Without loss of generality, assume that $\mathbf{E}[\tilde{L}(\hat{X}_0)] < \infty$, where $\tilde{L}$ is the function in Assumption 2.5.

Lemma 4.13. $\sup_k \mathbf{E}[\tilde{L}(\hat{X}_k)] < \infty$.

Proof. Using Assumption 2.5, note that

$$\begin{aligned} \mathbf{E}[\tilde{L}(\hat{X}_{k+1})] &= \mathbf{E}[\mathbf{E}[\tilde{L}(\hat{X}_{k+1})|\hat{X}_k, \theta_{k-1}]] \\ &= \mathbf{E}[\mathbf{E}_{\theta_{k-1}, \hat{X}_k}[\tilde{L}(X_1)]] \\ &\leq \rho \mathbf{E}[\tilde{L}(\hat{X}_k)] + b\mathbf{P}(\hat{X}_k \notin \mathbb{X}_0). \end{aligned}$$

The result follows because $\rho < 1$. □

Assumptions 3.1(2) and (3) follow from our Assumption 4.2 on the step-size-sequence $\{\gamma_k\}$. Therefore we will concentrate only on the remaining assumptions in the next two sections.

4.4.1 TD(1) Critic

Define a process Z_k in terms of the process $\{(X_k, U_k)\}$ of Chapter 2 (in which the policy is fixed) as follows:

$$Z_0 = \phi_\theta(X_0, U_0), \quad Z_{k+1} = I_{\mathbf{X}\setminus\{x^*\}}(X_{k+1})Z_k + \phi_\theta(X_{k+1}, U_{k+1}),$$

where I is the indicator function. Note that the process $\{Z_k\}$ depends on the parameter θ. Whenever we use this process inside an expectation or a probability measure we will assume that the parameter of this process is the same as the parameter of the probability or expectation. It is easy to see that (X_k, U_k, Z_k) is a Markov chain and that

$$\begin{aligned} &\mathbf{P}((\hat{X}_{k+1}, \hat{U}_{k+1}, \hat{Z}_{k+1}) \in S \times A \times B | \mathcal{F}_k) \\ &= \mathbf{P}((\hat{X}_{k+1}, \hat{U}_{k+1}, \hat{Z}_{k+1}) \in S \times A \times B | \hat{X}_k, \hat{U}_k, \hat{Z}_k, \theta_k), \end{aligned} \tag{4.4}$$

for $S \in \mathcal{B}(\mathbb{X})$, $A \in \mathcal{B}(\mathbb{U})$ and Borel set B. Therefore the update satisfies Assumption 3.1(1).

Let τ be the stopping time defined by

$$\tau = \min\{k > 0 | X_k = x^*\}.$$

For each $\theta \in \mathbb{R}^n$, define $T_\theta \in \mathcal{L}_\theta^2$ as

$$T_\theta(x, u) = \mathbf{E}_{\theta,x}[\tau | U_0 = u].$$

The fact that $T_\theta \in \mathcal{L}^2(\eta_\theta)$ follows from the assumption that $\mathbb{X}_0 = x^*$ (Assumption 4.4), and the uniform ergodicity Assumption 2.5. For each $\theta \in \mathbb{R}^n$, define

$$\begin{aligned} Q_\theta(x, u) &= \mathbf{E}_{\theta,x}\left[\textstyle\sum_{k=0}^{\tau-1}(g(X_k, U_k) - \bar{\alpha}(\theta)) \,\middle|\, U_0 = u\right], \\ \bar{h}_1(\theta) &= \begin{pmatrix} M\bar{\alpha}(\theta) \\ \bar{h}(\theta) + \bar{\alpha}(\theta)\bar{Z}(\theta) \end{pmatrix}, \\ \bar{G}_1(\theta) &= \begin{pmatrix} 1 & 0 \\ \bar{Z}(\theta)/M & \bar{G}(\theta) \end{pmatrix}, \end{aligned}$$

where

$$\begin{aligned} \bar{h}(\theta) &= \langle Q_\theta, \phi_\theta \rangle_\theta, \\ \bar{Z}(\theta) &= \langle T_\theta, \phi_\theta \rangle_\theta, \\ \bar{G}(\theta) &= \langle \phi_\theta, \phi'_\theta \rangle_\theta. \end{aligned}$$

As we will show later, $\bar{h}_1(\theta)$, $\bar{G}_1(\theta)$ and $\bar{Z}(\theta)$ are steady state averages of $\tilde{h}_\theta(Y_k)$, $\tilde{G}_\theta(Y_k)$ and Z_k if the value of the actor parameter was fixed at θ. To prove that $\bar{h}_1(\cdot)$ and $\bar{G}_1(\cdot)$ are bounded (Assumption 3.1(6)) we need the following lemma.

Lemma 4.14. *For any $d > 0$, there exists $K_d > 0$ such that*

1. $\mathbf{E}_{\theta,x}\left[|Q_\theta(x, U_0)|^d\right] \leq K_d \tilde{L}(x)$.

2. $\mathbf{E}_{\theta,x}\left[|T_\theta(x,U_0)|^d\right] \leq K_d\tilde{L}(x)$.

Proof. Since it is sufficient to prove the above results for $d \geq 1$ (this is because $\tilde{L}(x) \geq 1$ for all x), fix some $d \geq 1$. Using Jensen's inequality and (2.3) we have

$$\begin{aligned}\mathbf{E}_{\theta,x}[\tilde{L}(X_1)^{\frac{1}{d}}] &\leq \mathbf{E}_{\theta,x}[\tilde{L}(X_1)]^{\frac{1}{d}} \\ &\leq \rho^{\frac{1}{d}}\tilde{L}(x) + b^{\frac{1}{d}}I\{x \neq x^*\}.\end{aligned}$$

Since, by Assumption 2.9 and Eq. (2.7)

$$|\mathbf{E}_{\theta,x}[g(x,U_0)]|^d \leq \mathbf{E}_{\theta,x}[|g(x,U_0)|^d] \leq K_d L(x),$$

it follows from Theorem 15.2.5 of (Meyn & Tweedie, 1993) that $\mathbf{E}_{\theta,x}[Q_\theta(x,U_0)]^d \leq K_d' L(x)$ for some $K_d' > 0$. Since

$$Q_\theta(x,u) = g(x,u) - \bar{\alpha}(\theta) + \mathbf{E}_{\theta,x}[Q_\theta(X_1,U_1)|U_0 = u],$$

we have

$$\begin{aligned}|Q_\theta(x,u)|^d &\leq C\left(|g(x,u)|^d + 1 + \mathbf{E}_{\theta,x}\left[|\mathbf{E}_{\theta,X_1}[Q_\theta(X_1,U_1)]|^d\,|U_0 = u\right]\right) \\ &\leq C_1\left(|g(x,u)|^d + 1 + \mathbf{E}_{\theta,x}\left[\tilde{L}(X_1)\,|U_0 = u\right]\right).\end{aligned}$$

Taking expectations on both sides we have the required result. The proof of the second part is similar. □

Lemma 4.15. *$\bar{G}_1(\cdot)$ and $\bar{h}_1(\cdot)$ are bounded.*

Proof. From Assumptions 4.1, Lemmas 4.14 and 4.13, it follows that $||\phi_\theta||_\theta$, $||T_\theta||_\theta$ and $||Q_\theta||_\theta$ are all bounded and therefore their inner products are also bounded. □

Lemma 4.16. *For each $\theta \in \mathbb{R}^n$,*

1. $\mathbf{E}_{\theta,x^*}\left[\sum_{k=0}^{\tau-1}((g(X_k,U_k) - \bar{\alpha}(\theta))Z_k - \langle Q_\theta, \phi_\theta\rangle_\theta)\right] = 0$,
2. $\mathbf{E}_{\theta,x^*}\left[\sum_{k=0}^{\tau-1}(Z_k(\phi_\theta'(X_k,U_k) - \phi_\theta'(X_{k+1},U_{k+1})) - \langle\phi_\theta,\phi_\theta'\rangle_\theta)\right] = 0$.

Proof. We have

$$\begin{aligned}&\mathbf{E}_{\theta,x^*}\left[\sum_{k=0}^{\tau-1}(g(X_k,U_k) - \bar{\alpha}(\theta))Z_k\right] \\ &= \mathbf{E}_{\theta,x^*}\left[\sum_{k=0}^{\tau-1}(g(X_k,U_k) - \bar{\alpha}(\theta))\sum_{l=0}^{k}\phi_\theta(X_k,U_k)\right] \\ &= \mathbf{E}_{\theta,x^*}\left[\sum_{l=0}^{\tau-1}\phi_\theta(X_l,U_l)\sum_{k=l}^{\tau-1}(g(X_k,U_k) - \bar{\alpha}(\theta))\right] \\ &= \mathbf{E}_{\theta,x^*}\left[\sum_{l=0}^{\tau-1}\phi_\theta(X_l,U_l)Q_\theta(X_l,U_l)\right] \\ &= \mathbf{E}_{\theta,x^*}[\tau]\langle\phi_\theta,Q_\theta\rangle_\theta.\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbf{E}_{\theta,x^*}\left[\textstyle\sum_{k=0}^{\tau-1} Z_k(\phi'_\theta(X_k,U_k)-\phi'_\theta(X_{k+1},U_{k+1}))\right]\\
&=\mathbf{E}_{\theta,x^*}\left[\textstyle\sum_{l=0}^{\tau-1}\phi_\theta(X_l,U_l)\sum_{k=l}^{\tau-1}(\phi'_\theta(X_k,U_k)-\phi'_\theta(X_{k+1},U_{k+1}))\right]\\
&=\mathbf{E}_{\theta,x^*}\left[\textstyle\sum_{l=0}^{\tau-1}\phi_\theta(X_l,U_l)(\phi'_\theta(X_l,U_l)-\phi'_\theta(X_\tau,U_\tau))\right]\\
&=\mathbf{E}_{\theta,x^*}[\tau]\langle\phi_\theta,\phi'_\theta\rangle_\theta.
\end{aligned}$$

where the last equality follows from Assumption 4.5 and Equation (2.6). □

This lemma suggests the following regenerative representations. For $y=(x,u,z)$, define

$$\hat{h}_\theta(y)=\mathbf{E}_{\theta,x}\left[\sum_{k=0}^{\tau-1}(\tilde{h}_\theta(Y_k)-\bar{h}(\theta))\,\middle|\,U_0=u,Z_0=z\right],$$

$$\hat{G}_\theta(y)=\mathbf{E}_{\theta,x}\left[\sum_{k=0}^{\tau-1}(\tilde{G}_\theta(Y_k)-\bar{G}(\theta))\,\middle|\,U_0=u,Z_0=z\right].$$

Using the previous lemma, it is a little algebraic exercise to verify that $\hat{h}_\theta(\cdot)$ and $\hat{G}_\theta(\cdot)$ satisfy Assumption 3.1(5). To prove that these functions satisfy Assumption 3.1(7) and (9) we will need the following result.

Lemma 4.17. *For each $d>1$, $\sup_k \mathbf{E}[|\hat{Z}_k|^d]<\infty$.*

Proof. Let $\hat{Y}_k$ denote the vector $(\hat{X}_k,\hat{U}_k,\hat{Z}_k,r_k,\lambda_k,\theta_k)$. Since the step size sequences $\{\gamma_k\}$ and $\{\beta_k\}$ are deterministic, $\{\hat{Y}_k\}$ forms a time varying Markov chain. For each k, let $\mathbf{P}_{k,\hat{y}}$ denote the conditional law of the process $\{\hat{Y}_n\}$ given that $\hat{Y}_k=\hat{y}$. Define a sequence of stopping times for the process $\{\hat{Y}_n\}$ as follows:

$$\hat{\tau}_k=\min\{n>k:\hat{X}_n=x^*\}.$$

For $1<t<\frac{1}{\rho}$, define

$$V_k^{(d)}(\hat{y})=\mathbf{E}_{k,\hat{y}}\left[\sum_{l=k}^{\tau_k-1}t^{l-k}(1+|\hat{Z}_l|^d)\right],$$

which can be verified to be finite, due to uniform geometric ergodicity and the bounds on $\phi_\theta(x,u)$. It is easy to see that $V_k^{(d)}(\hat{Y}_k)\geq|\hat{Z}_k|^d$. Therefore it is sufficient to prove that

$$\sup_k \mathbf{E}[V_k^{(d)}(\hat{Y}_k)]<\infty.$$

We will now show that $V_k^{(d)}(\hat{y})$ acts as a Lyapunov function for the algorithm. Indeed,

$$\begin{aligned} V_k^{(d)}(\hat{y}) &\geq \mathbf{E}_{k,\hat{y}}\left[\sum_{l=k+1}^{\tau_k-1} t^{l-k}(1+|\hat{Z}_l|^d)\right] \\ &= \mathbf{E}_{k,\hat{y}}\left[\sum_{l=k+1}^{\tau_k-1} t^{l-k}(1+|\hat{Z}_l|^d) I\{\hat{X}_{k+1} \neq x^*\}\right] \\ &= t\mathbf{E}_{k,\hat{y}}\left[V_{k+1}^{(d)}(\hat{Y}_{k+1}) I\{\hat{X}_{k+1} \neq x^*\}\right] \\ &= t\mathbf{E}_{k,\hat{y}}\left[V_{k+1}^{(d)}(\hat{Y}_{k+1})\right] - t\mathbf{E}_{k,\hat{y}}\left[V_{k+1}^{(d)}(\hat{Y}_{k+1}) I\{X_{k+1} = x^*\}\right]. \end{aligned}$$

Using (2.3), some algebraic manipulations and the bounds on $\phi_\theta(\cdot,\cdot)$, we can verify that

$$\mathbf{E}_{k,\hat{y}}\left[V_{k+1}^{(d)}(\hat{Y}_{k+1}) I\{X_{k+1} = x^*\}\right]$$

is bounded. Finally, since $\mathbf{E}[V_{k+1}^{(d)}(\hat{Y}_{k+1})|\hat{Y}_k = \hat{y}] = \mathbf{E}_{k,\hat{y}}[V_{k+1}^{(d)}(\hat{Y}_{k+1})]$, we have

$$\mathbf{E}[V_{k+1}^{(d)}(\tilde{Y}_{k+1})] \leq \frac{1}{t}\mathbf{E}[V_k^{(d)}(\tilde{Y}_k)] + C$$

for some constant $C > 0$. The rest follows as $t > 1$. □

Using the above result it is easy to verify Assumption 3.1(4). To verify Assumption 3.1(7), note that $\hat{h}_\theta(\cdot), \hat{G}_\theta(\cdot), \tilde{h}_\theta(\cdot)$ and $\tilde{G}_\theta(\cdot)$ are affine in z and therefore can be expressed as

$$f_\theta^{(1)}(x,u) + z f_\theta^{(2)}(x,u)$$

for some functions $f_\theta^{(i)}$ for which

$$\mathbf{E}_{\theta,x}\left[|f_\theta^{(i)}(x,U_0)|^d\right] \leq K_d \tilde{L}(x)$$

for some $K_d > 0$. Therefore, Holder's inequality and the previous results can be used to see that Assumption 3.1(7) is indeed satisfied. As in the proof of Theorem 2.10, likelihood ratio methods can be used to verify Assumptions 3.1 (8) and (9).

Before we move on, note that the verification of Assumptions 3.1(9) involves upper bounding "derivatives" of $f_\theta^{(i)}(x,u)$ by functions independent of θ. The only function of (x,u) we introduced till now that is independent of θ is $L(x,u)$. But the functions $f_\theta^{(i)}(x,u)$ cannot be expected to be upper bounded by L. However, another family of functions that serves the purpose is the following:

$$\hat{L}_d(x,u) = \mathbf{E}_{\theta,x}[\tilde{L}(X_1)|U_0 = u]^{\frac{1}{d}}, \quad d > 0.$$

First, it is easy to see that $\hat{L}_d$ is in fact independent of θ as the expectation on the r.h.s. of the above equation is actually the expectation with respect to the transition kernel of the MDP. Furthermore, to see that this function is an appropriate one note that a combination of L and $\hat{L}$ can be used to bound quantities of the form $\mathbf{E}_{\theta,x}[\cdot|U_0 = u]$ and that

$$\mathbf{E}_{\theta,x}\left[\hat{L}_d(x,U_0)^d\right] = \mathbf{E}_{\theta,x}\left[\tilde{L}(X_1)\right] \leq \rho\tilde{L}(x) + b.$$

For example, it follows from the proof of Lemma 4.14 that Q_θ is bounded above by $C(1 + L + L_d)$ for some constant $C > 0$.

Finally, the following lemma verifies Assumption 3.1(10).

Lemma 4.18. *There exist $M, \epsilon > 0$ such that for all $\theta \in \mathbb{R}^n$ and $R \in \mathbb{R}^{m+1}$,*

$$R'\bar{G}(\theta)R \geq \epsilon|R|^2.$$

Proof. Let

$$R = \begin{pmatrix} \lambda \\ r \end{pmatrix}$$

for some $r \in \mathbb{R}^m$. Then, using Assumption 4.1(3) for the first inequality,

$$\begin{aligned} R'\bar{G}(\theta)R &= ||r'\phi_\theta||_\theta^2 + |\lambda|^2 + \tfrac{r'\bar{Z}(\theta)\lambda}{M} \\ &\geq a|r|^2 + |\lambda|^2 - \tfrac{r'\bar{Z}(\theta)\lambda}{M} \\ &\geq \min(a,1)|R|^2 - \tfrac{|\bar{Z}(\theta)|(|r|^2+|\lambda|^2)}{2M} \\ &= \left(\min(a,1) - \tfrac{\bar{Z}(\theta)}{2M}\right)|R|^2. \end{aligned}$$

Choose $M > \sup_\theta |\bar{Z}(\theta)|/\min(a,1)$, which is possible since $\bar{Z}(\theta)$ is bounded (cf. Lemma 4.15). □

4.4.2 TD(λ) Critic, $\lambda < 1$

To analyze the TD(λ) critic with $0 < \lambda < 1$, redefine the process Z_k as

$$Z_{k+1} = \lambda Z_k + \phi_\theta(X_{k+1}, U_{k+1}).$$

It is easy to see that, with this definition, Equation (4.4) holds for the TD(λ) critic also. This means that Assumption 3.1(1) is satisfied. For each $\theta \in \mathbb{R}^n$, let

$$\bar{h}_1(\theta) = \begin{pmatrix} M\bar{\alpha}(\theta) \\ \bar{h}(\theta) + \bar{\alpha}(\theta)\bar{Z}(\theta) \end{pmatrix},$$

$$\bar{G}_1(\theta) = \begin{pmatrix} 1 & 0 \\ \bar{Z}(\theta)/M & \bar{G}(\theta) \end{pmatrix},$$

where

$$\bar{h}(\theta) = \sum_{k=0}^{\infty} \lambda^k \langle \phi_\theta, P_\theta^k c - \bar{\alpha}(\theta)\underline{1}\rangle_\theta, \quad \bar{G}(\theta) = \sum_{k=0}^{\infty} \lambda^k \langle \phi_\theta, P_\theta^k (\phi'_\theta - P_\theta \phi'_\theta)\rangle_\theta,$$

and $\bar{Z}(\theta) = (1-\lambda)^{-1}\langle \underline{1}, \phi_\theta \rangle_\theta$. As in Assumption 4.7, let $\hat{\phi}_\theta = \phi_\theta - \langle \phi_\theta, \underline{1}\rangle_\theta \underline{1}$. Then, $P_\theta \phi_\theta - \phi_\theta = P_\theta \hat{\phi}_\theta - \hat{\phi}_\theta$ and therefore, $\bar{G}_1(\theta)$ can also be written as

$$\bar{G}(\theta) = \langle \hat{\phi}_\theta, \hat{\phi}'_\theta \rangle_\theta - (1-\lambda) \sum_{k=0}^{\infty} \lambda^k \langle \hat{\phi}_\theta, P_\theta^{k+1} \hat{\phi}'_\theta \rangle_\theta.$$

It is easy to see, using the Cauchy-Schwartz inequality and the boundedness of $||\hat{\phi}_\theta||_\theta$ (see the proof of Lemma 4.15), that $\bar{G}_1(\cdot)$ and $\bar{h}_1(\cdot)$ are bounded and therefore Assumption 3.1 (6) is satisfied. The following lemma is used to verify Assumption 3.1(5).

Lemma 4.19. *There exists $C > 0$ such that*

1. $|\mathbf{E}_{\theta,x}[(g(X_k, U_k) - \bar{\alpha}(\theta))Z_k] - \bar{h}_1(\theta)| \leq Ck \max(\lambda, \rho)^k \tilde{L}(x)$,
2. $|\mathbf{E}_{\theta,x}[Z_k(\phi'_\theta(X_k, U_k) - \phi'_\theta(X_{k+1}, U_{k+1}))] - \bar{G}(\theta)| \leq Ck \max(\lambda, \rho)^k \tilde{L}(x)$.

Proof. We will only prove part 1 since the proof of the other part is similar. We have

$$\begin{aligned}
&|\mathbf{E}_{\theta,x}[(g(X_k, U_k) - \bar{\alpha}(\theta))Z_k] - \bar{h}(\theta)| \\
&\leq \left[\textstyle\sum_{l=0}^{k} \lambda^l \left|\mathbf{E}_{\theta,x}[(g(X_k, U_k) - \bar{\alpha}(\theta))\phi_\theta(X_{k-l}, U_{k-l})] - \langle P_\theta^l c - \bar{\alpha}(\theta)\underline{1}, \phi_\theta\rangle_\theta\right|\right] \\
&\quad + C'\lambda^k \\
&\leq \textstyle\sum_{l=0}^{k} C'\lambda^l \rho^{k-l} \tilde{L}(x) + C'\lambda^k \\
&\leq \textstyle\sum_{l=0}^{k} C' \max(\lambda, \rho)^k \tilde{L}(x) + C'\lambda^k.
\end{aligned}$$

where the second inequality follows from Assumptions 2.5 and 4.1. □

From the previous lemma, it is clear that for $\theta \in \mathbb{R}^n$ and $y = (x, u, z)$,

$$\hat{h}_\theta(y) = \sum_{k=0}^{\infty} \mathbf{E}_{\theta,x} \left[(\tilde{h}_\theta(Y_k) - \bar{h}(\theta)) \middle| U_0 = u, Z_0 = z \right],$$

$$\hat{G}_\theta(y) = \sum_{k=0}^{\infty} \mathbf{E}_{\theta,x} \left[(\tilde{G}_\theta(Y_k) - \bar{G}(\theta)) \middle| U_0 = u, Z_0 = z \right],$$

are well defined and it is easy to check that these satisfy Assumption 3.1(5).

The verification of Assumptions 3.1(8) and (9) is tedious and therefore will only be outlined here. The trick is to write $\hat{h}_\theta(\cdot), \hat{G}_\theta(\cdot)$ in the form:

$$\sum_{k=0}^{\infty} \lambda^k \mathbf{E}_{\theta,x}[f_\theta(Y_k)|U_0 = u, Z_0 = z]$$

and show that the map $\theta \to \mathbf{E}_\theta[f_\theta(Y_k)|U_0 = u, Z_0 = z]$ is Lipschitz continuous with Lipschitz constant at most polynomial in k. The "forgetting" factor λ^k dominates the polynomial in k and thus the sum will be Lipschitz continuous in θ.

To verify Assumption 3.1(7) we have the following counterpart of Lemma 4.17.

Lemma 4.20. *For any $d > 1$, $\sup_k \mathbf{E}[|\hat{Z}_k|^d] < \infty$.*

Proof. We have

$$\begin{aligned} |\hat{Z}_k|^d &= \tfrac{1}{(1-\lambda)^d} \left| (1-\lambda) \textstyle\sum_{l=0}^{k} \lambda^{k-l} \phi_{\theta_k}(\hat{X}_k, \hat{U}_k) \right|^d \\ &\leq \tfrac{1}{(1-\lambda)^d} (1-\lambda) \textstyle\sum_{l=0}^{k} \lambda^{k-l} |\phi_{\theta_k}(\hat{X}_k, \hat{U})|^d, \end{aligned}$$

where the inequality follows from Jensen's inequality. The rest follows from the fact that

$$\sup_k \mathbf{E}[|\phi_{\theta_k}(\hat{X}_k, \hat{U}_k)|^d] < \infty.$$

□

Finally, we will verify Assumption 3.1(10). The following lemma helps us in verifying this assumption. It will also be used to derive various bounds in the future chapters. It says that under any policy, for certain functions f of state, not only does the conditional mean of $f(X_k)$ given X_0 go to the steady state expectation of $f(X_k)$ at geometric rate, the steady state variance of the conditional mean of $f(X_k)$ given X_0 goes to zero at a geometric rate. To make this precise, recall state-decision value functions Q_θ and state value functions V_θ defined as follows:

$$
\begin{aligned}
Q_\theta(x,u) &= \sum_{k=0}^{\infty} \mathbf{E}_{\theta,x}\left[(g(X_k,U_k)-\bar{\alpha}(\theta))|\, U_0=u\right], \\
V_\theta(x) &= \mathbf{E}_{\theta,x}\left[Q_\theta(x,U_0)\right].
\end{aligned}
$$

Lemma 4.21. *For each $\theta \in \mathbb{R}^n$, let $\| P_\theta^k \|_\theta$ denote the norm of the operator P_θ^k restricted to the span of Q_θ, V_θ $\hat{\phi}_\theta^i, \tilde{\phi}_\theta^i, i = 1, \ldots, m$, where for each θ,*

$$
\tilde{\phi}_\theta(x) = \mathbf{E}_{\theta,x}\left[\hat{\phi}_\theta(x,U_0)\right].
$$

Then for some $\rho_0 < 1$ and $C > 0$ we have

$$
\| P_\theta^k \|_\theta \leq C\rho_0^k, \quad \forall \theta, k.
$$

Proof. Note that for any $f_\theta \in \mathcal{L}_\theta^2$,

$$
\begin{aligned}
P_\theta^k f_\theta &= \mathbf{E}_{\theta,x}[f_\theta(X_k,U_k)|U_0=u] \\
&= \mathbf{E}_{\theta,x}\left[\mathbf{E}_{\theta,x}[f_\theta(X_k,U_k)|X_k]\right] \\
&= P_\theta^k \tilde{f}_\theta,
\end{aligned}
$$

where

$$
\tilde{f}(x,u) = \mathbf{E}_{\theta,x}[f(x,U_0)],
$$

Furthermore, using Jensen's inequality it is easy to see that $\| f_\theta \|_\theta \geq \| \tilde{f}_\theta \|_\theta$. Therefore, it is enough to compute the norm of P_θ^k further restricted to the span of V_θ, $\tilde{\phi}_\theta^i, i = 1, \ldots, n$. Let $\mathcal{L}^{\sqrt{\tilde{L}}}$ denote set of all functions f of state such that

$$
\sup_x \frac{|f(x)|}{\sqrt{\tilde{L}(x)}} < \infty,
$$

with the corresponding weighted norm denoted by $\| \cdot \|_{\sqrt{\tilde{L}}}$. Since the functions that span the subspace under consideration are finite in number and are in a bounded (uniformly in θ) $\mathcal{L}^{\sqrt{\tilde{L}}}$-ball, the unit $\mathcal{L}_\theta^2$-balls in the subspace are all contained in a $\mathcal{L}^{\sqrt{\tilde{L}}}$-ball. Furthermore, it is easy to see that $\sqrt{\tilde{L}}$ satisfies Lyapunov condition with ρ replaced by $\sqrt{\rho}$. Therefore, from Theorem 16.0.1 of (Meyn & Tweedie, 1993) it follows that for some constant $C > 0$,

$$
P_\theta^k f_\theta \leq C\rho^{k/2}\sqrt{\tilde{L}},
$$

for all f_θ such that $\pi_\theta(f_\theta) = 0$, $\| f_\theta \|_\theta = 1$. The result follows as $\pi_\theta(\tilde{L})$ is bounded uniformly in θ. □

Lemma 4.22. *There exists $L, \epsilon > 0$ such that for all $\theta \in \mathbb{R}^n$ and $R \in \mathbb{R}^{m+1}$*

$$R'\bar{G}_1(\theta)R \geq \epsilon |R|^2.$$

Proof. Recall the projected feature vectors $\hat{\phi}_\theta^j, j = 1, \ldots, m$. Due to the previous lemma, we have for some constant $C > 0$,

$$\| P_\theta^k(r'\hat{\phi}_\theta^j) \|_\theta \leq C\rho_0^k \| r'\hat{\phi}_\theta \|_\theta .$$

Therefore,

$$\begin{aligned}
r'\bar{G}(\theta)r &= r'\langle \hat{\phi}_\theta, \hat{\phi}'_\theta \rangle_\theta r - (1-\lambda)\sum_{k=0}^{\infty} \lambda^k r'\langle \hat{\phi}_\theta, P_\theta^{k+1}\hat{\phi}'_\theta \rangle_\theta r \\
&= \| r'\hat{\phi}_\theta \|_\theta^2 - (1-\lambda)\sum_{k=0}^{\infty} \lambda^k \langle r'\hat{\phi}_\theta, P_\theta^{k+1}(r'\hat{\phi}_\theta)\rangle_\theta \\
&\geq \| r'\hat{\phi}_\theta \|_\theta^2 - (1-\lambda)\left\{ \sum_{k=0}^{k_0-1} \lambda^k \| r'\hat{\phi}_\theta \|_\theta^2 + \sum_{k \geq k_0} C_1 \lambda^k \rho_0^{k+1} \| r'\hat{\phi}_\theta \|_\theta^2 \right\} \\
&\geq \| r'\hat{\phi}_\theta \|_\theta^2 - (1-\lambda^{k_0}) \| r'\hat{\phi}_\theta \|_\theta^2 - C_1(\lambda\rho_0)^{k_0} \tfrac{\rho_0(1-\lambda)}{(1-\rho_0\lambda)} \| (r'\hat{\phi}_\theta) \|_\theta^2 \\
&\geq \| r'\hat{\phi}_\theta \|_\theta^2 \lambda^{k_0} \left(1 - \tfrac{C_2\rho_0^{k_0+1}(1-\lambda)}{(1-\rho_0\lambda)} \right).
\end{aligned}$$

Take k_0 such that

$$\rho_0^{k_0+1} < \frac{(1-\rho_0\lambda)}{C_2(1-\lambda)}.$$

The rest is similar to proof of Lemma 4.18. □

4.5 Closing Remarks

The TD algorithm for the discounted reward problem is new. Note that, unlike the algorithm proposed in (Tsitsiklis & Van Roy, 1997), the convergence of this variant does not require the Markov chain to be ergodic. Furthermore, the mechanism for sampling of states in our variant is different from simulation of a single trajectory.

Our convergence result of the critic is central to our thesis and allows us to design actor-critic algorithms by abstracting the critic as a black box that outputs an approximation to the value function of the current policy of the actor with "asymptotically negligible error". This approximation is the one to which it would have converged if the actor parameters were frozen at the current values. Crucial to our result is the fact that actor changes slowly with respect to the critic.

This is the first result on TD learning with changing policies. The proof techniques used here also provide a unified approach to converge analysis of several variants of TD. In particular, these methods can be used to prove convergence of "replace trace" methods (Singh & Sutton, 1996). The proof techniques as complicated as the ones used in this chapter are needed only for the analysis of the algorithms in which the policy changes at each time step. We haven't considered here episodic variants of TD in which the parameters of both the critic and the actor are updated only when the system visits a certain state in average reward case or the terminal state in total

reward case. For those times between the visits to the terminal state, the policy used is the one to which the actor was updated after the last visit to the terminal state. The analysis of these variants is much simpler than those considered here as the noise in the estimate corresponds to martingale differences which are much easier to handle. Furthermore, the uniform ergodicity assumption can be relaxed for episodic variants. Due to their simplicity, only these variants will be taken up for study of rate of convergence later in the thesis.

Chapter 5

Actor-Critic Algorithms

In this chapter, we propose several variants of actor-critic algorithms which can be viewed as stochastic gradient algorithms on the parameter space of the actor. They differ from related algorithms like REINFORCE (Williams, 1992), likelihood ratio methods (Glynn, 1987; Glynn & L'Ecuyer, 1995), direct gradient methods (Marbach & Tsitsiklis, 2001; Baxter & Barlett, 1999) which are also stochastic gradient methods. While the gradient estimate depends on the value function approximation provided by the critic in actor-critic methods, the gradient is directly estimated from simulation in other methods. Therefore, we refer to these other methods as actor-only methods. The actor-only methods do not store any additional parameters other than the policy parameter vector θ (and average reward estimate in average reward problems). However, in actor-only methods, the variance in the estimate of the gradient can be very large for systems with large state spaces. The principal reason for this is that the estimate depends on the simulated path of the system starting from a particular state until the system's first return back to that state. For systems with large state spaces and for systems which take too long to reach steady state, the variance of the gradient estimate can be very large for the estimate to be useful. To alleviate this situation, several variance reduction techniques have been proposed. All these techniques amount to "throwing off" some part of the trajectory. For example one technique (Marbach, 1998) is to discount exponentially the contribution of the states visited in the distant past. Such variance reduction techniques introduce bias in the estimate and therefore there is always a trade off between the bias and the variance in the estimate. One of the main contributions of this thesis is to show that the addition of a critic to actor only methods (which makes them actor-critic methods) potentially improves the bias-variance trade off. In other words, depending on the capability of the critic to approximate value functions, the actor-critic algorithms can have much better bias for the same variance as their counterpart actor-only methods.

In the next two sections, we present our two different classes of actor-critic algorithms. The intuition for the algorithms is described using the connection between the gradient formulas of Chapter 2 and the convergence results on TD and the critic (Chapter 4). The next section also states and proves the convergence result for these algorithms.

5.1 Actor without Eligibility Traces

The variants presented in this section are motivated by the observation that the quantity

$$Q_{\theta_k}(X_k, U_k)\psi_{\theta_k}(X_k, U_k) \tag{5.1}$$

can be viewed as an estimate of $\nabla\bar{\alpha}(\theta)$, as its steady state expectation under the policy associated with θ_k is the gradient (cf. Theorem 2.10). Here Q_θ is a state-decision value function under policy θ and $\psi_\theta = \nabla \ln \mu_\theta(x, u)$. Since a critic can provide only an approximation $\hat{Q}_\theta$ of Q_θ, even if it converges instantly, the question is whether (5.1) remains a viable estimate of the gradient if Q_θ is replaced by an approximation $\hat{Q}_\theta$. It is obvious that the answer is yes if the approximation error is small. A more subtle point is that the estimate remains viable even when the approximation error is large but "almost" orthogonal to the functions

$$\psi_\theta^i(x, u) = \frac{\partial \ln \mu_\theta(u|x)}{\partial \theta_i} \tag{5.2}$$

in $\mathcal{L}_\theta^2$ (the space $\mathcal{L}_\theta^2$ was defined in Chapter 2). To see this, note that the bias contributed by the error $Q_\theta - \hat{Q}_\theta$ in the approximation of Q_θ is given by

$$\langle Q_\theta - \hat{Q}_\theta, \psi_\theta \rangle$$

which is zero if $Q_\theta - \hat{Q}_\theta$ is orthogonal to the $\psi_\theta^i, i = 1, \ldots n$. Therefore, we arrive at the following important conclusion. As far as convergence of gradient methods over a parametric family of policies is concerned, the objective of value function approximation should be to approximate the projection of the value function on the subspace Ψ_θ spanned by the basis functions $\psi_\theta^i, i = 1, \ldots, n$, rather than to approximate the value function itself.

The analysis in (Tsitsiklis & Van Roy, 1997; Tsitsiklis & Van Roy, 1999a) shows that this is precisely what Temporal Difference (TD) learning algorithms try to accomplish, i.e., to compute the projection of an exact value function onto the subspace spanned by the basis functions. This allows us to implement the critic by using a TD algorithm. (Note, however, that other types of critics are possible, e.g., based on batch solution of least squares problems, as long as they aim at computing the same projection.) However, we present a family of algorithms which also includes methods in which the critic tries to approximate the exact value function. The reason is that the computation of the exact projection needs the use of TD(1) critic which suffers from large variance. As we will show later, if we use TD(λ) with $\lambda < 1$ for the critic, bias is introduced into the estimate which depends on the error in the approximation of the exact value function rather than the projected value function. Therefore there is a trade off between the variance and prior knowledge about the value function as described by the critic's ability to approximate it. This trade off is discussed later in more detail.

Since the approximation of the state-decision value function provided by the critic

at time k is $r_k'\phi_{\theta_k}$, the actor parameter in this variant is updated along the direction of

$$r_k'\phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1})\psi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1})$$

at time k. More precisely, the actor parameter update for this variant is given by

$$\theta_{k+1} = \theta_k + \beta_k \Gamma(r_k) r_k'\phi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1})\psi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}).$$

Note that the step-size for the actor, at time k, is $\Gamma(r_k)\beta_k$ which depends on the parameter vector of the critic. The relation between the step-size of the critic, γ_k, and that of the actor is described in the following assumption.

Assumption 5.1.

1. *The step-sizes β_k and γ_k are deterministic, non-increasing and satisfy*

$$\sum_k \beta_k = \sum_k \gamma_k = \infty,$$

$$\sum_k \beta_k^2 < \infty, \quad \textit{and} \quad \sum_k \gamma_k^2 < \infty.$$

2. *For some $d > 0$, we have*

$$\sum_k \left(\frac{\beta_k}{\gamma_k}\right)^d < \infty.$$

3. *The function $\Gamma(\cdot)$ is assumed to satisfy the following inequalities for some positive constants $C_1 < C_2$:*

$$\begin{aligned} |r|\Gamma(r) &\in [C_1, C_2], \quad \forall r \in \mathbb{R}^m, \\ |\Gamma(r) - \Gamma(\hat{r})| &\leq \frac{C_2}{1 + |r| + |\hat{r}|} \quad \forall r, \hat{r} \in \mathbb{R}^n. \end{aligned} \tag{5.3}$$

The first part of the above assumption is standard for stochastic approximation algorithms. The second assumption is one of the central assumptions of this thesis. It implies that the critic is updated on a faster time-scale than the actor, where the separation of time-scales is achieved by using different step sizes γ_k and β_k for the critic and actor respectively. For finite MDP's the second part can be weakened to $\beta_k/\gamma_k \to 0$. The role of the factor $\Gamma(r_k)$ in the actor's step-size is to prevent the actor from taking large steps in a "wrong" direction. It represents the user's intuition about the "right size" of critic parameters when the critic has converged. A simple example of the function $\Gamma(\cdot)$ satisfying the above assumption is the following: for

$r = (r^1, \ldots, r^m)$,

$$\Gamma(r) = \begin{cases} 1 & \text{if} \quad \sum_i |r^i| < C, \\ \frac{1+C}{(1+\sum_i |r^i|)} & \text{otherwise,} \end{cases}$$

where r^i is the i-th component of vector r.

The critic for the variants of this section is as described in the previous chapter, with basis functions $\phi_\theta^j, j = 1, \ldots, m$, chosen to satisfy the following assumption.

Assumption 5.2. *For each $\theta \in \mathbb{R}^n$, the subspace Φ_θ in $\mathcal{L}_\theta^2$ spanned by the basis functions $\phi_\theta^i, i = 1, \ldots, m$, of the critic contains the subspace Ψ_θ spanned by the functions $\psi_\theta^j, j = 1, \ldots, n$, i.e.,*

$$\Phi_\theta \supset \Psi_\theta, \quad \forall \theta \in \mathbb{R}^n.$$

As we have argued earlier, ideally, the critic should compute the projection of value function onto the subspace Ψ_θ for any given $\theta \in \mathbb{R}^n$. Therefore it is sufficient for the critic to use $\psi_\theta^i, i = 1, \ldots n$, as basis functions for the critic. Nevertheless, we allow the possibility that $m > n$ and that Φ_θ properly contains Ψ_θ, so that the critic uses more features than are actually necessary. This added flexibility may turn out to be useful in a number of ways:

1. It is possible that for certain values of θ, the feature vectors ψ_θ^i are either close to zero or are almost linearly dependent. For these values of θ, the matrix $\bar{G}(\theta)$ defined in Theorem 4.8 becomes ill-conditioned which can have a negative effect on the performance of the algorithms. This might be avoided by using a richer set of features ϕ_θ^i.

2. When the critic uses TD(λ) with $\lambda < 1$, it can only compute an approximate - rather than exact - projection. The use of additional features can result in a reduction of the approximation error.

To understand what the additional features should be, let us consider the bias, denoted by $b_\theta(\lambda)$, in the estimate of the gradient of the average reward for policy θ. Due to Theorem 4.8, we can assume that the actor updates its parameter along

$$\bar{r}(\theta)' \phi_\theta(X_k, U_k) \psi_\theta(X_k, U_k)$$

when the actor parameter is θ. Therefore, the bias $b_\theta(\lambda)$ in the update direction is the difference between the gradient at θ and the steady state expectation of the update direction. Using Theorem 2.10, the bias $b_\theta(\lambda)$ can be seen to be

$$\left\langle Q_\theta - \bar{r}(\theta)' \hat{\phi}_\theta, \psi_\theta \right\rangle_\theta,$$

$\hat{\phi}_\theta$ is defined as

$$\hat{\phi}_\theta = \phi_\theta - \langle \phi_\theta, \underline{1} \rangle_\theta.$$

Consider the above with ψ_θ replaced by $\hat{\phi}_\theta$:

$$\begin{aligned} \left\langle Q_\theta - \bar{r}(\theta)'\hat{\phi}_\theta, \hat{\phi}_\theta \right\rangle_\theta &= \left\langle \hat{\phi}_\theta, Q_\theta \right\rangle_\theta - \left\langle \hat{\phi}_\theta, \hat{\phi}_\theta' \bar{r}(\theta) \right\rangle_\theta \\ &= \left\langle \hat{\phi}_\theta, Q_\theta \right\rangle_\theta - \bar{h}(\theta) + \left(\bar{G}(\theta) - \left\langle \hat{\phi}_\theta, \hat{\phi}_\theta' \right\rangle_\theta \right) \bar{r}(\theta). \end{aligned}$$

The second equality follows from the definition of $\bar{r}(\theta) = \bar{G}(\theta)^{-1}\bar{h}(\theta)$ of $\bar{r}(\theta)$. Using the fact that

$$Q_\theta = \sum_{k=0}^{\infty} P_\theta^k (g - \bar{\alpha}(\theta)\underline{1}),$$

is state-decision value function, the definition of $\bar{h}$, $\bar{G}$ from Theorem 4.8 and the identity $(1-\lambda^k) = (1-\lambda)(1+\cdots+\lambda^{k-1})$, the above expression for the bias can be reduced to

$$(1-\lambda) \sum_{k=0}^{\infty} \lambda^k \left\langle \hat{\phi}_\theta, P_\theta^{k+1}(Q_\theta - \bar{r}(\theta)'\hat{\phi}_\theta) \right\rangle_\theta.$$

Since ψ_θ is orthogonal to the constant $\underline{1}$, Assumption 5.2 implies that each ψ_θ^i can be expressed as a linear combination of the $\hat{\phi}_\theta^i$'s. Furthermore, for a function $f(\cdot,\cdot)$ of state and decision we have

$$P_\theta f = P_\theta \tilde{f},$$

where $\tilde{f}(\cdot)$ is the function defined by

$$\tilde{f}(x) = \mathbf{E}_{\theta,x}[f(x, U_0)].$$

Therefore, the expression for the bias can be rewritten as

$$b_\theta(\lambda) = (1-\lambda) \sum_{k=0}^{\infty} \lambda^k \left\langle \psi_\theta, P_\theta^{k+1}(V_\theta - \bar{r}(\theta)'\tilde{\phi}_\theta) \right\rangle_\theta,$$

where V_θ is the state value function corresponding to policy θ and $\tilde{\phi}$ is given by

$$\tilde{\phi}_\theta(x) = \mathbf{E}_{\theta,x}[\hat{\phi}_\theta(x, U_0)].$$

For future reference, we will call the functions $\tilde{\phi}_\theta$ as the value function component of $\hat{\phi}_\theta$. The following lemma gives a bound on the bias in terms of λ and the error in the value function approximation.

Lemma 5.3. *There exists constants $0 \leq \rho_0 < 1$ and $C > 0$ such that for all*

$$|b_\theta(\lambda)| \leq \frac{C(1-\lambda)}{(1-\rho_0\lambda)} \parallel V_\theta - \bar{r}(\theta)'\tilde{\phi}_\theta \parallel_\theta, \quad \forall \theta, \lambda.$$

Proof. It follows from Lemma 4.21 that there exists constants $\rho_0 < 1$ and $C > 0$ such that

$$\left\| P_\theta^k (V_\theta - \bar{r}(\theta)'\tilde{\phi}_\theta) \right\|_\theta \leq C\rho_0^k \parallel V_\theta - \bar{r}(\theta)'\tilde{\phi}_\theta \parallel_\theta .$$

Therefore, we have

$$\begin{aligned} |b_\theta(\lambda)| &\leq (1-\lambda) \sum_{k=0}^{\infty} \lambda^k \left| \left\langle \psi_\theta, P_\theta^{k+1}(V_\theta - \bar{r}(\theta)'\tilde{\phi}_\theta) \right\rangle_\theta \right| \\ &\leq (1-\lambda) \sum_{k=0}^{\infty} \lambda^k \parallel \psi_\theta \parallel_\theta \left\| P_\theta^{k+1}(V_\theta - \bar{r}(\theta)'\tilde{\phi}_\theta) \right\|_\theta \\ &\leq \frac{C(1-\lambda)}{(1-\rho_0\lambda)} \parallel V_\theta - \bar{r}(\theta)'\tilde{\phi}_\theta \parallel_\theta, \end{aligned}$$

where the second inequality follows from Cauchy-Schwartz inequality. □

The above bound captures the qualitative dependence of the bias on the critic parameter λ, ρ_0 which represents the mixing time of the Markov chain, and the ability of the critic to approximate the state value functions. The bound shows that actor-critic algorithms work well only when one of the following hold:

1. The critic can approximate value functions fairly well.
2. The critic parameter λ is set close to 1.
3. The Markov chain reaches steady state fast.

The mixing time ρ_0 of the Markov chain depends on the basis functions ϕ_θ^i for the critic and value function V_θ (cf. Lemma 4.21).

Let us now study the implications of the formula for the bias on the choice of basis functions for our critic. When $\lambda < 1$, the above discussion implies that for the bias to be less, the value function component of the critic's approximate state-decision value function should be close to the true value function. However, the value function component of the functions $\psi_\theta^j, j = 1, \ldots, n$, are all zero as

$$\mathbf{E}_{\theta,x}[\psi_\theta(x, U_0)] = 0, \quad \forall \theta, x.$$

Therefore, the subspace Φ_θ spanned by the set of basis functions $\phi_\theta^i, i = 1, \ldots, m$ must be strictly larger than the subspace Ψ_θ spanned by the functions $\psi_\theta^j, j = 1, \ldots n$ (Assumption 5.2), and the number of features for the critic m must be strictly greater than the number of actor parameters n. Suppose the first n basis functions for the

critic are $\psi_\theta^i, i = 1, \ldots, n$, and the remaining basis functions are $\tilde{\phi}_\theta^i, i = 1, \ldots, m-n$, which depend only on the state. For such a choice of the features, the update for the critic and the actor can be modified without changing the asymptotic behavior of our actor-critic algorithms. For convenience of notation, let r and $\tilde{r}$ denote the coefficient vectors for ψ_θ and $\tilde{\phi}_\theta$ respectively. Since $P_\theta \psi_\theta$ is zero, the temporal differences d_k in the critic (see Eq. 4.2) can be replaced by

$$g(\hat{X}_k, \hat{U}_k) + \tilde{r}_k' \tilde{\phi}_{\theta_k}(\hat{X}_{k+1}) - \tilde{r}_k' \tilde{\phi}_{\theta_k}(\hat{X}_k) - r_k' \psi_{\theta_k}(\hat{X}_k, \hat{U}_k).$$

Similarly, the actor update can be modified so that it uses only $r_k' \psi_{\theta_k}$ instead of $r_k' \psi_{\theta_k} + \tilde{r}_k' \tilde{\phi}_{\theta_k}$ as the estimate of state-decision value function. Furthermore, since for $l \leq k$,

$$\begin{aligned} \mathbf{E}[\psi_{\theta_k}(\hat{X}_k, \hat{U}_k) \tilde{\phi}_{\theta_l}(\hat{X}_l)] &= \mathbf{E}[\mathbf{E}_{\theta_k, \hat{X}_k}[\psi_{\theta_k}(\hat{X}_k, \hat{U}_k)] \tilde{\phi}_{\theta_l}(\hat{X}_l)] \\ &= 0, \end{aligned}$$

the term $r' \psi_{\theta_k}(\hat{X}_k, \hat{U}_k)$ in d_k can be removed from the update for the parameters $\tilde{r}$. With these modifications, the critic update breaks down into two parts as shown below.

$$\begin{aligned} r_{k+1} &= r_k + \gamma_k (g(\hat{X}_k, \hat{U}_k) + \tilde{r}_k' \tilde{\phi}_{\theta_k}(\hat{X}_{k+1}) - \tilde{r}_k' \tilde{\phi}_{\theta_k}(\hat{X}_k) - r_k' \phi_{\theta_k}(\hat{X}_k, \hat{U}_k)) Z_k \\ \tilde{r}_{k+1} &= \tilde{r}_k + \gamma_k (g(\hat{X}_k, \hat{U}_k) + \tilde{r}_k' \tilde{\phi}_{\theta_k}(\hat{X}_{k+1}) - \tilde{r}_k' \tilde{\phi}_{\theta_k}(\hat{X}_k)) \tilde{Z}_k \end{aligned}$$

where Z_k and $\tilde{Z}_k$ represent the eligibility trace vectors for the feature vectors ψ_θ and $\tilde{\phi}_\theta$ respectively. Therefore, the update for $\tilde{r}$ becomes the usual TD to approximate the value function and the update for r can be thought of as "advantage" learning with the basis functions ψ_θ used for approximating the "advantage"

$$Q_\theta(x, u) - V_\theta(x).$$

For $\lambda = 1$, since the bias is zero no matter what the additional $\tilde{\phi}_\theta$ features are, pure advantage learning is sufficient for convergence of the algorithms whereas for $\lambda < 1$, the critic must learn both the advantages as well as the state values for the actor's gradient estimate to be unbiased. Note that the functions $\psi_\theta^i, i = 1, \ldots, n$, cannot be used as basis functions when they do not satisfy the uniform linear independence assumptions (Assumption 4.7). However, when they do satisfy the assumptions, the actor update can be further modified as follows:

$$\theta_{k+1} = \theta_k - \beta_k \Gamma(r_k) r_k.$$

With this modification, the actor-critic algorithms become quasi-Newton (Bertsekas & Tsitsiklis, 1996) methods which perform better than gradient methods in some cases.

Before the convergence result of our first variant of actor-critic algorithm can be presented, the following assumption on the vector valued function ψ_θ is needed. Recall the function $L(x, u)$ defined in Chapter 2 for average reward problem. We assume

that the function ψ_θ is differentiable with derivatives bounded by L.

Assumption 5.4. *There exists $K > 0$ such that for each (x, u) the map $\theta \mapsto \psi_\theta(x, u)$, is differentiable with*

$$|\nabla \psi_\theta(x, u)| \leq K \tilde{L}(x, u).$$

The converge result for our first variant of actor-critic algorithms is the following:

Theorem 5.5. *(Convergence of Actor-Critic algorithms)*

$$\liminf_k [|\nabla \bar{\alpha}(\theta_k)| - |b_{\theta_k}(\lambda)|] \leq 0 \quad w.p.1.$$

In other words, the sequence $\{\theta_k\}$ of actor parameter values obtained by an actor-critic algorithm visits any neighborhood of the set

$$\{\theta : |\nabla \bar{\alpha}(\theta)| \leq |b_\theta(\lambda)|\}$$

infinitely often.

5.1.1 Convergence Analysis

In this subsection, we present a proof of Theorem 5.5. We start with the following notation. For each $\theta \in \mathbb{R}^n$ and $(x, u) \in \mathbb{X} \times \mathbb{U}$, let

$$H_\theta(x, u) = \psi_\theta(x, u)\phi'_\theta(x, u), \qquad \bar{H}(\theta) = \langle \psi_\theta, \phi'_\theta \rangle_\theta.$$

The recursion for the actor parameter θ can be written as

$$\begin{aligned} \theta_{k+1} &= \theta_k + \beta_k H_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1})(r_k \Gamma(r_k)) \\ &= \theta_k + \beta_k \bar{H}(\theta_k)\left(\bar{r}(\theta_k)\Gamma(\bar{r}(\theta_k))\right) \\ &\quad + \beta_k (H_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) - \bar{H}(\theta_k))(r_k \Gamma(r_k)) \\ &\quad + \beta_k \bar{H}(\theta_k)(r_k \Gamma(r_k) - \bar{r}(\theta_k)\Gamma(\bar{r}(\theta_k))). \end{aligned}$$

Let

$$\begin{aligned} f(\theta) &= \bar{H}(\theta)\bar{r}(\theta), \\ e_k^{(1)} &= (H_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}) - \bar{H}(\theta_k)) r_k \Gamma(r_k), \\ e_k^{(2)} &= \bar{H}(\theta_k)(r_k \Gamma(r_k) - \bar{r}(\theta_k)\Gamma(\bar{r}(\theta_k))). \end{aligned}$$

Using Taylor's series expansion, one can see that

$$\begin{aligned} \bar{\alpha}(\theta_{k+1}) \geq\ & \bar{\alpha}(\theta_k) + \beta_k \Gamma(\bar{r}(\theta)) \nabla \bar{\alpha}(\theta_k) \cdot f(\theta_k) + \beta_k \nabla \bar{\alpha}(\theta_k) \cdot e_k^{(1)} \\ & + \beta_k \nabla \bar{\alpha}(\theta_k) \cdot e_k^{(2)} - C\beta_k^2 \left| H_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1})(r_k \Gamma(r_k)) \right|^2, \end{aligned} \tag{5.4}$$

where C reflects a bound on the Hessian of $\bar{\alpha}(\theta)$.

The following lemma states that all terms except the one involving $f(\cdot)$ are negligible.

Lemma 5.6. *(Convergence of the noise terms)*

1. $\sum_{k=0}^{\infty} \beta_k \nabla\bar{\alpha}(\theta_k) \cdot e_k^{(1)}$ *converges w.p.1.*
2. $\lim_k e_k^{(2)} = 0$ *w.p.1.*
3. $\sum_k \beta_k^2 |H_{\theta_k}(\hat{X}_k, \hat{U}_k) r_k \Gamma(r_k)|^2 < \infty$ *w.p.1.*

Proof. Since r_k is bounded and $\Gamma(\cdot)$ satisfies the condition (5.3), it is easy to see that $r\Gamma(r)$ is bounded and $|r\Gamma(r) - \hat{r}\Gamma(\hat{r})| < C|r - \hat{r}|$ for some constant C. Therefore the proof of Part 1 is similar to the proof of Lemma 3.10. Similarly, the proof of Part 2 follows from identifying $\left|H_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1})(r_k\Gamma(r_k))\right|$ with H_k in Assumption 4.2, the fact that $\bar{H}(\cdot)$ is bounded, and Theorems 4.6 and 4.8. Proof of Part 3 is similar to the proof of Lemma 3.11 as

$$|H_{\theta_k}(\hat{X}_k, \hat{U}_k) r_k \Gamma(r_k)| \leq C\beta_k |H_{\theta_k}(\hat{X}_k, \hat{U}_k)|$$

for some $C > 0$. □

Proof of Theorem 5.5

Since the proof is standard, we will only outline it. For $T > 0$, define a sequence of random variables k_j by

$$k_0 = 0, \qquad k_{j+1} = \min\left\{ k \geq k_j \,\middle|\, \sum_{l=k_j}^{k} \bar{\beta}_k \geq T \right\}, \quad \text{for} \quad j > 0.$$

Then, using Eq. (5.4) and the fact that $f(\theta) = \nabla\bar{\alpha}(\theta) + b_\theta(\lambda)$ we have

$$\bar{\alpha}(\theta_{k_{j+1}}) \geq \bar{\alpha}(\theta_{k_j}) + \sum_{k=k_j}^{k_{j+1}-1} \beta_k \Gamma(\bar{r}(\theta_k)) \left(|\nabla\bar{\alpha}(\theta_k)|^2 - |b_{\theta_k}(\lambda)||\nabla\bar{\alpha}(\theta_k)| \right) + \delta_j,$$

where δ_j is defined as

$$\delta_j = \sum_{k=k_j}^{k_{j+1}-1} \left[\beta_k \nabla\bar{\alpha}(\theta_k) \cdot (e_k^{(1)} + e_k^{(2)}) - C\beta_k^2 \left| H_{\theta_k}(\hat{X}_k, \hat{U}_k) r_k \Gamma(r_k) \right|^2 \right].$$

Lemma 5.6 implies that δ_j goes to zero. The result follows easily.

In the variant presented in this section, the factor λ controls both the accuracy of value function approximation as well as the robustness of the algorithms to errors in value function approximation. In the next section, we present another variant in which these two can be controlled by two separate parameters, one for the critic and one for the actor.

5.2 Actor with Eligibility Traces

In this variant, the critic uses basis functions which depend only on the state and the actor uses eligibility traces. The actor is updated as follows:

$$\theta_{k+1} = \theta_k + \beta_k \Gamma(r_k) \tilde{d}_{k+1} \tilde{Z}_{k+1}$$

where $\tilde{d}_{k+1}$ represents the temporal difference

$$\tilde{d}_{k+1} = g(\hat{X}_{k+1}, \hat{U}_{k+1}) - \alpha_{k+1} + r_k' \phi_{\theta_k}(\hat{X}_{k+2}) - r_k' \phi_{\theta_k}(\hat{X}_{k+1}),$$

and $\tilde{Z}_{k+1}$ represents eligibility traces for the actor. As in the case of the critic, the update of eligibility traces involves a parameter we denote by $0 \leq \tilde{\lambda} \leq 1$. The update of the eligibility vector $\tilde{Z}_k$ for $\tilde{\lambda} = 1$ and $\tilde{\lambda} < 1$ is different and requires different assumptions. Therefore these two cases will be described separately in the following subsections.

5.2.1 $\tilde{\lambda} = 1$

For this variant, we need a state x^* that is hit with positive probability. Therefore we assume the following:

Assumption 5.7. *The set $\mathbb{X}_0$ consists of a single state x^*.*

The eligibility traces are updated as follows:

$$\tilde{Z}_{k+1} = \begin{cases} \tilde{Z}_k + \psi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}), & \text{if} \quad \hat{X}_{k+1} \neq x^*, \\ \psi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}), & \text{otherwise.} \end{cases}$$

Note that the actor with this update of eligibility traces is similar to the actor-only method of (Marbach & Tsitsiklis, 2001) except for the additional term $r_k' \phi_{\theta_k}(\hat{X}_{k+1}) - r_k' \phi_{\theta_k}(\hat{X}_k)$. This term can be thought of as changing the cost function from g to $g + P_\theta(\bar{r}(\theta)'\phi_\theta) - \bar{r}(\theta)'\phi_\theta$ whose average cost is same as that of the former. Therefore, the convergence result (Marbach & Tsitsiklis, 2001) of actor-only methods implies that this variant of actor-critic methods converge no matter what the critic features are.

5.2.2 $\tilde{\lambda} < 1$

In this case, the eligibility traces are updated as:

$$\tilde{Z}_{k+1} = \tilde{Z}_k + \tilde{\lambda} \psi_{\theta_k}(\hat{X}_{k+1}, \hat{U}_{k+1}).$$

Note that when there is no error in the critic's approximation and $\tilde{\lambda} = 0$, the update direction of the actor parameter is

$$(g(\hat{X}_{k+1}, \hat{U}_{k+1}) - \alpha_{k+1} + r_k' \phi_{\theta_k}(\hat{X}_{k+2}) - r_k' \phi_{\theta_k}(\hat{X}_{k+1})) \psi_{\theta_k},$$

and its steady state expectation is given by $\langle Q_{\theta_k} - V_{\theta_k}, \psi_{\theta_k} \rangle_\theta$. Since V_θ depends only on state and

$$\mathbf{E}_{\theta,x}[\psi_\theta(x, U_0)] = 0, \quad \forall x, \theta,$$

it is easy to see that the steady state expectation of the actor's update direction is the gradient direction. The use of eligibility traces mitigates the bias in the update direction contributed by the error in value function approximation. To see this, consider the steady state update direction for the actor when $1 > \tilde{\lambda} > 0$:

$$\sum_{k=0}^{\infty} \tilde{\lambda}^k \left\langle \psi_\theta, P_\theta^k (g - \bar{\alpha}(\theta)\underline{1} + P_\theta(\bar{r}(\theta)'\phi_\theta) - \bar{r}(\theta)'\phi_\theta) \right\rangle_\theta .$$

It is easy to see, through algebraic manipulations, that this is equal to

$$\sum_{k=0}^{\infty} \tilde{\lambda}^k \left\langle \psi_\theta, P_\theta^k (g - \bar{\alpha}(\theta)\underline{1}) \right\rangle_\theta + \langle \psi_\theta, \bar{r}(\theta)'\phi_\theta \rangle_\theta - (1-\tilde{\lambda}) \sum_{k=0}^{\infty} \tilde{\lambda}^k \left\langle \psi_\theta, P_\theta^{k+1} (\bar{r}(\theta)'\hat{\phi}_\theta) \right\rangle_\theta .$$

Therefore, the bias in the direction is given by

$$\sum_{k=0}^{\infty} (1-\tilde{\lambda}^k) \left\langle \psi_\theta, P_\theta^k (g - \bar{\alpha}(\theta)\underline{1}) \right\rangle_\theta - (1-\tilde{\lambda}) \sum_{k=0}^{\infty} \tilde{\lambda}^k \left\langle \psi_\theta, P_\theta^{k+1} (\bar{r}(\theta)'\phi_\theta) \right\rangle_\theta ,$$

which can be reduced to

$$(1-\tilde{\lambda}) \sum_{k=0}^{\infty} \tilde{\lambda}^k \left\langle \psi_\theta, P_\theta^{k+1} (Q_\theta - \bar{r}(\theta)'\hat{\phi}_\theta) \right\rangle_\theta .$$

Since for each θ, $P_\theta Q_\theta = P_\theta V_\theta$ the expression for the bias can be rewritten as

$$(1-\tilde{\lambda}) \sum_{k=0}^{\infty} \tilde{\lambda}^k \left\langle \psi_\theta, P_\theta^{k+1} (V_\theta - \bar{r}(\theta)'\hat{\phi}_\theta) \right\rangle_\theta .$$

Note that the formula for the bias is the same as that of the previous variant except that the bias now depends on the parameter $\tilde{\lambda}$ of the actor rather than λ of the critic. Therefore we use $b_\theta(\tilde{\lambda})$ to denote this bias. Furthermore, Theorem 5.5 applies to these algorithms as well with λ replaced by $\tilde{\lambda}$. One advantage of this method over the previous one is that the parameter $\tilde{\lambda}$ that controls robustness to approximation errors in the value function is different from the λ that controls the accuracy of critic's approximation.

Before closing this section, let us delve into the implications of the formula for the bias in the estimate of the gradient and the upper bound on it. Let $e_\theta(x)$ denote the the error in value function approximation of the critic for state x. Let τ be a geometric random variable independent of everything else with parameter $1-\tilde{\lambda}$. Then

the bias $b_\theta(\tilde{\lambda})$ in the gradient estimate can be written as

$$\begin{aligned} b_\theta(\tilde{\lambda}) &= \mathbf{E}_\theta \left[\psi_\theta(X_0, U_0) e(X_\tau)\right] \\ &= \int \nabla \mu_\theta(x,u) \mathbf{E}_{\theta,x} \left[e_\theta(X_\tau) \middle| U_0 = u\right] \nu(du), \end{aligned}$$

where $\mathbf{E}_\theta$ denotes the expectation with respect to the stationary distribution of the MDP controlled by RSP θ. If $\tilde{\lambda}$ is chosen close to one, the random time τ is large with high probability and therefore $\mathbf{E}_{\theta,x} \left[e_\theta(X_\tau) \middle| U_0 = u\right]$ is very small due to geometric ergodicity. Therefore, the parameter $\tilde{\lambda}$ can be thought of as follows.

- $1 - \tilde{\lambda}$ represents the confidence the user has on the ability of the critic to approximate the value function. Because, if $\tilde{\lambda}$ is set close to one, the bias contributed by the error in the critic's approximation of the value function is negligible.
- $1/\tilde{\lambda}$ denotes the user's estimate of the time for the Markov chain to reach steady state. To see this, note that the bias is small if the Markov chain reaches steady state with high probability in time τ.

Since, both the mixing time and the error in critic's function approximation vary with the policy, the parameter λ can be changed from policy to policy. That is, we can choose λ to be a function of θ and this can potentially improve both the transient and the asymptotic behavior of the actor-critic algorithms. Finally, these variants can also be analysed using the techniques presented in this thesis provided the function $\lambda(\cdot)$ is well behaved.

5.3 Closing Remarks

The convergence result of this chapter is quite weak. The best one can hope for in gradient methods with errors is that the lim sup $|\nabla \bar{\alpha}(\theta_k)|$ is less than the given bound. While such a result is provable (Bertsekas & Tsitsiklis, 2000) for methods without any gradient errors, it requires certain special structure on the noise sequences which is not present in our case in general. However, our results can be strengthened if we assume a priori that the sequence θ_k is suitably bounded and that $\bar{\alpha}(\cdot)$ is well behaved on this bounded set. This, in turn, might be guaranteed if the iterates θ_k are projected back to a compact set whenever they exit this set.

We have only proved the convergence of actor-critic algorithms without eligibility traces. To analyze actors with eligibility traces, we can write them as

$$\theta_{k+1} = \theta_k + \beta_k \Gamma(r_k) H_{\theta_k}(\tilde{Y}_{k+1}) r_k,$$

where $\tilde{Y}_k = (\hat{X}_k, \hat{U}_k, \tilde{Z}_k)$ forms a Markov chain. The rest of the analysis is similar. Note that the definition of temporal differences used in the update of critic and the actor are different. This is just an artifact of the convention adopted in the thesis that the decision at time $k+1$ is generated using policy corresponding to θ_k. This

convention is not at all essential for the working of the algorithms proposed in this thesis. In fact the policy use at time k can correspond to the actor parameter in the near past, that is, to $\theta_{k-\tau}$ where τ is a random variable with certain finite moments. Similarly, the approximation to the value function at time k can be taken to be $r'_{k-\tau_1}\phi_{\theta_{k-\tau_2}}$ where τ_1, τ_2 are random variables with finite moments. This is because, the change in the parameters of the actor and the critic are of the order of the step-sizes employed and the total change in the values of parameters over a random time of finite expectation is negligible.

In this chapter, we considered the actor updates only for the average reward problem. The algorithms for the other criteria differ only in the simulation of $\hat{X}_k, \hat{U}_k$ and the definition of temporal differences d_k or $\tilde{d}_k$. The differences in the simulation were already described in Chapter 4. The definition of temporal differences is standard and can be found, for example, in (Bertsekas & Tsitsiklis, 1996).

Finally, algorithms similar to ours have been proposed for the discounted reward criterion in (Kimura & Kobayashi, 1998). They also arrive at the conclusion that when eligibility traces are used to their maximum extent, the actor's update direction is the gradient of the discounted reward, no matter what the critic's approximation is. However, the reasoning and the gradient formula they base their arguments on are inaccurate.

Chapter 6

Rate of Convergence of Temporal Difference Learning

In this chapter, the rate of convergence of temporal difference (TD) and related algorithms is studied. Only algorithms for finite state autonomous systems are considered here. However, the results can be extended to Markov chains with general state spaces under appropriate conditions. The aim of this chapter is to understand the rate of convergence of TD in general and the issues relevant to actor-critic algorithms in particular.

Consider an aperiodic and irreducible Markov chain $\{X_k\}$ on a finite state space $\mathbb{X}$ with transition probabilities P_{xy} and stationary distribution π. Let $g : \mathbb{X} \to \mathbb{R}$ be a reward function. Let

$$\bar{\alpha} = \sum_x \pi(x) g(x)$$

be the average expected reward and $J : \mathbb{X} \to \mathbb{R}$ be the differential reward function:

$$J(x) = \sum_{k=0}^{\infty} E[g(X_k) - \bar{\alpha} | X_0 = x].$$

TD algorithms approximate J by a linear combination

$$\hat{J}(x; r) = \sum_{i=1}^{m} r^i \phi^i(x)$$

of m linearly independent basis functions $\phi^i, i = 1, \ldots, m$. Alternatively, the vector

$$\phi(x) = (\phi^1(x), \ldots, \phi^m(x))$$

can be thought of as the feature vector corresponding to the state x. In all the variants of TD algorithms, a sequence of approximations $\hat{J}(x; r_k)$ to the differential reward function are obtained, where the approximation at time k depends on the trajectory of the Markov chain up to time k. For several variants of TD, there are results (Van

Roy, 1998; Tsitsiklis & Van Roy, 1997; Tsitsiklis & Van Roy, 1999a; Tsitsiklis & Van Roy, 1999b) showing that the sequence of parameters r_k and therefore the sequence of approximations $\hat{J}(x; r_k)$ converge. A natural figure of merit for such algorithms is the rate at which the the expected distance between the approximations and the limit to which they converge tends to zero. In general, the best rate at which the expected distance decreases is $1/\sqrt{k}$. Therefore, if $\bar{r}$ denotes the limit of the parameters r_k, the figure of merit of these methods is taken to be

$$\lim_k k\mathbf{E}\left[\| r_k'\phi - \bar{r}'\phi \|^2\right],$$

where $\| \cdot \|$ denotes the weighted norm of functions on $\mathbf{X}$ with the weights corresponding to stationary probabilities π. That is, $\| \cdot \|$ denotes the norm corresponding to the inner product defined as follows: for two functions f_1 and f_2 on $\mathbf{X}$,

$$\langle f_1, f_2 \rangle = \sum_x \pi(x) f_1(x) f_2(x).$$

In the next two sections, the figures of merit of two different TD variants; recursive TD and least squares TD are studied. In particular, it is shown that the least squares variants are always better than the recursive variants and that some small modifications to the recursive variants leads to algorithms that are as good as the least squares variants. Therefore, we take the figure of merit of the least squares variants to be intrinsic to TD methods. Bounds are derived on the intrinsic variance of TD that capture qualitatively the effect of the parameter λ on the rate of convergence of TD. Similar analysis is possible for episodic variants and for other objective criteria, which lead to the same qualitative conclusions.

6.1 Recursive TD

The recursive variants of TD store and update the following parameters:

- α_k, the estimate of average reward,
- r_k, the coefficients of the basis functions,
- Z_k, the eligibility trace vector.

The parameters α_k and r_k are updated as

$$\begin{aligned} \alpha_{k+1} &= \alpha_k + \gamma_k(g(X_k) - \alpha_k), \\ r_{k+1} &= r_k + \gamma_k(g(X_k) - \alpha_k + r_k'\phi(X_{k+1}) - r_k'\phi(X_k))Z_k. \end{aligned}$$

where Z_k is different in different variants of TD. In this chapter, we restrict ourselves to the analysis of the variants in which the eligibility traces are given by

$$Z_k = \sum_{i=0}^{k} \lambda^{k-i}\phi(X_i),$$

where $0 < \lambda < 1$ parameterizes the family of recursive variants of TD. However, the analysis can be extended to other variants as well.

The update for α_k does not depend on the update for r_k and is common for all variants of TD. For all practical purposes, α_k can be taken to be

$$\alpha_k = \frac{1}{k} \sum_{l=0}^{k-1} g(X_k)$$

in all methods. The different TD methods differ only in their update of r_k. It is not difficult to see that α_k converges to $\bar{\alpha}$ w.p.1. Therefore, to simplify the analysis, we will only consider a modified iteration for r_k in which α_k is replaced by $\bar{\alpha}$:

$$r_{k+1} = r_k + \gamma_k(g(X_k) - \bar{\alpha} + r_k'\phi(X_{k+1}) - r_k'\phi(X_k))Z_k.$$

However, note that the rate of convergence of r_k in the original iterations depends on the rate at which α_k converges to $\bar{\alpha}$ and therefore a careful analysis should include this dependence. To avoid cumbersome notation, we study the modified instead of the original iterations. The qualitative conclusions we arrive at hold also for the original iterations.

The first step to analyze TD is to see that the triple $W_k = (X_k, X_{k+1}, Z_k)$ is a Markov chain. The modified iteration can be expressed in the form

$$r_{k+1} = r_k + \gamma_k(h(W_k) - G(W_k)r_k), \tag{6.1}$$

where for $w = (x, y, z)$

$$\begin{aligned} h(w) &= z(g(x) - \bar{\alpha}), \\ G(w) &= z(\phi(y)' - \phi(x)'). \end{aligned}$$

It has been shown (Tsitsiklis & Van Roy, 1997; Tsitsiklis & Van Roy, 1999a) that r_k converges to $\bar{r}$, the unique solution of a system of linear equations

$$\bar{G}r = \bar{h},$$

where $\bar{G}$ and $\bar{h}$ are steady state expected values of $G(W_k)$ and $h(W_k)$, respectively.

To compute the figure of merit of TD, we need the following theorem. This theorem can be viewed as an extension of Proposition 4.8 from (Bertsekas & Tsitsiklis, 1996) and as a special case of more general theorems (Benveniste *et al.*, 1990) on Gaussian approximations to recursive algorithms driven by Markov noise.

Theorem 6.1. *Consider a recursive algorithm of the form (6.1). Assume*

1. *$\{W_k\}$ is a Markov chain.*

2. *The step-sizes γ_k are deterministic, nonnegative, nonincreasing and satisfy*

$$\sum_k \gamma_k = \infty, \quad \sum_k \gamma_k^2 < \infty, \quad \textit{and} \quad \lim_k \left(\gamma_{k+1}^{-1} - \gamma_k^{-1}\right) = \bar{\gamma},$$

for some $\bar{\gamma} \geq 0$.

3. *There exists a deterministic bound on the sequences* $\{h(W_k)\}$ *and* $\{G(W_k)\}$.

4. *There exists a vector* $\bar{h}$, *matrix* $\bar{G}$, *scalars* C *and* ρ, *with* $0 \leq \rho < 1$, *such that*

$$\begin{aligned} \left|\mathbf{E}[G(W_k)|W_0 = w] - \bar{G}\right| &\leq C\rho^k, \\ \left|\mathbf{E}[h(W_k)|W_0 = w] - \bar{h}\right| &\leq C\rho^k, \end{aligned}$$

for all k, w.

5. $\bar{G} - \frac{\bar{\gamma}}{2} I$ *is positive definite.*

6. *There exists a summable sequence of matrices* $\{\hat{\Gamma}_l\}$ *such that for the above constants* C, ρ, *we have*

$$\left|\mathbf{E}\left[F_k F'_{k+l} \middle| W_0 = w\right] - \hat{\Gamma}_l\right| \leq C\rho^k,$$

for all k, w, *where*

$$\begin{aligned} F_k &= (h(W_k) - G(W_k)\bar{r}), \\ \bar{r} &= \bar{G}^{-1}\bar{h}. \end{aligned}$$

Then the following hold:

1. r_k *converges to* $\bar{r}$ *w.p.1.*

2. *The sequence* $\{(h(W_k) - G(W_k)\bar{r})\}$ *obeys a central limit theorem with covariance matrix*

$$\Gamma = \hat{\Gamma}_0 + \sum_{l=1}^{\infty} (\hat{\Gamma}_l + \hat{\Gamma}'_l).$$

That is, the sequence

$$\frac{1}{\sqrt{k}} \sum_{l=0}^{k-1} (h(W_k) - G(W_k)\bar{r})$$

converges in distribution to $N(0, \Gamma)$.

3. $\gamma_k^{-1/2}(r_k - \bar{r})$ *converges in distribution to* $N(0, \Sigma)$ *where the matrix* Σ *satisfies*

$$\bar{G}\Sigma + \Sigma\bar{G}' - \bar{\gamma}\Sigma = \Gamma.$$

4. *For* $\Sigma_0 = \bar{G}^{-1}\Gamma\left(\bar{G}'\right)^{-1}$, $\Sigma - \bar{\gamma}\Sigma_0$ *is positive semi-definite.*

Proof. The proof of first conclusion follows from Proposition 4.8 in (Bertsekas & Tsitsiklis, 1996).

Part (3) follows from Theorem 13 on pg. 332 in (Benveniste *et al.*, 1990). Most of the assumptions for Theorem 13 are trivially satisfied as $h(\cdot)$ and $G(\cdot)$ are bounded and $\bar{G}$ is positive definite. The solution to the Poisson equation can be easily constructed using assumptions (4) and (6). To verify (4.5.11) in (Benveniste *et al.*, 1990), note that

$$\begin{aligned}\lim_k \frac{\sqrt{\gamma_k} - \sqrt{\gamma_{k+1}}}{\gamma_{k+1}^{3/2}} &= \frac{(\gamma_k - \gamma_{k+1})}{\gamma_k \gamma_{k+1}} \times \frac{\gamma_k \gamma_{k+1}}{\gamma_{k+1}^{3/2}(\sqrt{\gamma_k} + \sqrt{\gamma_{k+1}})} \\ &= \frac{\bar{\gamma}}{2}.\end{aligned}$$

Similarly, Eq. (4.5.31) in (Benveniste *et al.*, 1990) follows from the fact that

$$\gamma_{k+1}^{-1} - \gamma_k^{-1} = \bar{\gamma}$$

and the assumption (5). For the representation for Γ used in this theorem see Section 4.4 in (Benveniste *et al.*, 1990).

The second part of the theorem follows from the third as the sequence

$$s_k = \frac{1}{k} \sum_{l=0}^{k-1} (h(W_k) - G(W_k)\bar{r}),$$

satisfies

$$s_{k+1} = s_k + \frac{1}{k+1} \left((h(W_k) - G(W_k)\bar{r}) - s_k \right)$$

and

$$(h(W_k) - G(W_k)\bar{r}) = (h(W_k) - \bar{h}) - (G(W_k) - \bar{G})\bar{r}.$$

To see this, note that when $\bar{G}$ is the identity matrix and, $\gamma_k = 1/(k+1)$ then $\bar{\gamma} = 1$ and therefore, $\Sigma = \Gamma$.

Finally, the last conclusion follows from Proposition 4 on pg. 112 in (Benveniste *et al.*, 1990). □

A part of the above theorem has been used to prove the convergence of TD algorithms (Tsitsiklis & Van Roy, 1999a; Tsitsiklis & Van Roy, 1997). The proof proceeds by verifying Assumptions 1-4 of the above theorem. Assumptions 5 and 6 are needed only for conclusions 2-4. The verification of assumption (6) consists of guessing $\hat{\Gamma}_l$ (using some informal calculations) and proving that (6) indeed holds for that $\hat{\Gamma}_l$ using aperiodicity and irreducibility of Markov chain $\{X_k\}$. An exact expression for $\hat{\Gamma}_l$ and bounds on Γ will be given in the next subsection.

6.1.1 Bounds on the Variance of TD

In this section, we will derive bounds on the covariance matrix Γ which then determines the asymptotic covariance of TD (Theorem 6.1). In the context of TD, it is easy to see that

$$h(W_k) - G(W_k)\bar{r} = \bar{d}_k Z_k,$$

where

$$\bar{d}_k = g(X_k) - \bar{\alpha} + \bar{r}'\phi(X_{k+1}) - \bar{r}'\phi(X_k).$$

Consider the following expression for $\hat{\Gamma}_l$:

$$\begin{aligned}
\hat{\Gamma}_l &= \lim_k \mathbf{E}\left[\bar{d}_k \bar{d}_{k+l} Z_k Z'_{k+l}\right] \\
&= \lim_k \mathbf{E}\left[\sum_{i=0}^{k}\sum_{j=0}^{k+l} \lambda^{i+j} \bar{d}_k \bar{d}_{k+l} \phi(X_{k-i})\phi(X_{k+l-j})'\right] \\
&= \lim_k \sum_{i=0}^{k}\sum_{j=0}^{k+l} \lambda^{i+j} \mathbf{E}\left[\bar{d}_k \bar{d}_{k+l} \phi(X_{k-i})\phi(X_{k+l-j})'\right] \\
&= \lim_k \sum_{i=0}^{k}\sum_{j=0}^{k+l} \lambda^{i+j} \mathbf{E}_\pi\left[\bar{d}_k \bar{d}_{k+l} \phi(X_{k-i})\phi(X_{k+l-j})'\right] \\
&= \lim_k \sum_{i=0}^{k}\sum_{j=0}^{i+l} \lambda^{i+j} \mathbf{E}_\pi\left[\bar{d}_k \bar{d}_{k+l} \phi(X_{k-i})\phi(X_{k+l-j})'\right] \\
&\quad + \lim_k \sum_{i=0}^{k}\sum_{j=i+l+1}^{k+l} \lambda^{i+j} \mathbf{E}_\pi\left[\bar{d}_k \bar{d}_{k+l} \phi(X_{k-i})\phi(X_{k+l-j})'\right] \\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{i+l} \lambda^{i+j} \mathbf{E}_\pi\left[\bar{d}_i \bar{d}_{i+l} \phi(X_0)\phi(X_{i+l-j})'\right] \\
&\quad + \sum_{i=0}^{\infty}\sum_{j=i+l+1}^{\infty} \lambda^{i+j} \mathbf{E}_\pi\left[\bar{d}_{j-l} \bar{d}_j \phi(X_{j-i-l})\phi(X_0)'\right],
\end{aligned}$$

where $\mathbf{E}_\pi$ denotes expectation under the steady state distribution. The justification for the fourth of the above equalities is that the approximation by steady state expectation is close for large i and j, and the past is discounted by a factor of λ. Therefore, Assumption 6 can be verified using the above expression for $\hat{\Gamma}_l$.

We now derive a bound on $\hat{\Gamma}_l$ that captures its dependence on λ and the mixing time of the Markov chain. Consider the following calculation for $\hat{\Gamma}_l$.

$$\hat{\Gamma}_l = \lim_k \sum_{i=0}^{k}\sum_{j=0}^{k+l} \lambda^{i+j} \mathbf{E}\left[\bar{d}_k \bar{d}_{k+l} \phi(X_{k-i})\phi(X_{k+l-j})'\right]$$

$$= \lim_k \sum_{i=0}^{k} \sum_{j=l}^{k+l} \lambda^{i+j} \mathbf{E}\left[\bar{d}_k \bar{d}_{k+l} \phi(X_{k-i}) \phi(X_{k+l-j})'\right]$$
$$+ \lim_k \sum_{i=0}^{k} \sum_{j=0}^{l-1} \lambda^{i+j} \mathbf{E}\left[\bar{d}_k \bar{d}_{k+l} \phi(X_{k-i}) \phi(X_{k+l-j})'\right]$$

Since the Markov chain is aperiodic and irreducible, and $\mathbf{E}_\pi[\bar{d}_k] = 0, \quad \forall k$, we have constants C, $\rho < 1$ such that

$$\left|\mathbf{E}_\pi\left[\bar{d}_k \bar{d}_{k+l} \phi(X_i) \phi(X_j)'\right]\right| \leq C\rho^l$$

for $\max(i, j) \leq k + l$. Therefore, the first term in the above expression is bounded by

$$\frac{C\lambda^l \rho^l}{(1-\lambda)^2}.$$

A bound for the second term is given by:

$$\lim_k \sum_{i=0}^{k} \sum_{j=0}^{l-1} \lambda^{i+j} \mathbf{E}\left[\bar{d}_k \bar{d}_{k+l} \phi(X_{k-i}) \phi(X_{k+l-j})'\right]$$
$$\leq \lim_k \sum_{i=0}^{k} \lambda^i \mathbf{E}\left[\bar{d}_k \phi(X_{k-i}) \bar{d}_{k+l} \sum_{j=0}^{l-1} \lambda^j \phi(X_{k+l-j})'\right]$$
$$\leq \lim_k \sum_{i=0}^{k} \lambda^i \mathbf{E}\left[\bar{d}_k \phi(X_{k-i}) \mathbf{E}\left[\bar{d}_{k+l} \sum_{j=0}^{l-1} \lambda^j \phi(X_{k+l-j})' \middle| X_k\right]\right].$$

Since

$$\sum_j \lambda^j \mathbf{E}_\pi[\bar{d}_j \phi(X_0)] = 0,$$

and

$$\left|\mathbf{E}\left[\bar{d}_{k+l} \phi(X_{k+l-j})' \middle| X_k\right] - \mathbf{E}_\pi[\bar{d}_j \phi(X_0)]\right| \leq C\rho^{l-j},$$

we have

$$\left|\mathbf{E}\left[\bar{d}_{k+l} \sum_{j=0}^{l} \lambda^j \phi(X_{k+l-j})' \middle| X_k\right]\right| \leq C \sum_{j=0}^{l-1} \lambda^j \rho^{l-j} + \sum_{j=l}^{\infty} \lambda^j \mathbf{E}_\pi[\bar{d}_j \phi(X_0)]$$
$$\leq C\left[\frac{\lambda^l - \rho^l}{\lambda - \rho} + \frac{(\lambda\rho)^l}{1 - \lambda\rho}\right].$$

Therefore, we have the following bound on $\hat{\Gamma}_l$:

$$C\left[\frac{\lambda^l\rho^l}{(1-\lambda)^2}+\frac{\lambda^l-\rho^l}{(\lambda-\rho)(1-\lambda)}+\frac{(\lambda\rho)^l}{(1-\lambda\rho)(1-\lambda)}\right]$$

Finally, putting together the bounds on $\hat{\Gamma}_l$ for different l, we have the following bound:

$$|\Gamma| \leq C\left[\frac{1}{(1-\lambda)^2(1-\lambda\rho)}+\frac{1}{(1-\lambda)^2(1-\rho)}+\frac{1}{(1-\lambda\rho)^2(1-\lambda)}\right].$$

Note that the above bound suggests that the variance $|\Gamma|$ of the update direction in TD methods increases with λ and the mixing factor ρ of the Markov chain. We will have more to say about these bounds later in this chapter.

6.2 Rate of convergence of LSTD

We have remarked earlier that the parameters in recursive TD converge to a solution of linear equation whose coefficients and constants are steady state expectations of some functions on the state space of the Markov chain. Another conceivable approach to approximate the solution of such equations is to solve the approximate equation

$$G_k\tilde{r}_k = h_k,$$

where G_k and h_k are approximations to steady state expectations. For TD methods, these approximations are given by

$$\begin{aligned}
h_k &= \frac{1}{k+1}\sum_{l=0}^{k+1}(g(X_k)-\tilde{\alpha}_k)Z_k), \\
G_k &= \frac{1}{k+1}\sum_{l=0}^{k+1}Z_k(\phi(X_{k+1})'-\phi(X_k)'), \\
\tilde{\alpha}_k &= \frac{1}{k+1}\sum_{l=0}^{k}g(X_l).
\end{aligned}$$

It is easy to see that $(\tilde{\alpha}_k, \tilde{r}_k)$ converges to $(\bar{\alpha}, \bar{r})$ since (G_k, h_k) converges to $(\bar{G}, \bar{h})$ (Law of Large Numbers (LLN)) and $\bar{G}$ is invertible. This algorithm is the so called Least squares TD (LSTD) (Boyan, 1999). In this section, we study and compare the rate of convergence of LSTD with recursive TD. For a fair comparison with modified TD given by (6.1) we only study a modified LSTD in which $\tilde{\alpha}_k$ is replaced by $\bar{\alpha}$.

Note that the sequence of matrices G_k and the sequence of vectors h_k are the time averages of the evaluation of a function on the sequence of states visited by a Markov chain. It follows from Theorem 6.1 that the sequence $\{(G(W_k), h(W_k)\}$ obeys the law of large numbers and the central limit theorem. Given this, we wish to obtain Gaussian approximations to $\sqrt{k}(\tilde{r}_k - \bar{r})$. To accomplish this, we recall the following

result from (Duflo, 1997).

Theorem 6.2. *Let U_k be a sequence of random variables in $\mathbb{R}^p$ converging in probability to u. Let a_k be a deterministic nonnegative sequence increasing to ∞. Let $\sqrt{a_k}(U_k - u)$ converge in distribution to $N(0, \Gamma)$. Let $f : \mathbb{R}^p \to \mathbb{R}^q$ be a function twice continuously differentiable in a neighborhood of u. Then, denoting the Jacobian of f at u by $\nabla f(u)$ we have*

1. *$f(U_k)$ converges in probability to $f(u)$,*
2. *$\sqrt{a_k}(f(U_k) - f(u))$ converges in distribution to $N(0, \nabla f(u)\Gamma\nabla f(u)')$.*

The above theorem has the following intuitive interpretation. Suppose $\{U_k\}$ is a sequence of random variables converging to u, with $\sqrt{k}(U_k - u)$ converging in distribution to $N(0, \Gamma)$. Then, the limiting distribution of $\sqrt{k}(f(U_k) - f(u))$ is the same as that of its linear part: $\sqrt{k}\nabla f(u)(U_k - u)$.

To apply the above result to obtain the limiting distribution of $\sqrt{k}(\tilde{r}_k - \bar{r})$ let

$$\begin{aligned} U_k &= (G_k, h_k), \\ u &= (\bar{G}, \bar{h}), \\ f(G, h) &= G^{-1}h. \end{aligned}$$

Note that $\tilde{r}_k = f(G_k, h_k)$ and f is infinitely differentiable around $(\bar{G}, \bar{h})$ as $\bar{G}$ is invertible. To compute linearization of f around $(\bar{G}, \bar{h})$ let g_{ij} denote ij^{th} entry of the matrix G and h_i denote i^{th} component of the vector h. The linearization of f is

$$\sum_i \left.\frac{\partial f}{\partial h_i}\right|_{(\bar{G},\bar{h})} (h_i - \bar{h}_i) + \sum \left.\frac{\partial f}{\partial g_{ij}}\right|_{(\bar{G},\bar{h})} (g_{ij} - \bar{g}_{ij}). \tag{6.2}$$

It is easy to see that the first term is $\bar{G}^{-1}(h - \bar{h})$. To compute the second term, note that f satisfies the equation

$$Gf = h.$$

Differentiating with respect to g_{ij} on both sides of the above equation, we have

$$E_{ij}f + G\frac{\partial f}{\partial g_{ij}} = 0,$$

where E_{ij} is the matrix with ij^{th} entry equal to one and all other entries equal to zero. Using this it is easy to see that the second term in (6.2) is

$$-\sum_{ij} \bar{G}^{-1}E_{ij}f(\bar{G}, \bar{h})(g_{ij} - \bar{g}_{ij}) = -\bar{G}^{-1}(G - \bar{G})\bar{r}.$$

Therefore, the linearization of f is

$$\begin{aligned}\bar{G}^{-1}\left[(h-\bar{h})-(G-\bar{G})\bar{r}\right] &= \bar{G}^{-1}(h-G\bar{r})-\bar{G}^{-1}(\bar{h}-\bar{G}\bar{r}) \\ &= \bar{G}^{-1}(h-G\bar{r}).\end{aligned}$$

Therefore, we have the following theorem on the asymptotic distribution of $\sqrt{k}(\tilde{r}_k-\bar{r})$.

Theorem 6.3. *If the sequence $(G(W_k)\bar{r}-h(W_k))$ obeys a central limit theorem with covariance matrix Γ, i.e., if*

$$\frac{1}{\sqrt{k}}\sum_{l=0}^{k}(G(W_k)\bar{r}-h(W_k))$$

converges in distribution to $N(0,\Gamma)$, then the sequence $\sqrt{k}(\tilde{r}_k-\bar{r})$ in LSTD converges in distribution to $N\left(0,\bar{G}^{-1}\Gamma\left(\bar{G}'\right)^{-1}\right)$.

Recall that we denoted this matrix $\bar{G}^{-1}\Gamma\left(\bar{G}'\right)^{-1}$ in Theorem 6.1 by Σ_0. We will use the same notation throughout this chapter.

Although, we stated the above theorem in the context of LSTD, it is easy to see that it holds in a much more general context - in which one is trying to solve a linear equation whose coefficients and constant on r.h.s. are steady state expectations of a process. We now address some issues which are specific to LSTD. The first is the effect of the choice of basis functions on the "rate of convergence of LSTD".

6.2.1 Effect of Features

We compare two instances of LSTD each using a different set of linearly independent basis functions spanning the same subspace. Let $\{\phi_1^i\}$ and $\{\phi_2^i\}$ be the set of basis functions in the first and the second instances respectively. Similarly, let $\tilde{r}_k^{(1)}, Z_k^{(1)}$ and $\tilde{r}_k^{(2)}, Z_k^{(2)}$ be the iterates in the first and second instances respectively. It is easy to see that the estimates of average reward $\{\alpha_k\}$ are the same in both instances. Since the basis functions span the same subspace in the two instances, there exists a non-singular matrix A such that $\phi_2 = A\phi_1$. Therefore, for each k, $Z_k^{(2)} = AZ_k^{(1)}$. Furthermore, $\tilde{r}_k^{(1)}, \tilde{r}_k^{(2)}$ satisfy linear equations

$$G_k^{(1)}\tilde{r}_k^{(1)} = h_k^{(1)}, \quad G_k^{(2)}\tilde{r}_k^{(2)} = h_k^{(2)},$$

where

$$\begin{aligned}G_k^{(2)} &= AG_k^{(1)}A', \\ h_k^{(2)} &= Ah_k^{(1)}.\end{aligned}$$

It is then easy to see that $\tilde{r}_k^{(2)} = (A')^{-1}\tilde{r}_k^{(1)}$ and that

$$\tilde{r}_k^{(2)'}\phi_2 = \tilde{r}_k^{(1)'}A^{-1}A\phi_1 = \tilde{r}_k^{(1)'}\phi_1.$$

Therefore, *the approximation to the differential reward function obtained by LSTD does not depend on the exact basis functions but depends only on the subspace spanned by them.* An important consequence of this observation is that for LSTD, the **limit**

$$\bar{r} = \lim_k r_k' \phi$$

as well as the **asymptotic variance** defined as

$$\lim_k kE[\| \tilde{r}_k' \phi - \bar{r}' \phi \|^2] \tag{6.3}$$

of the sequence of approximations to differential reward function obtained by LSTD depends only on the subspace spanned by the basis functions. Therefore the quantity (6.3) is called the **intrinsic variance of TD**. Note that if the basis functions ϕ^j's are orthonormal, the expression for intrinsic variance of TD reduces to

$$\begin{aligned} \lim_k kE[\| \tilde{r}_k' \phi - \bar{r}' \phi \|^2] &= \lim_k kE[|\tilde{r}_k - \bar{r}|^2] \\ &= \text{tr}(\Sigma_0). \end{aligned}$$

We now compare the variance of recursive TD and LSTD. Consider the asymptotic covariance matrix of $\sqrt{k}(r_k - \bar{r})$ which is

$$\lim_k kE[(r_k - \bar{r})(r_k - \bar{r})'] = \Sigma \lim_k k\gamma_k.$$

Assumption (2) of Theorem 6.1 implies that $\gamma_k^{-1} \approx k\bar{\gamma}$ which in turn implies that $\lim_k(k\gamma_k) = \bar{\gamma}^{-1}$. Therefore, the asymptotic covariance of TD is $\bar{\gamma}^{-1}\Sigma$ and is worse than (in the sense of positive semi-definiteness) asymptotic covariance of LSTD Σ_0 (cf. Conclusion 4 of the Theorem 6.1). Since the rate of convergence of LSTD depends only on the subspace spanned by basis functions, it is not possible to make the rate of convergence of TD better than LSTD by choosing a different set of basis (spanning the same subspace) functions for TD.

However, there is another small modification of recursive TD that can make it as good as LSTD, at least in the limit. This is Polyak's averaging (Polyak, 1990) applied to TD. In this method, the average reward estimate α_k is updated using step sizes $\{1/(k+1)\}$ and the coefficient vector r_k of the basis functions is updated using a step size γ_k such that $k\gamma_k \to \infty$. The actual estimate $\hat{r}_k$ of the coefficient vector is taken to be the average of the estimates r_k. In other words, the estimates $\alpha_k, r_k, \hat{r}_k$ are given by

$$\begin{aligned} \alpha_k &= \frac{1}{k+1} \sum_{l=0}^{k} g(X_l), \\ r_{k+1} &= r_k + \gamma_k (g(X_k) - \alpha_k + r_k' \phi(X_{k+1}) - r_k' \phi(X_k)) Z_k, \\ \hat{r}_k &= \frac{1}{k+1} \sum_{l=0}^{k} r_l. \end{aligned}$$

The limiting covariance of the vector $(\alpha_k, \hat{r}_k)$ is the same as that of LSTD. The reason why Polyak's averaging is as good as LSTD is explained in the next chapter. Note that any implementation of LSTD would involve inversion of matrices whereas Polyak's averaging does not involve any matrix inversion. Since only a slight modification of recursive TD yields a method with as good rate of convergence as LSTD, we will take the intrinsic variance of TD to be the common figure of merit for both recursive and least squares variants of TD.

6.2.2 Bounds on the Intrinsic Variance of TD

We will now comment on how the parameter λ and the mixing factor ρ affect the intrinsic variance of TD. Since the intrinsic variance of TD depends only on the subspace spanned by the basis functions, it is no loss of generality to assume that the basis functions are orthonormal. In this case, the intrinsic variance is $\mathrm{tr}(\Sigma_0)$ where $\Sigma_0 = \bar{G}^{-1}\Gamma\left(\bar{G}'\right)^{-1}$. The bounds on Γ were already derived in Subsection 6.1.1. To derive a bound on $\bar{G}^{-1}$, recall the expression for $\bar{G}$ from Chapter 4:

$$\begin{aligned}\bar{G} &= \langle\hat{\phi},\hat{\phi}'\rangle - (1-\lambda)\sum_{k=0}^{\infty}\lambda^k\langle\hat{\phi},P^{k+1}\hat{\phi}'\rangle \\ &= (1-\lambda)\sum_{k=0}^{\infty}\lambda^k\langle\hat{\phi},(I-P^{k+1})\hat{\phi}'\rangle.\end{aligned}$$

Let ρ be the second largest eigenvalue of P. It is well-known that $\rho < 1$ as the Markov chain $\{X_k\}$ is irreducible and aperiodic. Therefore, it is easy to see that for some $C > 0$ we have

$$r'\bar{G}r \geq C\left(\frac{1-\rho}{1-\rho\lambda}\right)|r|^2 \quad \forall r.$$

Using this we have the following upper bound for the intrinsic variance of TD: for some $C_1 > 0$,

$$C_1\left(\frac{(1-\rho\lambda)}{(1-\rho)}\right)^2\left(\frac{1}{(1-\lambda)^2(1-\lambda\rho)}+\frac{1}{(1-\lambda)^2(1-\rho)}+\frac{1}{(1-\lambda\rho)^2(1-\lambda)}\right).$$

The following are some of the qualitative properties of TD suggested by this bound. As expected, the eligibility traces become unstable as λ becomes close to 1 and the variance in this regime is of the order of $(1-\lambda)^{-2}$. Note that the above bounds indicate that $\bar{G}$ decreases as λ decreases and the variance Γ increases with λ. This means that both the noise and "signal" in the update direction increase with λ. However, the overall intrinsic variance also increases with λ as

$$\frac{1-\rho\lambda}{1-\lambda} \geq 1.$$

The rate at which the increase happens is monotonic in the mixing factor ρ of the Markov chain. Therefore, for Markov chains with large ρ (i.e., ρ close to 1), using TD(λ) with λ as close as possible to zero is advantageous in terms of the rate of convergence of TD. However, the bounds on the quality of the limiting approximation (Tsitsiklis & Van Roy, 1999a) show that if no linear combination of the basis functions can represent the value function exactly, then the approximation error can be large for small λ (see (Bertsekas & Tsitsiklis, 1996)). Therefore, there is a trade off between how fast TD converges to its limit and the quality of the limit.

6.3 Closing Remarks

In this chapter, we have studied the asymptotic variance of several variants of TD methods. We have not attempted any study of the transient behavior. There were earlier works (Singh & Dayan, 1998; Kearns & Singh, 2000) that attempted such a study for TD with look-up table representations . While (Singh & Dayan, 1998) considers the mean squared error curves calculated numerically for certain Markov chains, (Kearns & Singh, 2000) studies several variants of TD by deriving inequalities that the error satisfies with very high probability. Among other things, these works argue that for any finite k, the graph of the mean square error (i.e., the "variance" of r_k) versus λ is U-shaped with the minimum mean square error λ decreasing to zero as k goes to infinity. Note that we have not observed the U-dependence of variance on λ as we restricted ourselves to the asymptotics. However, our bounds suggest that $\lambda = 0$ corresponds to the least *asymptotic* variance.

Our analysis is the first step towards understanding TD with function approximation. The major drawback of our approach would be that it gives only asymptotic variance of TD. However, we need a prior estimate of the length of the transient period to determine which of analyses (asymptotic or transient) is relevant to a particular situation. Many other issues related to the rate of convergence of TD remain open. For example, the tightness of bounds established in this chapter needs to be studied with some examples. Moreover, there are many other observations made in (Singh & Dayan, 1998) that lack analytical explanations.

Chapter 7

Rate of Convergence of two-time-scale stochastic approximation

In this chapter, we consider two-time-scale stochastic approximation methods and study their rate of convergence. We characterize the asymptotic variance of these algorithms and compare it with that of single time-scale algorithms. These results are to be used later to study the convergence rate of actor-critic methods. In addition, these results are of independent interest, as they generalize the existing analysis of averaging methods. They are also applicable to the analysis of the rate of convergence of existing two-time-scale algorithms.

7.1 Introduction

Two-time-scale stochastic approximation methods (Borkar, 1996) are recursive algorithms in which some of the components are updated using step-sizes that are very small compared to those of the remaining components. Over the past few years, several such algorithms have been proposed for various applications (Konda & Borkar, 1999; Bhatnagar *et al.*, 1999; Konda & Tsitsiklis, 2000a; Konda & Tsitsiklis, 2000b; Bhatnagar *et al.*, 2000; Baras & Borkar, 1999).

The general setting for two-time-scale algorithms is as follows. Let $f(\theta, r)$ and $g(\theta, r)$ be two unknown functions and let (θ^*, r^*) be the unique solution to the equations

$$f(\theta, r) = 0, \qquad g(\theta, r) = 0. \tag{7.1}$$

The functions $f(\cdot, \cdot)$ and $g(\cdot, \cdot)$ are accessible only by simulating or observing a stochastic system which, given θ and r as input, produces $F(\theta, r, V)$ and $G(\theta, r, W)$. Here, V, W are random variables, representing noise, whose distribution satisfies

$$f(\theta, r) = E[F(\theta, r, V)], \quad g(\theta, r) = E[G(\theta, r, W], \quad \forall \theta, r.$$

Assume that the noise (V, W) in each simulation or observation of the stochastic system is independent of the noise in all other simulations. In other words, assume that we have access to an independent sequence of functions $F(\cdot, \cdot, V_k)$ and $G(\cdot, \cdot, W_k)$. Suppose that for any given θ, the stochastic iteration

$$r_{k+1} = r_k + \gamma_k G(\theta, r_k, W_k) \tag{7.2}$$

is known to converge to some $h(\theta)$. Furthermore, assume that the stochastic iteration

$$\theta_{k+1} = \theta_k + \gamma_k F(\theta_k, h(\theta_k), V_k) \tag{7.3}$$

is known to converge to θ^*. Given this information, we wish to construct an algorithm that solves the system of equations (7.1).

Note that the iteration (7.2) has only been assumed to converge when θ is held fixed. This assumption allows us to fix θ at a current value θ_k, run the iteration (7.2) for a long time, so that r_k becomes approximately equal to $h(\theta_k)$, use the resulting r_k to update θ_k in the direction of $F(\theta_k, r_k, W_k)$, and repeat this procedure. While this is a sound approach, it requires an increasingly large time between successive updates of θ_k. Two-time-scale stochastic approximation methods circumvent this difficulty by using different step sizes $\{\beta_k\}$ and $\{\gamma_k\}$ and update θ_k and r_k, according to

$$\begin{aligned} \theta_{k+1} &= \theta_k + \beta_k F(\theta_k, r_k, V_k), \\ r_{k+1} &= r_k + \gamma_k G(\theta_k, r_k, W_k), \end{aligned}$$

where β_k is very small relative to γ_k. This makes θ_k "quasi-static" compared to r_k and has an effect similar to fixing θ_k and running the iteration (7.2) forever. In turn, θ_k sees r_k as a close approximation of $h(\theta_k)$ and therefore its update looks almost the same as (7.3).

How small should the ratio β_k/γ_k be for the above scheme to work? The answer generally depends on the functions $f(\cdot, \cdot)$ and $g(\cdot, \cdot)$, which are typically unknown. This leads us to consider a safe choice whereby $\beta_k/\gamma_k \to 0$. The subject of this chapter is the convergence rate analysis of the two-time-scale algorithms that result from this choice. We note here that the analysis is significantly different from the case where $\lim_k \beta_k/\gamma_k > 0$, which can be handled using existing techniques.

Two-time-scale algorithms have been proved to converge in a variety of contexts (Borkar, 1996; Konda & Borkar, 1999; Konda & Tsitsiklis, 2000b). However, except for the special case of Polyak's averaging, there are no results on their rate of convergence. The existing analyses (Polyak, 1990; Polyak & Juditsky, 1992; Kushner & Yang, 1993) of Polyak's methods rely on special structure and are not applicable to the more general two-time-scale iterations considered here.

The main result of this chapter is a rule of thumb for calculating the asymptotic covariance of two-time-scale stochastic iterations. For example, consider the special case of linear iterations:

$$\theta_{k+1} = \theta_k + \beta_k(b_1 - A_{11}\theta_k - A_{12}r_k + V_k), \tag{7.4}$$

$$r_{k+1} = r_k + \gamma_k(b_2 - A_{21}\theta_k - A_{22}r_k + W_k). \tag{7.5}$$

We show that the asymptotic covariance matrix of $\beta_k^{-1/2}\theta_k$ is the same as that of $\beta_k^{-1/2}\bar{\theta}_k$, where $\bar{\theta}_k$ evolves according to the single-time-scale stochastic iteration:

$$\begin{aligned} \bar{\theta}_{k+1} &= \bar{\theta}_k + \beta_k(b_1 - A_{11}\bar{\theta}_k - A_{12}\bar{r}_k + V_k), \\ 0 &= b_2 - A_{21}\bar{\theta}_k - A_{22}\bar{r}_k + W_k. \end{aligned}$$

Besides the calculation of the asymptotic covariance of $\beta_k^{-1/2}\theta_k$ (Theorem 7.8), we also establish that the distribution of $\beta_k^{-1/2}(\theta_k - \theta^*)$ converges to a Gaussian with mean zero and with the above asymptotic covariance (Theorem 7.10). We discuss extensions of these results to more general non-linear iterations with Markov noise. There are other possible extensions of these results (such as weak convergence of paths to a diffusion process) which we do not consider here.

Our results also explain why Polyak's averaging is optimal. Again, for the sake of simplicity, consider the linear case. Suppose that we are looking for the solution of the linear system

$$Ar = b$$

in a setting where we only have access to noisy measurements of $b - Ar$. The standard algorithm in this setting is

$$r_{k+1} = r_k + \gamma_k(b - Ar_k + W_k), \tag{7.6}$$

and is known to converge under suitable conditions. (Here, W_k represents zero-mean noise at time k.) In order to improve the rate of convergence, Polyak (Polyak, 1990; Polyak & Juditsky, 1992) suggests using the average

$$\theta_k = \frac{1}{k}\sum_{l=0}^{k-1} r_l \tag{7.7}$$

as an estimate of the solution, instead of r_k. It was shown in (Polyak, 1990) that if $k\gamma_k \to \infty$, the asymptotic covariance of $\sqrt{k}\theta_k$ is $A^{-1}\Gamma\,(A')^{-1}$, where Γ is the covariance of W_k. Furthermore, this asymptotic covariance matrix is known to be optimal (Kushner & Yin, 1997).

The calculation of the asymptotic covariance in (Polyak, 1990) uses the special averaging structure. We provide here an alternative calculation based on our results. Note that θ_k satisfies the recursion

$$\theta_{k+1} = \theta_k + \frac{1}{k+1}(r_k - \theta_k), \tag{7.8}$$

and the iteration (7.6)-(7.8) for r_k and θ_k is a special case of the two-time-scale iteration (7.4)-(7.5), with the correspondence $b_1 = 0$, $A_{11} = I$, $A_{12} = -I$, $V_k = 0$, $b_2 = b$, $A_{21} = 0$, $A_{22} = 0$. Furthermore, the assumption $k\gamma_k \to \infty$ corresponds to our

general assumption $\beta_k/\gamma_k \to 0$.

By applying our rule of thumb to the iteration (7.6)-(7.8), we see that the asymptotic covariance of $\left(\sqrt{k+1}\right)\theta_k$ is same as that of $\left(\sqrt{k+1}\right)\bar{\theta}_k$ where $\bar{\theta}_k$ satisfies

$$\bar{\theta}_{k+1} = \bar{\theta}_k + \frac{1}{k+1}\big(-\bar{\theta}_k + A^{-1}(b+W_k)\big),$$

or

$$\bar{\theta}_k = \frac{1}{k}\sum_{l=0}^{k-1}(A^{-1}b + A^{-1}W_l).$$

It then follows that the covariance of $\sqrt{k}\bar{\theta}_k$ is $A^{-1}\Gamma\left(A'\right)^{-1}$, and we recover the result of (Polyak, 1990; Polyak & Juditsky, 1992).

Finally, we would like to point out the differences between the two-time-scale iterations we study here and those that arise in the study of the tracking ability of adaptive algorithms (see (Benveniste *et al.*, 1990)). There, the slow component represents the movement of underlying system parameters and the fast component represents the user's algorithm. The fast component, *i.e.*, the user's algorithm, does not affect the slow component. In contrast, we consider iterations in which the fast component affects the slow one and vice-versa. Furthermore the relevant figures of merit are different. For example, in (Benveniste *et al.*, 1990) one is mostly interested in the behavior of the fast component, whereas we focus on the asymptotic covariance of the slow component.

The outline of the chapter is as follows. In the next section, we consider linear iterations driven by i.i.d. noise and obtain expressions for the asymptotic covariance of the iterates. In Section 7.3, we provide a brief discussion of transient behavior. In Section 7.4, we compare the convergence rate of two-time-scale and their single-time-scale counterparts. In Section 7.5, we establish asymptotic normality of the iterates and, in Section 7.6, we discuss extensions to the case of nonlinear iterations driven by more general noise sequences.

Before proceeding, we introduce some notation. Throughout the chapter, $|\cdot|$ represents the Euclidean norm of vectors or the induced operator norm of matrices. Furthermore, I and 0 represent identity and null matrices, respectively. We use the abbreviation w.p.1. for "with probability one". We use $c, c_1, c_2, \ldots$ to represent some constants whose values are not important.

7.2 Linear Iterations

In this section, we consider iterations of the form

$$\theta_{k+1} = \theta_k + \beta_k(b_1 - A_{11}\theta_k - A_{12}r_k + V_k), \tag{7.9}$$

$$r_{k+1} = r_k + \gamma_k(b_2 - A_{21}\theta_k - A_{22}r_k + W_k), \tag{7.10}$$

where θ_k is in $\mathbb{R}^n$, r_k is in $\mathbb{R}^m$, and b_1, b_2, A_{11}, A_{12}, A_{21}, A_{22} are vectors and matrices of appropriate dimensions.

Before we present our results, we motivate various assumptions that we will need. The first two assumptions are standard.

Assumption 7.1. *The random variables (V_k, W_k), $k = 0, 1, \ldots$, are independent of r_0, θ_0, and of each other. They have zero mean and common covariance*

$$\begin{aligned} E[V_k V_k'] &= \Gamma_{11}, \\ E[V_k W_k'] &= \Gamma_{12} = \Gamma_{21}', \\ E[W_k W_k'] &= \Gamma_{22}. \end{aligned}$$

Assumption 7.2. *The step-size sequences $\{\gamma_k\}$ and $\{\beta_k\}$ are deterministic, positive, nonincreasing, and satisfy the following:*

1. $\sum_k \gamma_k = \sum_k \beta_k = \infty$.
2. $\beta_k, \gamma_k \to 0$.

The key assumption that the step sizes β_k and γ_k are of different orders of magnitude is subsumed by the following.

Assumption 7.3. *There exists some $\epsilon \geq 0$, such that*

$$\frac{\beta_k}{\gamma_k} \to \epsilon.$$

For the iterations (7.9)-(7.10) to be consistent with the general scheme of two-time-scale stochastic approximation described in the introduction, we need some assumptions on the matrices A_{ij}. In particular, we need iteration (7.10) to converge to $A_{22}^{-1}(b_2 - A_{21}\theta)$, when θ_k is held constant at θ. Furthermore, the sequence θ_k generated by the iteration

$$\theta_{k+1} = \theta_k + \beta_k(b_1 - A_{12}A_{22}^{-1}b_2 - (A_{11} - A_{12}A_{22}^{-1}A_{21})\theta_k + V_k),$$

which is obtained by substituting $A_{22}^{-1}(b_2 - A_{21}\theta_k)$ for r_k in iteration (7.9), should also converge. Our next assumption is needed for the above convergence to take place.

Let Δ be the matrix defined by

$$\Delta = A_{11} - A_{12}A_{22}^{-1}A_{21}. \tag{7.11}$$

Recall that a square matrix A is said to be Hurwitz if the real part of each eigenvalue of A is strictly negative.

Assumption 7.4. *The matrices $-A_{22}$, $-\Delta$ are Hurwitz.*

It is not difficult to show that, under the above assumptions, (θ_k, r_k) converges in mean square to (θ^*, r^*). The objective of this chapter is to capture the rate at which

this convergence takes place. Obviously, this rate depends on the step-sizes β_k, γ_k, and this dependence can be quite complicated in general. The following assumption ensures that the rate of mean square convergence of (θ_k, r_k) to (θ^*, r^*) bears a simple relationship (asymptotically linear) with the step-sizes β_k, γ_k.

Assumption 7.5.

1. *There exists a constant* $\bar{\beta} \geq 0$ *such that*

$$\lim_k (\beta_{k+1}^{-1} - \beta_k^{-1}) = \bar{\beta}.$$

2. *If* $\epsilon = 0$ *then*

$$\lim_k (\gamma_{k+1}^{-1} - \gamma_k^{-1}) = 0.$$

3. *The matrix* $-\left(\Delta - \frac{\bar{\beta}}{2} I\right)$ *is Hurwitz.*

Note that when $\epsilon > 0$, the iterations (7.9)-(7.10) are essentially single-time-scale algorithms and therefore can be analyzed using existing techniques (Nevel'son & Has'minskii, 1973; Kushner & Clark, 1978; Benveniste *et al.*, 1990; Duflo, 1997; Kushner & Yin, 1997). We include this in our analysis as we would like to study the behavior of the rate of convergence as $\epsilon \downarrow 0$. The following is an example of sequences satisfying the above assumption with $\epsilon = 0$, $\bar{\beta} = 1/(\tau_1 \beta_0)$:

$$\gamma_k = \frac{\gamma_0}{(1 + k/\tau_0)^\alpha}, \quad \beta_k = \frac{\beta_0}{(1 + k/\tau_1)}, \quad 0 < \alpha < 1.$$

Let $\theta^* \in \mathbb{R}^m$ and $r^* \in \mathbb{R}^n$ be the unique solution to the system of linear equations

$$\begin{aligned} A_{11}\theta + A_{12} r &= b_1, \\ A_{21}\theta + A_{22} r &= b_2. \end{aligned}$$

For each k, let

$$\hat{\theta}_k = \theta_k - \theta^*, \qquad \hat{r}_k = r_k - A_{22}^{-1}(b_2 - A_{21}\theta_k), \tag{7.12}$$

and

$$\begin{aligned} \Sigma_{11}^k &= \beta_k^{-1} E[\hat{\theta}_k \hat{\theta}_k'], \\ \Sigma_{12}^k &= \left(\Sigma_{21}^k\right)' = \beta_k^{-1} E[\hat{\theta}_k \hat{r}_k'], \\ \Sigma_{22}^k &= \gamma_k^{-1} E[\hat{r}_k \hat{r}_k'], \\ \Sigma^k &= \begin{bmatrix} \Sigma_{11}^k & \Sigma_{12}^k \\ \Sigma_{21}^k & \Sigma_{22}^k \end{bmatrix}. \end{aligned}$$

Our main result is the following.

Theorem 7.6. *Under Assumptions 7.1-7.5, and when the constant ϵ of Assumption 7.3 is sufficiently small, the limit matrices*

$$\Sigma_{11}^{(\epsilon)} = \lim_k \Sigma_{11}^k, \qquad \Sigma_{12}^{(\epsilon)} = \lim_k \Sigma_{12}^k, \qquad \Sigma_{22}^{(\epsilon)} = \lim_k \Sigma_{22}^k \tag{7.13}$$

exist. Furthermore, the matrix

$$\Sigma^{(0)} = \begin{bmatrix} \Sigma_{11}^{(0)} & \Sigma_{12}^{(0)} \\ \Sigma_{21}^{(0)} & \Sigma_{22}^{(0)} \end{bmatrix}$$

is the unique solution to the following system of equations

$$\Delta\Sigma_{11}^{(0)} + \Sigma_{11}^{(0)}\Delta' - \bar{\beta}\Sigma_{11}^{(0)} + A_{12}\Sigma_{21}^{(0)} + \Sigma_{12}^{(0)}A_{12}' = \Gamma_{11}, \tag{7.14}$$

$$A_{12}\Sigma_{22}^{(0)} + \Sigma_{12}^{(0)}A_{22}' = \Gamma_{12}, \tag{7.15}$$

$$A_{22}\Sigma_{22}^{(0)} + \Sigma_{22}^{(0)}A_{22}' = \Gamma_{22}. \tag{7.16}$$

Finally,

$$\lim_{\epsilon \downarrow 0} \Sigma_{11}^{(\epsilon)} = \Sigma_{11}^{(0)}, \quad \lim_{\epsilon \downarrow 0} \Sigma_{12}^{(\epsilon)} = \Sigma_{12}^{(0)}, \quad \lim_{\epsilon \downarrow 0} \Sigma_{22}^{(\epsilon)} = \Sigma_{22}^{(0)}. \tag{7.17}$$

Proof. Let us first consider the case $\epsilon = 0$. The idea of the proof is to study the iteration in terms of transformed variables:

$$\begin{aligned} \tilde{\theta}_k &= \hat{\theta}_k, \\ \tilde{r}_k &= L_k\hat{\theta}_k + \hat{r}_k, \end{aligned} \tag{7.18}$$

for some sequence of $n \times m$ matrices $\{L_k\}$ which we will choose so that *the faster time-scale iteration does not involve the slower time-scale variables.* To see what the sequence $\{L_k\}$ should be, we rewrite the iterations (7.9)-(7.10) in terms of the transformed variables as shown below (see Subsection 7.7.1 for the algebra leading to these equations):

$$\begin{aligned} \tilde{\theta}_{k+1} &= \tilde{\theta}_k - \beta_k\left(B_{11}^k\tilde{\theta}_k + A_{12}\tilde{r}_k\right) + \beta_k V_k, \\ \tilde{r}_{k+1} &= \tilde{r}_k - \gamma_k\left(B_{21}^k\tilde{\theta}_k + B_{22}^k\tilde{r}_k\right) + \gamma_k W_k + \beta_k(L_{k+1} + A_{22}^{-1}A_{21})V_k, \end{aligned} \tag{7.19}$$

where

$$\begin{aligned} B_{11}^k &= \Delta - A_{12}L_k, \\ B_{21}^k &= \frac{L_k - L_{k+1}}{\gamma_k} + \frac{\beta_k}{\gamma_k}(L_{k+1} + A_{22}^{-1}A_{21})B_{11}^k - A_{22}L_k, \\ B_{22}^k &= \frac{\beta_k}{\gamma_k}(L_{k+1} + A_{22}^{-1}A_{21})A_{12} + A_{22}. \end{aligned}$$

We wish to choose $\{L_k\}$ so that B_{21}^k is eventually zero. To accomplish this, we define

the sequence of matrices $\{L_k\}$ by

$$\begin{aligned} L_k &= 0, \quad 0 \leq k \leq k_0, \\ L_{k+1} &= \left(L_k - \gamma_k A_{22} L_k + \beta_k A_{22}^{-1} A_{21} B_{11}^k\right) \left(I - \beta_k B_{11}^k\right)^{-1}, \quad \forall k \geq k_0, \end{aligned} \tag{7.20}$$

so that $B_{21}^k = 0$ for all $k \geq k_0$. For the above recursion to be meaningful, we need $(I - \beta_k B_{11}^k)$ to be non-singular for all $k \geq k_0$. This is handled by Lemma 7.11 in the Appendix, which shows that if k_0 is sufficiently large, then the sequence of matrices $\{L_k\}$ is well defined and also converges to zero.

For every $k \geq k_0$, we define

$$\begin{aligned} \tilde{\Sigma}_{11}^k &= \beta_k^{-1} E[\tilde{\theta}_k \tilde{\theta}_k'], \\ \left(\tilde{\Sigma}_{21}^k\right)' &= \tilde{\Sigma}_{12}^k = \beta_k^{-1} E[\tilde{\theta}_k \tilde{r}_k'], \\ \tilde{\Sigma}_{22}^k &= \gamma_k^{-1} E[\tilde{r}_k \tilde{r}_k']. \end{aligned}$$

Using the transformation (7.18), it is easy to see that

$$\begin{aligned} \tilde{\Sigma}_{11}^k &= \Sigma_{11}^k, \\ \tilde{\Sigma}_{12}^k &= \Sigma_{11}^k L_k' + \Sigma_{12}^k, \\ \tilde{\Sigma}_{22}^k &= \Sigma_{22}^k + \left(\frac{\beta_k}{\gamma_k}\right) \left(L_k \Sigma_{12}^k + \Sigma_{21}^k L_k' + L_k \Sigma_{11}^k L_k'\right). \end{aligned}$$

Since $L_k \to 0$, we obtain

$$\begin{aligned} \lim_k \Sigma_{11}^k &= \lim_k \tilde{\Sigma}_{11}^k, \\ \lim_k \Sigma_{12}^k &= \lim_k \tilde{\Sigma}_{12}^k, \\ \lim_k \Sigma_{22}^k &= \lim_k \tilde{\Sigma}_{12}^k, \end{aligned}$$

provided that the limits exist.

To compute $\lim_k \tilde{\Sigma}_{22}^k$, we use Eq. (7.19), the fact that $B_{21}^k = 0$ for large enough k, the fact that B_{22}^k converges to A_{22}, and some algebra, to arrive at the following recursion for $\tilde{\Sigma}_{22}^k$:

$$\tilde{\Sigma}_{22}^{k+1} = \tilde{\Sigma}_{22}^k + \gamma_k (\Gamma_{22} - A_{22} \tilde{\Sigma}_{22}^k - \tilde{\Sigma}_{22}^k A_{22}' + \delta_{22}^k(\tilde{\Sigma}_{22}^k)), \tag{7.21}$$

where $\delta_{22}^k(\cdot)$ is some matrix-valued affine function (on the space of matrices) such that

$$\lim_k \delta_{22}^k(\Sigma_{22}) = 0, \qquad \text{for all } \Sigma_{22}.$$

Since $-A_{22}$ is Hurwitz, it follows (see Lemma 7.12 in the Appendix) that the limit

$$\lim_k \Sigma_{22}^k = \lim_k \tilde{\Sigma}_{22}^k = \Sigma_{22}^{(0)}$$

exists, and $\Sigma_{22}^{(0)}$ satisfies Eq. (7.16).

Similarly, $\tilde{\Sigma}_{12}^k$ satisfies

$$\tilde{\Sigma}_{12}^{k+1} = \tilde{\Sigma}_{12}^k + \gamma_k(\Gamma_{12} - A_{12}\Sigma_{22}^{(0)} - \tilde{\Sigma}_{12}^k A_{22}' + \delta_{12}^k(\tilde{\Sigma}_{12}^k)) \tag{7.22}$$

where, as before, $\delta_{12}^k(\cdot)$ is an affine function that goes to zero. (The coefficients of this affine function depend, in general, on $\tilde{\Sigma}_{22}^k$, but the important property is that they tend to zero as $k \to \infty$.) Since $-A_{22}$ is Hurwitz, the limit

$$\lim_k \Sigma_{12}^k = \lim_k \tilde{\Sigma}_{12}^k = \Sigma_{12}^{(0)}$$

exists and satisfies Eq. (7.15). Finally, $\tilde{\Sigma}_{11}^k$ satisfies

$$\begin{aligned}\tilde{\Sigma}_{11}^{k+1} &= \tilde{\Sigma}_{11}^k + \beta_k\Big(\Gamma_{11} - A_{12}\Sigma_{21}^{(0)} - \Sigma_{12}^{(0)}A_{12}' - \Delta\tilde{\Sigma}_{11}^k - \tilde{\Sigma}_{11}^k\Delta' \\ &\qquad +\bar{\beta}\tilde{\Sigma}_{11}^k + \delta_{11}^k(\tilde{\Sigma}_{11}^k)\Big),\end{aligned} \tag{7.23}$$

where $\delta_{11}^k(\cdot)$ is some affine function that goes to zero. (Once more, the coefficients of this affine function depend, in general, on $\tilde{\Sigma}_{22}^k$ and $\tilde{\Sigma}_{12}^k$, but they tend to zero as $k \to \infty$.) Since $-\left(\Delta - \frac{\bar{\beta}}{2}I\right)$ is Hurwitz, the limit

$$\lim_k \Sigma_{11}^k = \lim_k \tilde{\Sigma}_{11}^k = \Sigma_{11}^{(0)}$$

exists and satisfies Eq. (7.14).

The above arguments show that for $\epsilon = 0$, the limit matrices in (7.13) exist and satisfy Eqs. (7.14)-(7.16). To complete the proof, we need to show that these limit matrices exist for sufficiently small $\epsilon > 0$ and that the limiting relations (7.17) hold. As this part of the proof uses standard techniques, we will only outline the analysis.

Define for each k,

$$Z_k = \begin{pmatrix} \hat{\theta}_k \\ \hat{r}_k \end{pmatrix}.$$

The linear iterations (7.9)-(7.10) can be rewritten in terms of Z_k as

$$Z_{k+1} = Z_k - \beta_k B_k Z_k + \beta_k U_k.$$

where U_k is a sequence of independent random vectors and $\{B_k\}$ is a sequence of deterministic matrices. Using the assumption that β_k/γ_k converges to ϵ, it can be shown that the sequence of matrices B_k converges to some matrix $B^{(\epsilon)}$ and, similarly, that

$$\lim_k E[U_k U_k'] = \Gamma^{(\epsilon)}$$

for some matrix $\Gamma^{(\epsilon)}$. Furthermore, when $\epsilon > 0$ is sufficiently small, it can be shown that $-\left(B^{(\epsilon)} - \frac{\bar{\beta}}{2} I\right)$ is Hurwitz. It then follows from standard theorems (see for e.g. (Polyak, 1976)) on the asymptotic covariance of stochastic approximation methods, that the limit

$$\lim_k \beta_k^{-1} E[Z_k Z_k']$$

exists and satisfies a *linear* equation whose coefficients depend smoothly on ϵ (the coefficients are infinitely differentiable w.r.t. ϵ). Since the components of the above limit matrix are $\Sigma_{11}^{(\epsilon)}$, $\Sigma_{12}^{(\epsilon)}$ and $\Sigma_{22}^{(\epsilon)}$ modulo some scaling, the latter matrices also satisfy a linear equation which depends on ϵ. The explicit form of this equation is tedious to write down and does not provide any additional insight for our purposes. We note however, that when we set ϵ to zero, this system of equations becomes the same as Eqs. (7.14)-(7.16) . Since Eqs. (7.14)-(7.16) have a unique solution, the system of equations for $\Sigma_{11}^{(\epsilon)}$, $\Sigma_{12}^{(\epsilon)}$ and $\Sigma_{22}^{(\epsilon)}$ also has unique solution for all sufficiently small ϵ. Furthermore, the dependence of the solution on ϵ is smooth because the coefficients are smooth in ϵ. □

Remark 7.7. The transformations used in the above proof are inspired by those used to study singularly perturbed ordinary differential equations (Kokotovic, 1984). However, most of these transformations were time-invariant because the perturbation parameter was constant. In such cases, the matrix L satisfies a static Riccati equation instead of the recursion (7.20). In contrast, our transformations are time-varying because our "perturbation" parameter β_k/γ_k is time-varying.

In most applications, the iterate r_k corresponds to some auxiliary parameters and one is mostly interested in the asymptotic covariance $\Sigma_{11}^{(0)}$ of θ_k. Note that according to Theorem 7.6, the covariance of the auxiliary parameters is of the order of γ_k, whereas the covariance of θ_k is of the order of β_k. With two time scales, one can potentially improve the rate of convergence of θ_k (compared to a single time-scale algorithm) by sacrificing the rate of convergence of the auxiliary parameters. To make such comparisons possible, we need an alternative interpretation of $\Sigma_{11}^{(0)}$, that does not explicitly refer to the system (7.14)-(7.16). This is accomplished by our next result, which provides a useful tool for the design and analysis of two-time-scale stochastic approximation methods.

Theorem 7.8. *The asymptotic covariance matrix $\Sigma_{11}^{(0)}$ of $\beta_k^{-1/2}\theta_k$ is the same as the asymptotic covariance of $\beta_k^{-1/2}\bar{\theta}_k$, where $\bar{\theta}_k$ is generated by*

$$\begin{aligned} \bar{\theta}_{k+1} &= \bar{\theta}_k + \beta_k(b_1 - A_{11}\bar{\theta}_k - A_{12}\bar{r}_k + V_k), \\ 0 &= b_2 - A_{21}\bar{\theta}_k - A_{22}\bar{r}_k + W_k. \end{aligned}$$

In other words,

$$\Sigma_{11}^{(0)} = \lim_k \beta_k^{-1} E[\bar{\theta}_k \bar{\theta}_k'].$$

Proof. We start with Eqs. (7.14)-(7.16) and perform some algebraic manipulations to eliminate $\Sigma_{12}^{(0)}$ and $\Sigma_{22}^{(0)}$. This leads to a single equation for $\Sigma_{11}^{(0)}$, of the form

$$\begin{aligned}\Delta\Sigma_{11}^{(0)} + \Sigma_{11}^{(0)}\Delta' - \bar{\beta}\Sigma_{11}^{(0)} &= \Gamma_{11} - A_{12}A_{22}^{-1}\Gamma_{21} - \Gamma_{12}\left(A_{22}'\right)^{-1}A_{12}' \\ &\quad + A_{12}A_{22}^{-1}\Gamma_{22}\left(A_{22}'\right)^{-1}A_{12}'.\end{aligned}$$

Note that the r.h.s. of the above equation is exactly the covariance of $V_k - A_{12}A_{22}^{-1}W_k$. Therefore, the asymptotic covariance of θ_k is same as the asymptotic covariance of the following stochastic approximation:

$$\bar{\theta}_{k+1} = \bar{\theta}_k + \beta_k(-\Delta\bar{\theta}_k + V_k - A_{12}A_{22}^{-1}W_k).$$

□

Remark. The single-time-scale stochastic approximation procedure in Theorem 7.8 is not implementable when the matrices A_{ij} are unknown. The theorem establishes that two-time-scale stochastic approximation performs as well as if these matrices are known.

7.3 Separation of Time-scales

The results of the previous section show that the asymptotic covariance matrix of $\beta_k^{-1/2}\theta_k$ is independent of the step-size schedule $\{\gamma_k\}$ for the fast iteration if

$$\frac{\beta_k}{\gamma_k} \to 0.$$

In this section, we want to understand, at least qualitatively, the effect of the step-sizes γ_k on the transient behavior. To do this, recall the recursions (7.21)-(7.23) satisfied by the covariance matrices $\tilde{\Sigma}^k$:

$$\begin{aligned}\tilde{\Sigma}_{11}^{k+1} &= \tilde{\Sigma}_{11}^k + \beta_k(\Gamma_{11} - A_{12}\Sigma_{21}^{(0)} - \Sigma_{12}^{(0)}A_{12}' \\ &\quad -\Delta\tilde{\Sigma}_{11}^k - \tilde{\Sigma}_{11}^k\Delta' - \bar{\beta}\tilde{\Sigma}_{11}^k + \delta_{11}^k(\tilde{\Sigma}_{11}^k)), \\ \tilde{\Sigma}_{12}^{k+1} &= \tilde{\Sigma}_{12}^k + \gamma_k(\Gamma_{12} - A_{12}\Sigma_{22}^{(0)} - \tilde{\Sigma}_{12}^k A_{22}' + \delta_{12}^k(\tilde{\Sigma}_{12}^k)), \\ \tilde{\Sigma}_{22}^{k+1} &= \Sigma_{22}^k + \gamma_k(\Gamma_{22} - A_{22}\Sigma_{22}^k - \Sigma_{22}^k A_{22}' + \delta_{22}^k(\Sigma_{22}^k)),\end{aligned}$$

where the $\delta_{ij}^k(\cdot)$ are affine functions that tend to zero as k tends to infinity. Using explicit calculations, it is easy to verify that the error terms δ_{ij}^k are of the form

$$\begin{aligned}\delta_{11}^k &= A_{12}(\tilde{\Sigma}_{21}^k - \Sigma_{21}^{(0)}) + (\tilde{\Sigma}_{12}^k - \Sigma_{12}^{(0)})A_{12}' + O(\beta_k), \\ \delta_{12}^k &= A_{12}(\Sigma_{22}^{(0)} - \tilde{\Sigma}_{22}^k) + O\left(\frac{\beta_k}{\gamma_k}\right), \\ \delta_{22}^k &= O\left(\frac{\beta_k}{\gamma_k}\right).\end{aligned}$$

To clarify the meaning of the above relations, the first one states that the affine function $\delta_{11}^k(\Sigma_{11})$ is the sum of the constant term $A_{12}(\tilde{\Sigma}_{21}^k - \Sigma_{21}^{(0)}) + (\tilde{\Sigma}_{12}^k - \Sigma_{12}^{(0)})A_{12}'$, and another affine function of Σ_{11}^k whose coefficients are proportional to β_k.

The above relations show that the rate at which $\tilde{\Sigma}_{11}^k$ converges to $\Sigma_{11}^{(0)}$ depends on the rate at which $\tilde{\Sigma}_{12}^k$ converges to $\Sigma_{12}^{(0)}$, through the term δ_{11}^k. The rate of convergence of $\tilde{\Sigma}_{12}^k$, in turn, depends on that of $\tilde{\Sigma}_{22}^k$, through the term δ_{12}^k. Since the step-size in the recursions for $\tilde{\Sigma}_{22}^k$ and $\tilde{\Sigma}_{12}^k$ is γ_k, and the error terms in these recursions are proportional to β_k/γ_k, the transients depend on both sequences $\{\gamma_k\}$ and $\{\beta_k/\gamma_k\}$. But each sequence has a different effect. When γ_k is large, instability or large oscillations of r_k are possible. On the other hand, when β_k/γ_k is large, the error terms δ_{ij}^k can be large and can prolong the transient period. Therefore, one would like to have β_k/γ_k decrease to zero quickly, while at the same time avoiding large γ_k. Apart from these loose guidelines, it appears difficult to obtain a characterization of desirable step-size schedules.

7.4 Single Time-scale vs. Two Time-scales

In this section, we compare the optimal asymptotic covariance of $\beta_k^{-1/2}\theta_k$ that can be obtained by a realizable single-time-scale stochastic iteration, with the optimal asymptotic covariance that can be obtained by a realizable two-time-scale stochastic iteration. The optimization is to be carried out over a set of suitable gain matrices that can be used to modify the algorithm, and the optimality criterion to be used is one whereby a matrix covariance matrix Σ is preferable to another covariance matrix $\tilde{\Sigma}$ if $\tilde{\Sigma} - \Sigma$ is nonzero and nonnegative definite.

Recall that Theorem 7.8 established that the asymptotic covariance of a two-time-scale iteration is the same as in a related single-time-scale iteration. However, the related single-time-scale iteration was unrealizable, unless the matrix A is known. In contrast, in this section we compare realizable iterations, that do not require explicit knowledge of A (although knowledge of A would be required in order to select the best possible realizable iteration).

We now specify the classes of stochastic iterations that we will be comparing.

1. We consider two-time-scale iterations of the form

$$\begin{aligned} \theta_{k+1} &= \theta_k + \beta_k G_1(b_1 - A_{11}\theta_k - A_{12}r_k + V_k), \\ r_{k+1} &= r_k + \gamma_k(b_2 - A_{21}\theta_k - A_{22}r_k + W_k). \end{aligned}$$

   Here, G_1 is a gain matrix, which we are allowed to choose in a manner that minimizes the asymptotic covariance of $\beta_k^{-1/2}\theta_k$.

2. We consider single-time scale iterations, in which we have $\gamma_k = \beta_k$, but in which we are allowed to use an arbitrary gain matrix G, in order to minimize the asymptotic covariance of $\beta_k^{-1/2}\theta_k$. Concretely, we consider iterations of the

form

$$\begin{bmatrix} \theta_{k+1} \\ r_{k+1} \end{bmatrix} = \begin{bmatrix} \theta_k \\ r_k \end{bmatrix} + \beta_k G \begin{bmatrix} b_1 - A_{11}\theta_k - A_{12}r_k + V_k \\ b_2 - A_{21}\theta_k - A_{22}r_k + W_k \end{bmatrix}.$$

We then have the following result.

Theorem 7.9. *Under Assumptions 7.1-7.5, and with $\epsilon = 0$, the minimal possible asymptotic covariance of $\beta_k^{-1/2}\theta_k$, when the gain matrices G_1 and G can be chosen freely, is the same for the two classes of stochastic iterations described above.*

Proof. The single-time-scale iteration is of the form

$$Z_{k+1} = Z_k + \beta_k G(b - AZ_k + U_k),$$

where

$$Z_k = \begin{bmatrix} \theta_k \\ r_k \end{bmatrix}, \quad U_k = \begin{bmatrix} V_k \\ W_k \end{bmatrix},$$

and

$$b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

As is well known (Kushner & Yin, 1997), the optimal (in the sense of positive definiteness) asymptotic covariance of $\beta_k^{-1/2}Z_k$ over all possible choices of G is the covariance of $A^{-1}U_k$. We note that the top block of $A^{-1}U_k$ is equal to $\Delta^{-1}(V_k - A_{12}A_{22}^{-1}W_k)$. It then follows that the optimal asymptotic covariance matrix of $\beta_k^{-1/2}\theta_k$ is the covariance of $\Delta^{-1}(V_k - A_{12}A_{22}^{-1}W_k)$.

For the two-time-scale iteration, Theorem 7.8 shows that for any choice of G_1, the asymptotic covariance is the same as for the single-time-scale iteration:

$$\theta_{k+1} = \theta_k + \beta_k G_1(b_1 - \Delta\theta_k + V_k - A_{12}A_{22}^{-1}W_k).$$

From this, it follows that the optimal asymptotic covariance of $\beta_k^{-1/2}\theta_k$ is the covariance of $\Delta^{-1}(V_k - A_{12}A_{22}^{-1}W_k)$, which is the same as for single-time-scale iterations. □

7.5 Asymptotic Normality

In Section 7.2, we showed that $\beta_k^{-1}E[\hat{\theta}_k\hat{\theta}_k']$ converges to $\Sigma_{11}^{(0)}$. The proof techniques used in that section do not extend easily (without stronger assumptions) to the nonlinear case. For this reason, we develop here a different result, namely, the asymptotic normality of $\hat{\theta}_k$, which is easier to extend to the nonlinear case. In particular, we show that the distribution of $\beta_k^{-1/2}\hat{\theta}_k$ converges to a zero-mean normal distribution with covariance matrix $\Sigma_{11}^{(0)}$. The proof is similar to the one presented in (Polyak, 1990) for stochastic approximation with averaging.

Theorem 7.10. *If Assumptions 7.1-7.5 hold with $\epsilon = 0$, then $\beta_k^{-1/2}\hat{\theta}_k$ converges in distribution to $N(0, \Sigma_{11}^{(0)})$.*

Proof. Recall the iterations (7.19) in terms of transformed variables $\tilde{\theta}$ and $\tilde{r}$. Assuming that k is large enough so that $B_{21}^k = 0$, these iterations can be written as

$$\begin{aligned}\tilde{\theta}_{k+1} &= (I - \beta_k\Delta)\tilde{\theta}_k - \beta_k A_{12}\tilde{r}_k + \beta_k V_k + \beta_k\delta_k^{(1)},\\ \tilde{r}_{k+1} &= (I - \gamma_k A_{22})\tilde{r}_k + \gamma_k W_k + \beta_k\delta_k^{(2)} + \beta_k(L_{k+1} + A_{22}^{-1}A_{21})V_k,\end{aligned}$$

where $\delta_k^{(1)}$ and $\delta_k^{(2)}$ are given by

$$\begin{aligned}\delta_k^{(1)} &= A_{12}L_k\tilde{\theta}_k,\\ \delta_k^{(2)} &= -(L_{k+1} + A_{22}^{-1}A_{21})A_{12}\tilde{r}_k.\end{aligned}$$

Using Theorem 7.6, $E[|\tilde{\theta}_k|^2]/\beta_k$ and $E[|\tilde{r}_k|^2]/\gamma_k$ are bounded, which implies that

$$\begin{aligned}E[|\delta_k^{(1)}|^2] &\leq c\beta_k|L_k|^2,\\ E[|\delta_k^{(2)}|^2] &\leq c\gamma_k,\end{aligned} \tag{7.24}$$

for some constant $c > 0$. Without loss of generality assume $k_0 = 0$ in (7.19). For each i, define the sequence of matrices Θ_j^i and R_j^i, $j \geq i$ as

$$\begin{aligned}\Theta_i^i &= I,\\ \Theta_{j+1}^i &= \Theta_j^i - \beta_j\Delta\Theta_j^i, \quad \forall j \geq i,\\ R_i^i &= I,\\ R_{j+1}^i &= R_j^i - \gamma_j A_{22}R_j^i, \quad \forall j \geq i.\end{aligned}$$

Using the above matrices, $\tilde{r}_k$ and $\tilde{\theta}_k$ can be rewritten as

$$\tilde{\theta}_k = \Theta_k^0\tilde{\theta}_0 - \sum_{i=0}^{k-1}\beta_i\Theta_k^i A_{12}\tilde{r}_i + \sum_{i=0}^{k-1}\beta_i\Theta_k^i V_i + \sum_{i=0}^{k-1}\beta_i\Theta_k^i\delta_i^{(1)}, \tag{7.25}$$

and

$$\tilde{r}_k = R_k^0\tilde{r}_0 + \sum_{i=0}^{k-1}\gamma_i R_k^i W_i + \sum_{i=0}^{k-1}\beta_i R_k^i\delta_i^{(2)} + \sum_{i=0}^{k-1}\beta_i R_k^i(L_{i+1} + A_{22}^{-1}A_{21})V_i. \tag{7.26}$$

Substituting the r.h.s. of Eq. (7.26) for $\tilde{r}_k$ in Eq. (7.25), and dividing by $\beta_k^{1/2}$, we have

$$\begin{aligned}\beta_k^{-1/2}\tilde{\theta}_k &= \frac{1}{\sqrt{\beta_0}}\tilde{\Theta}_k^0\tilde{\theta}_0 + \sum_{i=0}^{k-1}\beta_i\tilde{\Theta}_k^i A_{12}\left(\beta_i^{-1/2}R_i^0\tilde{r}_0\right) + \sum_{i=0}^{k-1}\beta_i\tilde{\Theta}_k^i\left(\beta_i^{-1/2}\delta_i^{(1)}\right)\\ &\quad + S_k^{(1)} + S_k^{(2)} + S_k^{(3)}\end{aligned}$$

$$+\sum_{i=0}^{k-1}\sqrt{\beta_i}\tilde{\Theta}_k^i(V_i+A_{12}A_{22}^{-1}W_i), \tag{7.27}$$

where

$$\begin{aligned}
\tilde{\Theta}_k^i &= \sqrt{\frac{\beta_i}{\beta_k}}\Theta_k^i, \quad \forall k \geq i,\\
S_k^{(1)} &= \sum_{i=0}^{k-1}\beta_i\tilde{\Theta}_k^i A_{12}\left(\beta_i^{-1/2}\sum_{j=0}^{i-1}\beta_j R_i^j\delta_j^{(2)}\right),\\
S_k^{(2)} &= \sum_{i=0}^{k-1}\beta_i\tilde{\Theta}_k^i A_{12}\left(\beta_i^{-1/2}\sum_{j=0}^{i-1}\beta_j R_i^j(L_{j+1}+A_{22}^{-1}A_{21})V_j\right),\\
S_k^{(3)} &= \sum_{i=0}^{k-1}\sqrt{\beta_i}\tilde{\Theta}_k^i A_{12}\sum_{j=0}^{i-1}\gamma_j R_i^j W_j - \sum_{j=0}^{k-1}\sqrt{\beta_j}\tilde{\Theta}_k^j A_{12}A_{22}^{-1}W_j.
\end{aligned}$$

We wish to prove that the various terms in Eq. (7.27), with the exception of the last one, converge in probability to zero. Note that the last term is a martingale and therefore, can be handled by appealing to a central limit theorem for martingales. Some of the issues we encounter in the remainder of the proof are quite standard, and in such cases we will only provide an outline.

To better handle each the various terms in Eq. (7.27), we need approximations of Θ_k^i and R_k^i. To do this, consider the nonlinear map $A \mapsto \exp(A)$ from square matrices to square matrices. A simple application of the inverse function theorem shows that this map is a diffeomorphism (differentiable, one-to-one with differentiable inverse) in a neighborhood of the origin. Let us denote the inverse of $\exp(\cdot)$ by $\ln(\cdot)$. Since $\ln(\cdot)$ is differentiable around $I = \exp(0)$, the function $\epsilon \mapsto \ln(I - \epsilon A)$ can be expanded into Taylor's series for sufficiently small ϵ as follows:

$$\ln(I-\epsilon A) = -\epsilon(A - E(\epsilon)),$$

where commutes with A and $\lim_{\epsilon\to 0} E(\epsilon) = 0$. Assuming, without loss of generality, that γ_0 and β_0 are small enough for the above approximation to hold, we have for $k \geq 0$,

$$\begin{aligned}
\Theta_k^i &= \exp\left(-\sum_{j=i}^{k-1}\beta_j(\Delta - E_j^{(1)})\right),\\
R_k^i &= \exp\left(-\sum_{j=i}^{k-1}\gamma_j(A_{22} - E_j^{(2)})\right),
\end{aligned} \tag{7.28}$$

for some sequence of matrices $\{E_k^{(i)}\}, i = 1, 2$, converging to zero. To obtain a similar

representation for $\tilde{\Theta}_k^i$, note that Assumption 7.5(1) implies

$$\frac{\beta_k}{\beta_{k+1}} = (1 + \beta_k(\varepsilon_k + \bar{\beta})), \tag{7.29}$$

for some $\varepsilon_k \to 0$. Therefore, using the fact that $1 + x = \exp(x(1 - o(x)))$ and Eq. (7.28), we have

$$\tilde{\Theta}_k^i = \exp\left(-\sum_{j=i}^{k-1} \beta_j \left(\left(\Delta - \frac{\bar{\beta}}{2} I\right) - E_j^{(3)}\right)\right), \tag{7.30}$$

for some sequences of matrices $E_k^{(3)}$ converging to zero. Furthermore, it is not difficult to see that the matrices $E_k^{(i)}, i = 1, 2, 3$, commute with the matrices Δ, A_{22} and $\Delta - (\bar{\beta}/2)I$ respectively. Since $-\Delta$, $-\left(\Delta - (\bar{\beta}/2)I\right)$ and $-A_{22}$ are Hurwitz, using standard Lyapunov techniques we have for some constants $c_1, c_2 > 0$,

$$\begin{aligned} \max(|\Theta_k^i|, |\tilde{\Theta}_k^i|) &\leq c_1 \exp\left(-c_2 \sum_{j=i}^{k-1} \beta_j\right), \\ |R_k^i| &\leq c_1 \exp\left(-c_2 \sum_{j=i}^{k-1} \gamma_j\right). \end{aligned} \tag{7.31}$$

Therefore it is easy to see that the first term in Eq. (7.27) goes to zero w.p.1. To prove that the second term goes to zero w.p.1., note that $\ln \beta_i \approx \bar{\beta} \sum_{j=0}^{i-1} \beta_j$ (cf. Eq. (7.29)) and therefore for some $c_1, c_2 > 0$

$$\left|\beta_i^{-1/2} R_i^0 \tilde{r}_0\right| \leq c_1 \exp\left(-c_2 \sum_{j=0}^{i-1} (\gamma_j - \frac{\bar{\beta}}{2}\beta_j)\right),$$

which goes to zero as $i \to \infty$ (Assumption 7.3). Therefore, it follows from Lemma 7.13 that the second term also converges to zero w.p.1. Using (7.24) and Lemma 7.13 it is easy to see that the third term in Eq. (7.27) converges in the mean (i.e., in L_1) to zero. Next, consider $E[|S_k^{(1)}|]$. Using (7.24) we have for some positive constants c_1, c_2 and c_3,

$$\begin{aligned} E\left[\left|\beta_i^{-1/2} \sum_{j=0}^{i-1} \beta_j R_j^i \delta_j^{(2)}\right|\right] \\ \leq c_1 \sum_{j=0}^{i-1} \gamma_j \exp\left(-\sum_{l=j}^{i-1} (c_2\gamma_l - c_3\beta_l)\right) \sqrt{\frac{\beta_j}{\gamma_j}}. \end{aligned}$$

Since $\beta_j/\gamma_j \to 0$, Lemma 7.13 implies that $S_k^{(1)}$ converges in the mean to zero. To

study $S_k^{(2)}$, consider

$$E\left[\left|\beta_i^{-1/2}\sum_{j=0}^{i-1}\beta_j R_i^j(L_{j+1}+A_{22}^{-1}A_{21})V_j\right|^2\right].$$

Since the V_k are zero mean i.i.d., the above term is bounded above by

$$c_1\sum_{j=0}^{i-1}\gamma_j\exp\left(-\sum_{l=j}^{i-1}(c_2\gamma_l-c_3\beta_l)\right)\frac{\beta_j}{\gamma_j}$$

for some constants c_1, c_2 and c_3. Lemma 7.13 implies that $S_k^{(2)}$ converges in the mean to zero. Finally, consider $S_k^{(3)}$. By interchanging the order of summation, it can be rewritten as

$$\sum_{j=0}^{k-1}\sqrt{\beta_j}\tilde{\Theta}_k^j\left[\frac{\gamma_j}{\beta_j}\sum_{i=j}^{k-1}\beta_i\left(\Theta_i^j\right)^{-1}A_{12}R_i^j-A_{12}A_{22}^{-1}\right]W_j. \tag{7.32}$$

Since $-A_{22}$ is Hurwitz, we have

$$A_{22}^{-1}=\int_0^\infty\exp(-A_{22}t)dt,$$

and we can rewrite the term inside the brackets in Eq. (7.32) as

$$\begin{aligned}&\sum_{i=j}^{k-1}\gamma_i\left(\frac{\gamma_j\beta_i}{\beta_j\gamma_i}\left(\Theta_i^j\right)^{-1}-I\right)A_{12}R_i^j\\&+A_{12}\left(\sum_{i=j}^{k-1}\gamma_i R_i^j-\int_0^{\sum_{i=j}^{k-1}\gamma_i}\exp(-A_{22}t)dt\right)\\&-A_{12}A_{22}^{-1}\exp\left(-\sum_{i=j}^{k-1}\gamma_i A_{22}\right).\end{aligned}$$

We consider each of these terms separately. To analyze the first term, we wish to obtain an "exponential" representation for $\gamma_j\beta_i/\beta_j\gamma_i$. It is not difficult to see from Assumptions 7.5(1),(2) that

$$\begin{aligned}\frac{\beta_{k+1}}{\gamma_{k+1}}&=\frac{\beta_k}{\gamma_k}(1-\varepsilon_k\gamma_k)\\&=\frac{\beta_k}{\gamma_k}\exp(-\varepsilon_k\gamma_k+O(\varepsilon_k^2\gamma_k^2)).\end{aligned}$$

where $\varepsilon_k\to 0$. Therefore, using Eqs. (7.28), (7.30), and the mean value theorem, we

have

$$\left|\frac{\gamma_j\beta_i}{\beta_j\gamma_i}\left(\Theta_i^j\right)^{-1} - I\right| \leq c_1 \sup_{l\geq j}(\varepsilon_l + \beta_l/\gamma_l)\left(\sum_{l=j}^{i-1}\gamma_l\right)\exp\left(c_2\sum_{l=j}^{i-1}(\varepsilon_l+\beta_l/\gamma_l)\gamma_l\right),$$

which in turn implies, along with Lemma 7.14 (with $p = 1$) and Assumption 7.3, that the first term is bounded in norm by $c\sup_{l\geq j}(\varepsilon_l+\gamma_l/\beta_l)$ for some constant $c > 0$. The second term is the difference between an integral and its Riemannian approximation and therefore is bounded in norm by $c\sup_{l\geq j}\gamma_l$ for some constant $c > 0$. Finally, since $-A_{22}$ is Hurwitz, the norm of the third term is bounded above by

$$c_1\exp\left(-c_2\sum_{i=j}^{k-1}\gamma_i\right)$$

for some constants $c_1, c_2 > 0$. An explicit computation of $E\left[\left|S_k^{(1)}\right|^2\right]$, using the fact that (V_k, W_k) is zero-mean i.i.d., and an application of Lemma 7.13 shows that $S_k^{(1)}$ converges to zero in the mean square. Therefore the distribution of $\beta_k^{-1/2}\tilde{\theta}_k$ converges to asymptotic distribution of the martingale comprising of the remaining terms. To complete the proof, we use the standard central limit theorem for martingales (see (Duflo, 1997)). The key assumption of this theorem is Lindberg's condition which, in our case, boils down to the following: for each $\epsilon > 0$,

$$\lim_k \sum_{i=0}^{k-1} E\left[|X_i^{(k)}|^2 I\{|X_i^{(k)}| \geq \epsilon\}\right] = 0,$$

where I is the indicator function and for each $i < k$,

$$X_i^{(k)} = \sqrt{\beta_i}\tilde{\Theta}_k^i(V_i + A_{12}A_{22}^{-1}W_i).$$

The verification of this assumption is quite standard. □

7.6 Nonlinear Iterations

In previous sections, we studied linear iterations driven by zero-mean i.i.d. noise. In this section, we discuss the extension of the asymptotic normality result to nonlinear iterations driven by Markov noise. We will only present an informal sketch of the extension because the details are quite tedious and technical. Although, we will not attempt to extend the result to the most general case, the iterations we consider here are more general than those considered in earlier sections.

We consider the iterations

$$\begin{aligned}\theta_{k+1} &= \theta_k + \beta_k F(\theta_k, r_k, X_k),\\ r_{k+1} &= r_k + \gamma_k G(\theta_k, r_k, X_k),\end{aligned}$$

which we assume to converge to (θ^*, r^*), where β_k, γ_k are nonnegative step-sizes with the same properties as before, and where $\{X_k\}$ is a stochastic-process evolving on some state space $\mathbf{X}$. For each k, the next state X_{k+1} is generated from X_k, θ_k, r_k using a transition probability kernel P_{θ_k, r_k} from a given family $\{P_{\theta,r}\}$. For each fixed $\theta, r,$, we assume that the Markov chain with transition probability kernel $P_{\theta,r}$ is ergodic, with steady state expectation denoted by $E_{\theta,r}$.

Since $\theta_k \to \theta^*$ and $\beta_k \to 0$, and using the same notation $\hat{\theta}_k = \theta_k - \theta^*$, the limiting distribution of $\beta_k^{-1/2}\hat{\theta}_k$ depends only on the "local" behavior of the algorithm around (θ^*, r^*). To study this local behavior, consider the steady state expectations of update-directions in the above iterations:

$$\begin{aligned} f(\theta, r) &= E_{\theta,r}[F(\theta, r, X_k)], \\ g(\theta, r) &= E_{\theta,r}[G(\theta, r, X_k)]. \end{aligned}$$

Suppose $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ are continuously differentiable around (θ^*, r^*) with

$$\begin{aligned} A_{11} &= -\frac{\partial f}{\partial \theta}(\theta^*, r^*), \\ A_{12} &= -\frac{\partial f}{\partial r}(\theta^*, r^*), \\ A_{21} &= -\frac{\partial g}{\partial \theta}(\theta^*, r^*), \\ A_{22} &= -\frac{\partial g}{\partial r}(\theta^*, r^*), \end{aligned}$$

satisfying Assumption 7.4 of Section 2. Linearizing the iterations around (θ^*, r^*) we have

$$\begin{aligned} \hat{\theta}_{k+1} &= \hat{\theta}_k - \beta_k(A_{11}\hat{\theta}_k + A_{12}\hat{r}_k - V_k - \varepsilon_k^{(1)}), \\ \hat{r}_{k+1} &= \hat{r}_k - \gamma_k(A_{21}\hat{\theta}_k + A_{22}\hat{r}_k - W_k - \varepsilon_k^{(2)}). \end{aligned}$$

Here, V_k, W_k represent the difference between the samples of F, G and their steady state expected values:

$$\begin{aligned} V_k &= f(\theta_k, r_k) - F(\theta_k, r_k, X_k), \\ W_k &= g(\theta_k, r_k) - G(\theta_k, r_k, X_k), \end{aligned}$$

and $\varepsilon_k^{(1)}, \varepsilon_k^{(2)}$ represent the errors due to linearization:

$$\begin{aligned} \varepsilon_k^{(1)} &= f(\theta_k, r_k) + A_{11}\hat{\theta}_k + A_{12}\hat{r}_k, \\ \varepsilon_k^{(2)} &= g(\theta_k, r_k) + A_{21}\hat{\theta}_k + A_{22}\hat{r}_k. \end{aligned}$$

Note that the above iterations are similar to (7.9) and (7.10) except that V_k and W_k are no more zero-mean i.i.d., and we also have the additional error terms $\varepsilon_k^{(1)}, \varepsilon_k^{(2)}$. Since (θ_k, r_k) converges to (θ^*, r^*) and the linear approximation errors are bounded above

by $c(|\hat{\theta}_k|^2 + |\hat{r}_k|^2)$ for (θ_k, r_k) sufficiently close to (θ^*, r^*), it is intuitive that these errors will not contribute to the asymptotic distribution of $\beta_k^{-1/2}\theta_k$. Furthermore, although the sequence $\{(V_k, W_k)\}$ is not i.i.d., it can be decomposed into a martingale difference term and some other terms whose contribution to the asymptotic behavior of θ_k and r_k is, again, negligible (see (Benveniste *et al.*, 1990),(Kushner & Yin, 1997)). Finally, note that the proof of asymptotic normality of Section 3 works even if (V_k, W_k) are martingale differences under appropriate, conditions. Therefore, one can expect that even in the nonlinear case $\beta_k^{-1/2}(\theta_k - \theta^*)$ converges in distribution to $N(0, \Sigma_{11}^{(0)})$ where the asymptotic covariance matrices $\Sigma_{ij}^{(0)}, i, j = 1, 2$, satisfy Eqs. (7.14), (7.15) and (7.16). To figure out what Γ should be, note that, in the linear case, it is the covariance matrix in the central limit theorem (CLT) for the sequence (V_k, W_k). For the case of Markov noise, it is the covariance matrix in the CLT for the sequence $H^*(X_k) = (F^*(X_k), G^*(X_k))$ where

$$\begin{aligned} F^*(\cdot) &= F(\theta^*, r^*, \cdot), \\ G^*(\cdot) &= G(\theta^*, r^*, \cdot). \end{aligned}$$

This covariance matrix has the following explicit representation in terms of the steady state expectation $E^*[\cdot] = E_{\theta^*, r^*}[\cdot]$ (see (Meyn & Tweedie, 1993)):

$$\Gamma = E^*[H^*(X_0)H^*(X_0)'] + \sum_{k=1}^{\infty} E^*[H^*(X_0)H^*(X_k)'] + \sum_{k=1}^{\infty} E^*[H^*(X_k)H^*(X_0)'].$$

An alternative representation for $\Sigma_{11}^{(0)}$, in the spirit of Theorem 7.8, is as the asymptotic covariance of $\bar{\theta}_k$ satisfying the linear iteration:

$$\bar{\theta}_{k+1} = \bar{\theta}_k + \beta_k(-\Delta\bar{\theta}_k + F^*(\bar{X}_k) - A_{12}A_{22}^{-1}G^*(\bar{X}_k)),$$

where $\{\bar{X}_k\}$ is the Markov chain corresponding to (θ^*, r^*) and Δ is given by (7.11).

7.7 Auxiliary Results

This section contains proofs of some auxiliary results used earlier in this chapter.

7.7.1 Verification of Eq. (7.19)

Without loss of generality, assume that $b_1 = b_2 = 0$. Then, $\theta^* = 0$ and

$$\tilde{\theta}_k = \hat{\theta}_k = \theta_k,$$

and, using the definition of $\tilde{r}_k$ [cf. Eqs. (7.12) and (7.18)], we have

$$\begin{aligned} \tilde{r}_k &= L_k\theta_k + \hat{r}_k \\ &= L_k\theta_k + r_k + A_{22}^{-1}A_{21}\theta_k \end{aligned} \tag{7.33}$$

$$= \quad r_k + M_k\theta_k,$$

where

$$M_k = L_k + A_{22}^{-1}A_{21}.$$

To verify the equation for $\tilde{\theta}_{k+1} = \theta_{k+1}$, we use the recursion for θ_{k+1}, to obtain

$$\begin{aligned}
\theta_{k+1} &= \theta_k - \beta_k\big(A_{11}\theta_k + A_{12}r_k - V_k\big) \\
&= \theta_k - \beta_k(A_{11}\theta_k + A_{12}\tilde{r}_k - A_{12}(L_k + A_{22}^{-1}A_{21})\theta_k - V_k\big) \\
&= \theta_k - \beta_k(A_{11}\theta_k - A_{12}A_{22}^{-1}A_{21}\theta_k - A_{12}L_k\theta_k + A_{12}\tilde{r}_k - V_k) \\
&= \theta_k - \beta_k(\Delta\theta_k - A_{12}L_k\theta_k + A_{12}\tilde{r}_k) + \beta_k V_k \\
&= \theta_k - \beta_k(B_{11}^k\theta_k + A_{12}\tilde{r}_k) + \beta_k V_k,
\end{aligned}$$

where the last step made use of the definition $B_{11}^k = \Delta - A_{12}L_k$.

To verify the equation for $\tilde{r}_{k+1}$, we first use the definition (7.33) of $\tilde{r}_{k+1}$, and then the update formulas for θ_{k+1} and r_{k+1}, to obtain

$$\begin{aligned}
\tilde{r}_{k+1} &= r_{k+1} + (A_{22}^{-1}A_{21} + L_{k+1})\theta_{k+1} \\
&= r_k - \gamma_k(A_{21}\theta_k + A_{22}r_k - W_k) + (A_{22}^{-1}A_{21} + L_{k+1})\theta_{k+1} \\
&= r_k - \gamma_k\Big(A_{21}\theta_k + A_{22}\big(\tilde{r}_k - (L_k + A_{22}^{-1}A_{21})\theta_k\big) - W_k\Big) \\
&\quad +(A_{22}^{-1}A_{21} + L_{k+1})\theta_{k+1} \\
&= r_k - \gamma_k(A_{22}\tilde{r}_k - A_{22}L_k\theta_k - W_k) + M_{k+1}\theta_{k+1} \\
&= r_k + M_{k+1}\theta_k - \gamma_k(A_{22}\tilde{r}_k - A_{22}L_k\theta_k - W_k) \\
&\quad -\beta_k M_{k+1}(B_{11}^k\theta_k + A_{12}\tilde{r}_k - V_k) \\
&= r_k + M_k\theta_k - \gamma_k\left[\frac{L_k - L_{k+1}}{\gamma_k} - A_{22}L_k + \frac{\beta_k}{\gamma_k}M_{k+1}B_{11}^k\right]\theta_k + \gamma_k W_k \\
&\quad -\gamma_k\left(A_{22} + \frac{\beta_k}{\gamma_k}M_{k+1}A_{12}\right)\tilde{r}_k + \beta_k M_{k+1}V_k \\
&= \tilde{r}_k - \gamma_k\left(B_{21}^k\tilde{\theta}_k + B_{22}^k\tilde{r}_k\right) + \gamma_k W_k + \beta_k M_{k+1}V_k,
\end{aligned}$$

which is the desired formula.

7.7.2 Convergence of the Recursion (7.20)

Lemma 7.11. *For k_0 sufficiently large, the (deterministic) sequence of matrices $\{L_k\}$ defined by Eq. (7.20) is well defined and converges to zero.*

Proof. The recursion (7.20) can be rewritten, for $k \geq k_0$, as

$$L_{k+1} = (I - \gamma_k A_{22})L_k + \beta_k(A_{22}^{-1}A_{21}B_{11}^k + (I - \gamma_k A_{22})L_k B_{11}^k)(I - \beta_k B_{11}^k)^{-1}, \quad (7.34)$$

which is of the form

$$L_{k+1} = (I - \gamma_k A_{22})L_k + \beta_k D_k(L_k),$$

for a matrix-valued function $D_k(L_k)$ defined in the obvious manner. This function has the following properties. When k_0 is chosen large enough and $k \geq k_0$, the step-size is small enough. As long as $|L_k| \leq 1$, we have $|B_{11}^k| = |\Delta - A_{12}L_k| \leq c$, for some absolute constant c. With β_k small enough, the matrix $I - \beta_k B_{11}^k$ is invertible, and satisfies $|(I - \beta_k B_{11}^k)^{-1}| \leq 2$. With $|B_{11}^k|$ bounded by c, we have

$$|A_{22}^{-1}A_{21}B_{11}^k + (I - \gamma_k A_{22})L_k B_{11}^k| \leq d(1 + |L_k|),$$

for some absolute constant d. To summarize, when k_0 is chosen small enough, and as long as $|L_k| \leq 1$, we have $|D_k(L_k)| \leq 2d$.

Recall now that the sequence L_k is initialized with $L_{k_0} = 0$. The unperturbed iteration $L_{k+1} = (1 - \gamma_k A_{22})L_k$ is stable, because $-A_{22}$ is assumed to be Hurwitz. Using a quadratic Lyapunov function for this unperturbed iteration, and invoking the assumption $\beta_k/\gamma_k \to 0$, a standard induction on k shows that the sequence $|L_k|$ generated by the perturbed iteration $L_{k+1} = (I - \gamma_k A_{22})L_k + \beta_k D_k(L_k)$ is bounded by 1, and that the quadratic Lyapunov function converges to zero, which then implies that L_k converges to zero. □

7.7.3 Linear Matrix Iterations

Consider a linear matrix iteration of the form

$$\Sigma_{k+1} = \Sigma_k + \beta_k(\Gamma - A\Sigma_k - \Sigma_k B + \delta_k(\Sigma_k))$$

for some square matrices A, B, step-size sequence β_k, and sequence of matrix-valued affine functions $\delta_k(\cdot)$. Assume

1. The real parts of the eigenvalues of A are positive and the real parts of the eigenvalues of B are nonnegative. (The roles of A and B can also be interchanged.)

2. β_k is positive and

$$\beta_k \to 0, \qquad \sum_k \beta_k = \infty.$$

3. $\lim_k \delta_k(\cdot) = 0$.

We then have the following standard result whose proof can be found, for example, in (Polyak, 1976).

Lemma 7.12. *For any Σ_0, $\lim_k \Sigma_k = \Sigma^*$ exists and is the unique solution to the equation*

$$A\Sigma + \Sigma B = \Gamma.$$

7.7.4 Convergence of Some Series

We provide here some lemmas that are used in the proof of asymptotic normality. Throughout this subsection, $\{\gamma_k\}$ is a positive sequence such that

1. $\gamma_k \to 0$, and
2. $\sum_k \gamma_k = \infty$.

Furthermore, $\{t_k\}$ is the sequence defined by

$$t_0 = 0, \quad t_k = \sum_{j=0}^{k-1} \gamma_k, \quad k > 0.$$

Lemma 7.13. *For any nonnegative sequence $\{\delta_k\}$ that converges to zero and any $p > 0$, we have*

$$\lim_k \sum_{j=0}^{k} \gamma_j \left(\sum_{i=j}^{k-1} \gamma_i \right)^p \exp\left(-\sum_{i=j}^{k-1} \gamma_i \right) \delta_j = 0. \tag{7.35}$$

Proof. Let $\delta(\cdot)$ be a nonnegative function on $[0, \infty)$ defined by

$$\delta(t) = \delta_k, \quad t_k \le t < t_{k+1}.$$

Then it is easy to see that for any $k_0 > 0$,

$$\sum_{j=k_0}^{k} \gamma_j \left(\sum_{i=j}^{k-1} \gamma_i \right)^p \exp\left(-\sum_{i=j}^{k-1} \gamma_i \right) \delta_j = \int_{t_{k_0}}^{t_k} (t_k - s)^p e^{-(t_k - s)} \delta(s) ds + e_k^{k_0}.$$

where

$$|e_k^{k_0}| \le c \sum_{j=k_0}^{k} \gamma_j^2 \left(\sum_{i=j}^{k-1} \gamma_i \right)^p \exp\left(-\sum_{i=j}^{k-1} \gamma_i \right) \delta_j$$

for some constant $c > 0$. Therefore for k_0 sufficiently large, we have

$$\lim_k \sum_{j=k_0}^{k} \gamma_j \left(\sum_{i=j}^{k-1} \gamma_i \right)^p \exp\left(-\sum_{i=j}^{k-1} \gamma_i \right) \delta_j \le \frac{\lim_t \int_0^t \delta(s)(t-s)^p e^{-(t-s)} ds}{1 - c \sup_{k \ge k_0} \gamma_k}.$$

To calculate the above limit, note that

$$\begin{aligned} \lim_t \left| \int_0^t (t-s)^p e^{-(t-s)} \delta(s) ds \right| &= \lim_t \left| \int_0^t s^p e^{-s} \delta(t-s) ds \right| \\ &\le \lim_t \left(\sup_{s \ge t-T} |\delta(s)| \right) \int_0^T s^p e^{-s} ds \end{aligned}$$

$$
\begin{aligned}
&+ \sup_s |\delta(s)| \int_T^\infty s^p e^{-s} ds \\
= \; & \sup_s |\delta(s)| \int_T^\infty s^p e^{-s} ds.
\end{aligned}
$$

Since T is arbitrary, the above limit is zero. Finally, note that the limit in Eq. (7.35) does not depend on the starting limit of the summation. □

Lemma 7.14. *For each $p > 0$, there exists $K_p > 0$ such that for any $k \geq j \geq 0$,*

$$
\sum_{i=j}^{k} \gamma_i \left(\sum_{l=j}^{i-1} \gamma_l \right)^p \exp\left(- \sum_{l=j}^{i-1} \gamma_l \right) \leq K_p.
$$

Proof. For all j sufficiently large, we have

$$
\sum_{i=j}^{k} \gamma_i \left(\sum_{l=j}^{i-1} \gamma_l \right)^p \exp\left(- \sum_{l=j}^{i-1} \gamma_l \right) \leq \frac{\int_0^{(t_k - t_j)} \tau^p e^{-\tau} d\tau}{1 - c \sup_{l \geq j} \gamma_l}
$$

for some $c \geq 0$. □

7.8 Closing Remarks

There are many ways of studying the rate of convergence of stochastic approximation:

1. One approach is based on the central limit theorem for martingales (Kushner & Clark, 1978; Duflo, 1997).

2. Another approach is to compare the asymptotic behavior of the algorithm with that of a diffusion (Nevel'son & Has'minskii, 1973; Kushner & Huang, 1979; Benveniste *et al.*, 1990).

3. Finally, rates of convergence can be obtained by using large deviation techniques (Kushner, 1984; Dupuis & Kushner, 1985; Dupuis & Kushner, 1987; Dupuis, 1988; Dupuis & Kushner, 1989).

Ours is the first attempt at the study of the rate of convergence of two-time-scale stochastic approximation using any of the above mentioned approaches. Although Polyak's averaging methods (Polyak, 1990; Polyak & Juditsky, 1992) are also two-time-scale iterations, the existing analysis is not extendable to general two-time scale iterations. Similarly, although two-time-iterations arise in the study of the tracking ability (Benveniste *et al.*, 1990) of adaptive algorithms, the iterations encountered in that context have a special structure that may not be present in the general iterations we studied in this chapter.

Finally, our results are contrary to the common belief that the introduction of two-time-scale slows down the convergence of θ_k. This discrepancy is due to the fact

that the two-time-scales are introduced by increasing the speed of evolution of r_k rather than slowing down the evolution of θ_k. This process might slow down the convergence of r_k but does not hamper the possibility of achieving an optimal rate of convergence for θ_k.

Chapter 8

Rate of convergence of Actor-Critic Algorithms

In this chapter, the results of the previous chapter are used to study the rate of convergence of the actor-critic algorithms proposed in this thesis. To simplify the presentation and the analysis, only **episodic variants** of algorithms for **total reward** problems are studied in this chapter. However, it is known that the asymptotic behavior of episodic versions is not too different from that of non-episodic versions (Marbach & Tsitsiklis, 2001; Marbach, 1998) and the qualitative conclusions of this chapter hold for other objective criteria as well.

In episodic versions, the parameters are updated only when the system visits the terminal state. Therefore, the policy used by the actor during the course of a single trajectory (or episode) is constant. Furthermore, the increment by which the parameters are updated is equal to the sum of all individual increments, each corresponding to a transition in the trajectory.

The rate of convergence of our algorithms will be compared with that of known actor-only methods (Williams, 1992; Marbach & Tsitsiklis, 2001; Baxter & Barlett, 1999; Glynn, 1987) which we discuss in the next section. Later, it is shown that the rate of convergence of actor-only and actor-critic algorithms with a TD(1) critic are the same. Therefore, it is argued that, for the rate of convergence of actor-critic algorithms to be better than that of actor-only methods, it is essential that the critic use TD(λ) with $\lambda < 1$ and basis functions to approximate the true value function.

8.1 Actor-only Methods

Like actor-critic methods, actor-only methods also optimize over a parametric family of policies. They too use the same formulas as in Chapter 2 to estimate the gradient of overall reward. However, they do not rely on a critic to estimate state-decision or state value functions. Instead, for each gradient estimate, they obtain estimates of state-decision values from simulation and use these in place of the actual state-decision values in the gradient formulas. Therefore, unlike actor-critic algorithms in which the estimates of state-decision values are stored and updated, these methods

generate a new independent estimate for each update of the policy parameters. For this reason, one suspects that, actor-critic algorithms make better use of simulation information and therefore should converge faster than the corresponding actor-only methods. We will show in the next section that this is not always true.

We consider the following actor-only method for the total reward problem from (Williams, 1992) which is essentially the same as the algorithms in (Marbach & Tsitsiklis, 2001), with appropriate modifications. This actor-only method will be compared to actor-critic algorithms. We consider the optimization of expected total reward over a parametric family of RSPs $\{\mu_\theta; \theta \in \mathbb{R}^n\}$ of the MDP described in Section 2.5.

During the course of the algorithm several trajectories ending in the terminal state t of the MDP are simulated sequentially. After each termination, the starting state of the next trajectory is chosen according to a fixed distribution ξ. Let τ_k denote the time step in which the terminal state t is visited for the k^{th} time. For each k, the k^{th} trajectory, from time τ_k to time τ_{k+1}, is simulated using the policy corresponding to θ_k. At time τ_{k+1}, the parameters of the algorithm are updated as follows:

$$\theta_{k+1} = \theta_k + \beta_k \sum_{l=\tau_k+1}^{\tau_{k+1}-1} g(\hat{X}_l, \hat{U}_l)\hat{Z}_l$$

where for $\tau_k \leq l \leq \tau_{k+1}$,

$$\hat{Z}_l = \sum_{j=\tau_k+1}^{l} \tilde{\lambda}^{l-j} \psi_{\theta_k}(\hat{X}_k, \hat{U}_k), \tag{8.1}$$

where $\tilde{\lambda}$ is an algorithm parameter. This algorithm was proposed in (Marbach & Tsitsiklis, 2001) with $\tilde{\lambda} = 1$. However, this reference also suggests the use of $\tilde{\lambda}$ less than but close to one to improve the rate of convergence of these algorithms at the cost of introducing a small bias into the estimate of the gradient.

It is not difficult to see that for $\tilde{\lambda} = 1$, the increment in the above algorithm is an estimate of the gradient of the total reward corresponding to policy θ_k. In other words, expectation of the update increment given that the policy is θ is exactly the gradient of the total reward $\nabla\bar{\alpha}(\theta)$. The rate of convergence of this algorithm, with $\tilde{\lambda} = 1$, is now studied using some informal calculations. However, these can be easily formalized using the results of (Benveniste *et al.*, 1990).

It is proved in (Marbach & Tsitsiklis, 2001) that the sequence $\{\nabla\bar{\alpha}(\theta_k)\}$ converges to zero. We will also assume that the sequence $\{\theta_k\}$ converges to θ^*. To analyse the algorithm, rewrite it in a form amenable to analysis. Note that the algorithm is of the form

$$\theta_{k+1} = \theta_k + \beta_k \nabla\bar{\alpha}(\theta_k) + \beta_k M_k,$$

where M_k denotes the noise in the estimate of $\nabla\bar{\alpha}(\theta_k)$ (cf. Section 2.5). Since θ_k converges to θ^*, we must have $\nabla\bar{\alpha}(\theta^*) = 0$, and therefore the noise term M_k can be

approximated by the noise in the update increment when the current policy is θ^*, and $\nabla\bar{\alpha}(\theta_k)$ can be approximated by its linear term $-\Delta(\theta_k - \theta^*)$ where $-\Delta$ is the Hessian of $\bar{\alpha}(\cdot)$ at θ^*. In other words, M_k can be approximated by a sequence of i.i.d. estimates of the gradient $\nabla\bar{\alpha}(\theta^*)$ (which is equal to zero). Suppose that θ^* is an isolated local maximum and Δ is a positive definite matrix. Then, the asymptotic variance of θ_k in the above algorithm should be the same as that of the following iteration:

$$\theta_{k+1} = \theta_k - \beta_k\Delta(\theta_k - \theta^*) + M_k, \tag{8.2}$$

where the M_k's are i.i.d. estimates of the gradient of the total reward corresponding to the policy θ^*. It is sufficient for our purposes to characterize the asymptotic variance of actor-only algorithms in such a context.

8.2 Actor-Critic Methods

In this section, we wish to obtain a linear iteration of the form (8.2) which characterizes the rate of convergence of episodic variants of actor-critic methods for the total reward problem described in the previous section. We assume that in the algorithms the policy parameter θ converges to θ^*. Although our convergence analysis does not imply such convergence in general, we make this assumption to simplify the analysis. Furthermore, a fair comparison of actor-only and actor-critic methods is possible only when both converge.

As the episodic variants of the actor-critic methods for total reward problems have not been considered so far, a brief description is given below. The simulation of the MDP for these variants is the same as for actor-only methods. The critic parameters and the actor parameters are updated only after the termination of a trajectory. The critic update is given by:

$$r_{k+1} = r_k + \gamma_k \sum_{l=\tau_k+1}^{\tau_{k+1}-1} d_l\hat{Z}_l$$

where for $\tau_k < l < \tau_{k+1}$, d_l represents the temporal difference at time l:

$$d_l = g(\hat{X}_l, \hat{U}_l) + r_k'\phi_{\theta_k}(\hat{X}_{l+1}, \hat{U}_{l+1}) - r_k'\phi_{\theta_k}(\hat{X}_l, \hat{U}_l),$$

and the $\hat{Z}_l$ represent eligibility traces:

$$\hat{Z}_l = \sum_{j=\tau_k+1}^{l} \lambda^{l-j}\phi_{\theta_k}(\hat{X}_j, \hat{U}_j).$$

The actor update is given by:

$$\theta_{k+1} = \theta_k + \beta_k \Gamma(r_k) \sum_{l=\tau_k+1}^{\tau_k - 1} r'_k \phi_{\theta_k}(\hat{X}_l, \hat{U}_l) \psi_{\theta_k}(\hat{X}_l, \hat{U}_l),$$

where the step-size of the actor β_k, the function Γ and the basis functions ϕ_θ are as described in Section 4.3. To study the rate of convergence of actor-critic methods we use the informal results of Section 7.6.

Note that the actor-critic algorithms can be expressed as

$$\begin{aligned} r_{k+1} &= r_k + \gamma_k (h_k(\theta_k) - G_k(\theta_k) r_k), \\ \theta_{k+1} &= \theta_k + \beta_k \Gamma(r_k) H_k(\theta_k) r_k, \end{aligned}$$

where for each k,

$$\begin{aligned} \mathbf{E}\left[h_k(\theta_k) | \, \theta_l, r_l, l \le k\right] &= \bar{h}(\theta_k), \\ \mathbf{E}\left[G_k(\theta_k) | \, \theta_l, r_l, l \le k\right] &= \bar{G}(\theta_k), \\ \mathbf{E}\left[H_k(\theta_k) | \, \theta_l, r_l, l \le k\right] &= \bar{H}(\theta_k), \end{aligned}$$

for some functions $\bar{h}(\theta)$, $\bar{G}(\theta)$ and $\bar{H}(\theta)$. To simplify analysis, we assume $\Gamma(r) = 1$ in a neighborhood of $\bar{r}(\theta^*)$. To simplify the notation we suppress the dependence of various quantities on θ^* in the following discussion.

Using the informal results of Section 7.6, the asymptotic variance of the above iterations can be seen to be the same as for the single time-scale iteration:

$$\theta_{k+1} = \theta_k - \beta_k \Delta(\theta_k - \theta^*) + \beta_k H_k \bar{r} + \beta_k \bar{H} \bar{G}^{-1} (h_k - G_k \bar{r})$$

where $-\Delta$ is the Jacobian of the steady state update direction $\bar{H}(\theta)\bar{r}(\theta)$ at θ^* and all the quantities that depend on θ_k are evaluated at θ^*.

The first property of the asymptotic variance of θ_k is that it depends only on the subspace spanned by the basis functions of the critic, as we now explain. Notice that the first two terms depend only on the limiting approximation of the state-decision value function obtained by the critic which depends only on the subspace spanned by the basis functions. The term remaining is

$$\bar{H} \bar{G}^{-1} (h_k - G_k \bar{r}).$$

Consider the transformation of features from ϕ to $A\phi$ for some invertible matrix A. Then the matrices H_k, $\bar{H}$, $\bar{G}$, h_k, G_k and $\bar{r}$ are transformed to $H_k A'$, $\bar{H} A'$, $A \bar{G} A'$, $A h_k$, $A G_k A'$ and $(A')^{-1} \bar{r}$. It is now easy to see that the above term is invariant under such transformations.

We will now use this property of actor-critic algorithms to study their rate of convergence when the critic uses TD(1) with features satisfying Assumption 5.2. Assume, without loss of generality, that the first n basis vectors are $\psi's$ and the basis functions $\phi_{n+1}, \ldots, \phi_m$ are orthogonal to $\psi's$. Then the matrix $\bar{G} = \langle \phi, \phi' \rangle$ has the

following block diagonal structure:

$$\bar{G} = \begin{bmatrix} \bar{G}_0 & 0 \\ 0 & \bar{G}_1 \end{bmatrix}$$

where $\bar{G}_0 = \langle \psi, \psi \rangle$. Similarly, the matrix $\bar{H} = \langle \psi, \phi' \rangle$ is of the form

$$\bar{H} = \begin{bmatrix} \bar{G}_0 & 0 \end{bmatrix}.$$

Therefore, we have

$$\bar{H}\bar{G}^{-1} = [I \quad 0],$$

which in turn implies that the above iteration can be rewritten as

$$\theta_{k+1} = \theta_k - \beta_k \Delta(\theta_k - \theta^*) + \beta_k(H_k - \tilde{G}_k)\bar{r} + \beta_k \tilde{h}_k,$$

where $\tilde{h}_k$ and $\tilde{G}_k$ are the first n rows of h_k and G_k. Since the first n components of the feature vectors ϕ form the vector ψ, it is easy to see that $\tilde{h}_k$ is the same as actor-only methods. Using simple algebraic manipulations and the fact that all the features of the terminal state are zeroes (Assumption 4.11), it is also easy to see that $H_k = \tilde{G}_k$. Therefore it follows that the rate of convergence of actor-critic algorithms with TD(1) critic is the same as actor-only methods. Although these results are applicable only to episodic variants, we do not expect the rate of convergence of non-episodic variants to be much different. Therefore, in order to improve the rate of convergence of actor-only methods we have to use TD(λ) with $\lambda < 1$. In the next section, we will illustrate, through a numerical example, that the rate of convergence of actor-critic methods can be substantially better than that of actor-only methods if proper features are used for the critic.

8.3 Numerical Example

In this section, we use a well-known academic example from (Bertsekas & Tsitsiklis, 1996) to numerically study the rate of convergence of actor-critic methods. Our objective is to illustrate various issues involved in the design of actor-critic methods.

A driver is looking for inexpensive parking on the way to his destination. The parking area contains N spaces. The driver starts at space N and traverses the parking spaces sequentially, that is, from space x he goes next to space $x - 1$, etc. The destination corresponds to parking space 0. Each parking space is free with probability p independently of whether other parking spaces are free or not. The driver can observe whether a parking space is free only when he reaches it, and then, if it is free, he makes a decision to park in that space or not. If he parks in space x, he incurs a cost $c(x) > 0$. If he reaches the destination he must park in a garage, which is expensive and costs $C > 0$. The problem is to computationally approximate the optimal policy.

It is easy to prove that the optimal policy for this problem is a threshold policy. That is, there exists an integer x^* such that it is optimal to park at x if and only if $x \leq x^*$. For computational purposes, we formulate the problem as a total reward problem. The state space consists of integers $0, \ldots, N$ and a termination state t where the state x corresponds to the driver being at space x and the space x being free. The decision to park is denoted by u_0 and the decision not to park is denoted by u_1. The reward function for this problem is given by

$$g(x, u_0) = \begin{cases} -c(x) & \text{if} \quad 1 \leq x \leq N, \\ -C & \text{if} \quad x = 0, \\ 0 & \text{otherwise} \end{cases}$$

and $g(x, u_1) = 0, \quad \forall x$. An approximation to a threshold policy with threshold θ is provided by the following randomized stationary policy:

$$\mu_\theta(u_0|x) = \frac{1}{2}\left(1 - \tanh\left(\frac{x-\theta}{T}\right)\right), \quad \mu_\theta(u_1|x) = 1 - \mu_\theta(u_0|x),$$

where the parameter T controls the accuracy of approximation. To compute an approximation to the optimal policy, we can optimize the total reward over the above family of parameterized policies. In order for the problem to be completely defined, we need a probability distribution for the starting state. If we were to capture fully the dynamics of the original problem, we would take the starting state to be $N - \tau$ where τ is a geometric random variable with parameter p. However, for the sake of simplicity we take the starting state to be N.

To illustrate the advantages of actor-critic methods, we consider the case where

$$p = 0.05, \quad N = 200, \quad c(x) = x, \quad 1 \leq x \leq N, \quad C = 100, \quad T = 15.$$

For the problem with these parameter values, Figure 8-1 shows a plot for the total reward as a function of θ using exact computations. The maximum of the total reward occurs at the threshold value of 35.7. It is known from (Bertsekas & Tsitsiklis, 1996) that the optimal policy is a threshold policy with threshold value 35. Note that if the randomized threshold policy with threshold 35.7 is rounded off to a deterministic threshold policy, the resulting policy would be optimal for the original problem. In this case, we were lucky that the optimal solution to the approximate problem is also an optimal solution to the original problem of finding an optimal deterministic threshold policy. The following subsections describe computational results for different algorithms applied to this problem.

8.3.1 Actor-only Methods

The actor-only method described in the Section 8.1 was applied to this problem. The algorithm was tried with two different values for the parameter $\tilde{\lambda}$. In the first case, $\tilde{\lambda}$ was set to 1 and therefore, the algorithm updates the policy parameter in a direction that is an unbiased estimate of the gradient of the total reward. The step sizes β_k

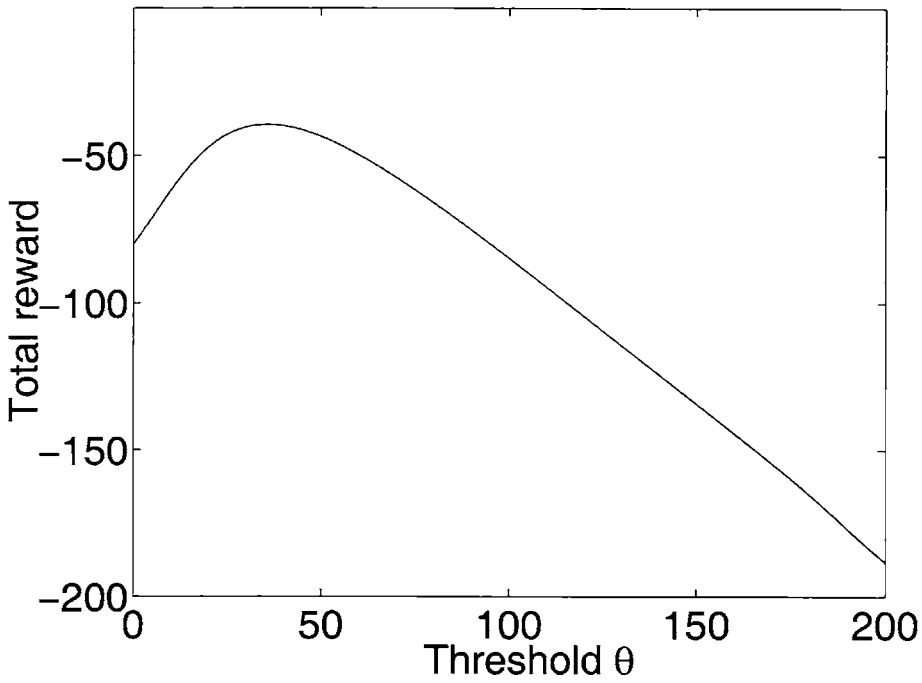

Figure 8-1: Total reward as a function of the threshold parameter.

used were of the form

$$\beta_k = \frac{\beta_0}{\left(1 + \frac{k}{k_a}\right)}, \tag{8.3}$$

where

$$\begin{aligned} \beta_0 &= 0.04, \\ k_a &= 40. \end{aligned}$$

The threshold parameter θ was initialized to 100 and the algorithm was run for 10000 trajectories. Figure 8-2 shows the sequence of thresholds obtained during the course of the algorithm. The figure shows that the threshold parameter "settled" down in a region between 30 and 40 after 2000 iterations and kept moving slowly (due to small-step sizes) until it started to converge around 8000 iterations. Although, it is difficult to say whether the convergence of the algorithm after 8000 trajectories is due to small step-sizes, the key point is that the parameter has not converged till 8000 iterations. The threshold to which the algorithm appears to have converged is 35.92 which is 0.22 away from the optimal threshold.

In the second case, $\tilde{\lambda}$ was set to 0.7 and the rest of the parameters were left unchanged. In (Marbach, 1998; Marbach & Tsitsiklis, 2001), choosing a value less than 1 for $\tilde{\lambda}$ was suggested to reduce the variance in the actor-only algorithm. The plot (Figure 8-3) shows that setting $\tilde{\lambda}$ to 0.7 has indeed led to faster convergence of

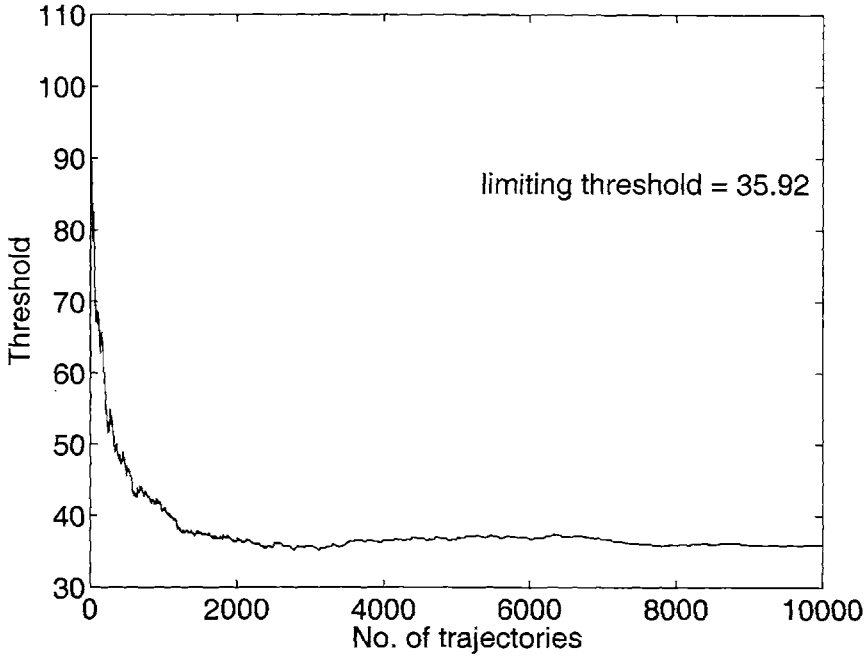

Figure 8-2: Actor-only method with $\tilde{\lambda} = 1$.

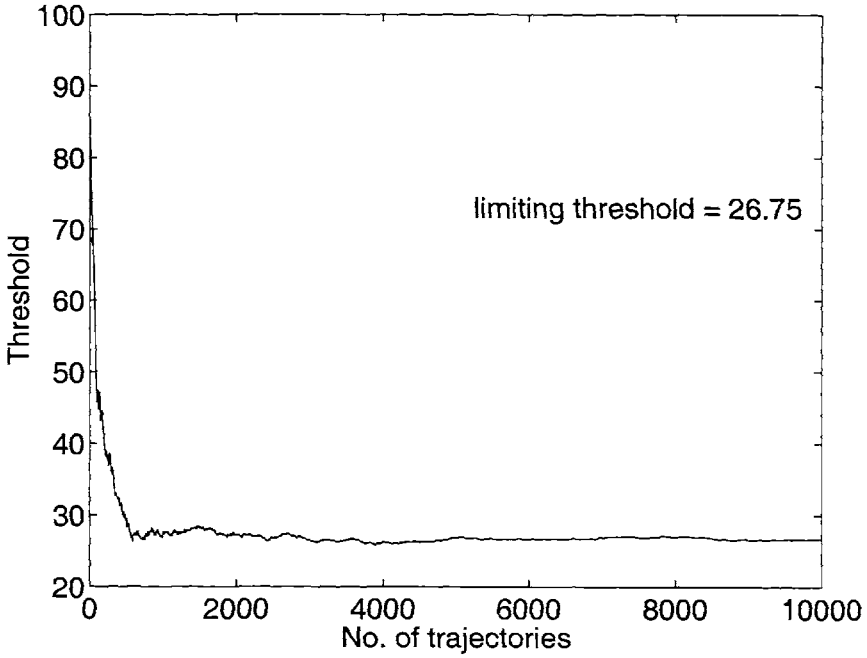

Figure 8-3: Actor-only method with $\tilde{\lambda} = 0.7$.

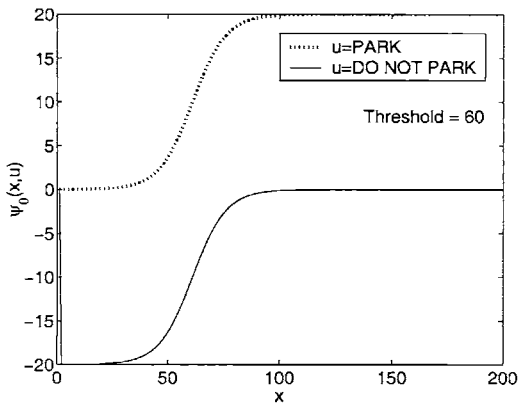

Figure 8-4: Basis function 1

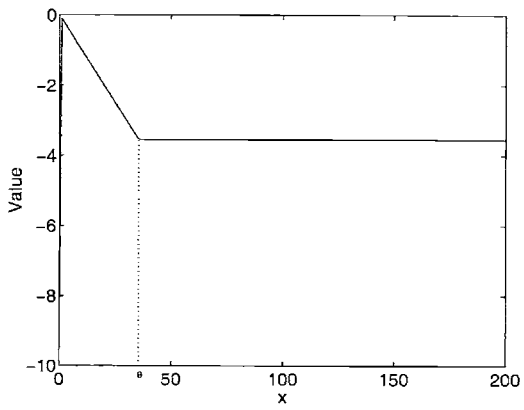

Figure 8-5: Basis function 2

the algorithm. However, the value to which the algorithm converged is 26.75 which is far from the optimal threshold. This is due to the large bias introduced by the choice of $\tilde{\lambda}$ less than 1. But, note that the reduction in the variance is quite remarkable. In the next subsection, we describe an actor-critic algorithm with comparable rate of convergence but converging to a much better threshold.

8.3.2 Actor-Critic Method

The actor-critic method without an eligibility trace for the actor presented in Section 5.1, was tried on this example. The parameter λ of the critic was taken to be 0.7.

As we have argued in Section 5.1, two sets of basis functions are needed for the critic in these algorithms: one set, called the basis functions for advantages, with the same span as the functions ψ_θ^i defined as

$$\psi_\theta^i(x,u) = \frac{\partial \ln \mu_\theta(u|x)}{\partial \theta_i}.$$

and an other set, called the basis functions for state-values, to approximate the state-value function. Since there is only one actor parameter in this problem, the natural choice for the first set is a properly scaled version of the function ψ_θ defined in this case by

$$\begin{aligned}
\psi_\theta(x,u_0) &= \frac{1}{2T\mu_\theta(u_0|x)}\text{sech}\left(\frac{x-\theta}{2T}\right)^2, \\
\psi_\theta(x,u_1) &= -\frac{1}{2T\mu_\theta(u_1|x)}\text{sech}\left(\frac{x-\theta}{2T}\right)^2.
\end{aligned}$$

Thus, the first basis function we use is $\phi_\theta = 10T\psi_\theta$. Figure 8-4 shows this basis function for $\theta = 60$.

The second set also consists of a single basis function that approximates the state-

value functions. To understand the shape or the form of the exact state-value function in this example, consider a deterministic threshold policy with threshold θ. Since the driver employing this policy will park immediately if he finds empty a parking space numbered below θ, the values of the states $x \leq \theta$ should be $r(x)$. When a parking space numbered $x > \theta$ is found empty, the driver using the threshold policy simply passes it and therefore the value of this space should be the same as that of the next parking space. Therefore for $x > \theta$, the value of x is taken to be $r(\theta)$. Thus, the state value basis functions are taken to be piecewise linear and appropriately scaled as shown in Figure 8-5. The step-size used for the critic is of the form

$$\gamma_k = \frac{\gamma_0}{\left(1 + \frac{k}{k_c}\right)^{\alpha}},$$

where

$$\begin{aligned} \gamma_0 &= 0.0005, \\ k_c &= 1000, \\ \alpha &= 0.75. \end{aligned}$$

Using these basis functions, the above step-sizes for the critic and $\lambda = 0.7$, the episodic variant of the actor-critic algorithms presented in the previous section was run for 5000 trajectories using the same step-sizes as (8.3). Figures 8-7 and 8-6 show the evolution of the parameters of the actor and critic during the course of this algorithm. The following are some observations.

- Note that the actor parameter has converged in spite of the critic parameters not converging due to the slowly decreasing step-sizes used. This is contrary to what the proofs of convergence would make one believe, namely that the actor converges because the critic converges. The reason for this discrepancy is the following. Since the actor uses a smaller step-size than that of the critic, only the long run average behavior of the critic is visible to the actor. In other word, if the critic keeps oscillating around a point r^*, due to the averaging in the actor's update, the actor moves in a direction that corresponds to the critic parameter r^*. Therefore, the critic is only needed to "settle down" around the "right values". This is a salient feature of actor-critic algorithms and two-time-scale stochastic approximation in general. Compare these algorithms with optimistic policy iteration (Bertsekas & Tsitsiklis, 1996), in which the convergence of the policy evaluation to the right value function is most critical for policy improvement.

- Another important feature is that the rate of convergence of the actor-critic algorithms is comparable to that of actor-only methods with variance reduction. However, the quality of the solution obtained is much better than that of the actor-only methods. For example, note that the limiting threshold obtained by the actor-critic algorithms is 36.3, which is 0.7 away from the optimal threshold.

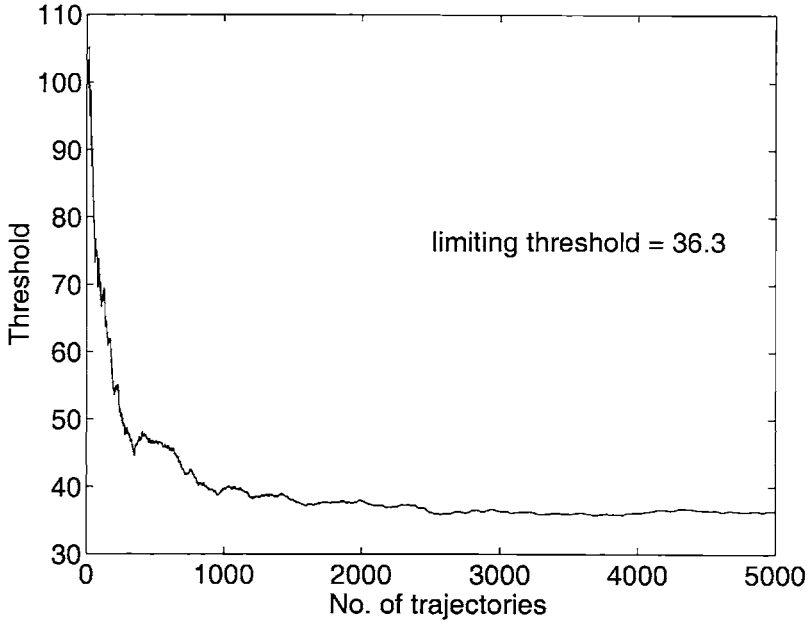

Figure 8-6: Evolution of the actor parameter in the actor-critic algorithm

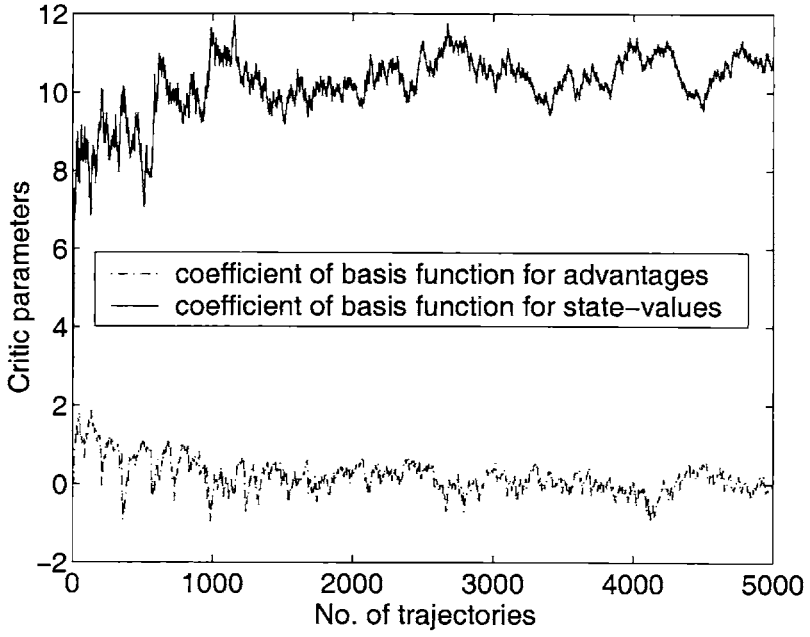

Figure 8-7: Evolution of the critic parameters in the actor-critic algorithm

In contrast, the actor-only method with $\tilde{\lambda} = 0.7$ converged to 26.75 which is far from optimal.

8.4 Closing Remarks

In this chapter, we have considered episodic variants of actor-critic algorithms (without eligibility traces for the actor) for the total reward problem and compared them with their actor-only counterparts. When the actor-critic algorithms use a TD(1) critic, we have shown that their asymptotic variance is the same as that of actor-only methods (without variance reduction). However, we do not know if there is substantial difference in the transient behaviors or which of the algorithms have better transient behavior.

To compare the actor-only algorithms with actor-critic algorithms in which the critic tries to approximate the value function to some extent, we tried both algorithms on a numerical example of small size. Even for such a small problem, the actor-critic algorithms performed better than the actor-only counterpart. More empirical and analytical results are needed to compare the performance of these algorithms. It will be interesting to see how actor-critic algorithms perform when applied to a large real world problem.

Finally, we have made an analytical comparison of rate of convergence of actor-critic algorithm with that of actor-only algorithms only when the former uses a critic with $\lambda = 1$. This comparison needs to be extended to the case when $\lambda < 1$.

Chapter 9

Summary and Future Work

In this thesis, a methodology for optimization of Markov decision processes over a parametric family of randomized stationary policies was presented. Unlike existing methods for solving such problems, the methods of this thesis use value function approximation which facilitates incorporation of user's prior knowledge about the problem (in the form of basis functions) into the solution method. Our belief is that this feature is essential for any methodology to be useful in solving large-scale real world problems. The methodology of this thesis is applicable to Markov decision processes with general state and decision spaces and for optimizing various objective criteria.

Several analytical tools were also developed and used for the analysis of the convergence and of the rate of convergence of some proposed as well as pre-existing algorithms. In particular, a new result on the tracking ability of linear stochastic approximation was proved. This thesis also studied the rate of convergence of two-time-scale linear iterations and provided the first results on this subject. These two results on linear stochastic approximation are applicable in a more general context than that of this thesis.

A major constraint of our methodology is that it is only applicable to optimization over a family of *randomized* policies. While the restriction to a family of randomized policies is justified to some extent, there are important applications in which randomized policies are either undesirable or unnatural. Therefore we need a similar methodology that is applicable to optimization over deterministic policies.

More precisely, consider MDP's with real Euclidean state and decision spaces with system dynamics of the form

$$X_{k+1} = f(X_k, U_k, W_k)$$

where the W_k are i.i.d. and $f(x, u, w)$ is a smooth function of x and u. Let $c(x, u)$ be a smooth one-stage reward function. Let $\{\mu_\theta; \theta \in \mathbb{R}^n\}$ denote a smooth parameterized family of *deterministic* stationary policies. That is, for each $\theta \in \mathbb{R}^n$, $\mu_\theta(\cdot)$ is a function mapping each state to a control where $\mu_\theta(x)$ is assumed to be smooth in x and θ. Let $\bar{\alpha}(\theta)$ be the overall cost when policy μ_θ is used. The following are some of the questions that need to be answered in this context:

- Is there a formula for the gradient of the overall cost $\nabla\bar{\alpha}$ in terms of the value function?
- What actor-critic algorithms are possible in this setting?
- How do these algorithms perform in comparison with existing methods such as IPA (Glasserman, 1991)?

Another important direction to explore is actor-critic algorithms for partially observed Markov decision process. Such techniques would consist of a combination of policy approximation, value function approximation and approximate filtering methods. It will be an interesting exercise to study the interplay of these different techniques.

Our results on the rate of convergence of TD and the comparison of the rate of convergence of actor-critic methods with that of actor-only methods are inconclusive. In particular, we have derived only bounds on the rate of convergence of TD. A more detailed study of TD is needed to test the tightness of these bounds and the qualitative properties inferred from them. Similarly, we have made an analytical comparison of rate of convergence of only a special case of actor-critic algorithm with that of actor-only algorithms. This comparison needs to be extended to other variants of actor-critic algorithms using at least some concrete examples.

Finally, the actor-critic algorithms need to be tested on large-scale real world problems. Further research is needed to come up with guidelines to fine tune these methods to specific applications. For example, guidelines are needed to apply these algorithms to hierarchical and distributed systems such as manufacturing systems, sensor/actuator networks, etc.

References

Abounadi, J., Bertsekas, D. P., & Borkar, V. S. 1998. *Approximation to Non-Expansive Maps: Application to Q-Learning Algorithms.* Preprint.

Abounadi, J., Bertsekas, D. P., & Borkar, V. S. 2001. Learning Algorithms for Markov Decision Processes with Average Cost. *SIAM Journal on Control and Optimization*, **40**(3), 681–698.

Anderson, C. W. 1986. *Learning and Problem Solving with Multilayer Connectionist Systems.*

Athreya, K. B., & Ney, P. 1978. A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.*, **245**, 493–501.

Baras, J. S., & Borkar, V. S. 1999. *A Learning Algorithm For Markov Decision Processes With Adaptive State Aggregation.* Tech. rept. Institute for Systems Research, University of Maryland.

Barto, A., Sutton, R., & Anderson, C. 1983. Neuron-like elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man and Cybernetics*, **13**, 835–846.

Baxter, J., & Barlett, P. L. 1999. *Direct Gradient-Based Reinforcement Learning: I. Gradient Estimation Algorithms.* Tech. rept. Research School of Information Sciences and Engineering, Australian National University.

Benveniste, A., Metivier, M., & Priouret, P. 1990. *Adaptive Algorithms and Stochastic Approximations.* Berlin-Heidelberg: Springer-Verlag.

Bertsekas, D. P. 1995a. A Counterexample to Temporal Differences Learning. *Neural Computation*, **7**(2), 270–279.

Bertsekas, D. P. 1995b. *Dynamic Programming and Optimal Control.* Belmont, MA: Athena Scientific.

Bertsekas, D. P., & Tsitsiklis, J. N. 1996. *Neuro-Dynamic Programming.* Belmont, MA: Athena Scientific.

Bertsekas, D. P., & Tsitsiklis, J. N. 2000. Gradient Convergence in gradient methods. *SIAM Journal in Optimization*, **10**(3), 627–642.

Bhatnagar, S., Fu, M. C., & Marcus, S. I. 1999. *Optimal multilevel feedback policies for ABR flow control using two timescale SPSA.* Tech. rept. Institute of Systems Research, University of Maryland.

Bhatnagar, S., Fu, M. C., Marcus, S. I., & Bhatnagar, S. 2000. *Randomized Difference Two-Timescale Simultaneous Perturbation Stochastic Approximation Algorithms for Simulation Optimization of Hidden Markov Models.* Tech. rept. TR 2000-13. Institute of Systems Research, University of Maryland.

Borkar, V. S. 1996. Stochastic approximation with two time scales. *Systems and Control Letters*, **29**, 291–294.

Borkar, V. S., & Meyn, S. P. 2000. The O.D.E. Method for Convergence of Stochastic Approximation and Reinforcement Learning. *SIAM Journal on Control and Optimization*, **38**(2), 447–469.

Boyan, J. A. 1999. Least-squares temporal difference learning. *Pages 49–56 of: Machine Learning: Proceedings of the Sixteenth International Conference.* San Francisco, CA: Morgan Kaufmann.

Cao, X. R., & Chen, H. F. 1997. Perturbation Realization, Potentials, and Sensitivity Analysis of Markov Processes. *IEEE Transactions on Automatic Control*, **42**, 1382–1393.

Duflo, M. 1997. *Random iterative models.* Springer-Verlag.

Dupuis, P. 1988. Large deviation analysis of some recursive algorithms with state-dependent noise. *Annals of Probabilitu*, **16**, 1506–36.

Dupuis, P., & Kushner, H.J. 1985. Stochastic approximations via large deviations: asymptotic properties. *SIAM Journal on Control and Optimization*, **23**, 675–96.

Dupuis, P., & Kushner, H.J. 1987. Asymptotic behavior of constrained stochastic approximations via the theory of large deviations. *Probability theory and Related fields*, **75**, 223–44.

Dupuis, P., & Kushner, H.J. 1989. Stochastic approximation and large deviations: upper bounds and w.p.1. convergence. *SIAM Journal on Control and Optimization*, **27**, 1108–35.

Eweda, E., & Machi, O. 1984. Convergence of an adaptive linear estimation algorithm. *IEEE Trans. on Automatic Control*, **AC-29**(2), 119–127.

Feinberg, Eugene A., & Schwartz, Adam (eds). 2001. *Handbook of Markov Decision Processes : Methods and Applications.* Kluwer Academic Publishers.

Glasserman, P. 1991. *Gradient estimation via perturbation analysis.* Kluwer Academic Publishers.

Glynn, P. W. 1986. Stochastic Approximation for Monte Carlo Optimization. *Pages 285–289 of: Proceedings of the 1986 Winter Simulation Conference.*

Glynn, P. W. 1987. Likelihood Ratio Gradient Estimation: an Overview. *Pages 366–375 of: Proceedings of the 1987 Winter Simulation Conference.*

Glynn, P. W., & L'Ecuyer, P. 1995. Likelihood ratio gradient estimation for stochastic recursions. *Advances in applied probability*, **27**, 1019–1053.

Jaakola, T., Jordan, M., & Singh, S. P. 1994. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, **6**, 1185–1201.

Jordan, S., & Varaiya, P. 1994. Control of Multiple Service, Multiple Reseource Communication Networks. *IEEE Transactions on Automatic Control*, **20**(11), 2979–2988.

Kearns, M., & Singh, S. 2000. Bias-Variance Error Bounds for Temporal Difference Updates. *Pages 142–147 of: Proceedings of the 13th Annual Conference on Computational Learning Theory.*

Kimura, H., & Kobayashi, S. 1998. An Analysis of Actor/Critic Algorithms using Eligibility Traces: Reinforcement Learning with Imperfect Value Function. *Pages 278–286 of: 15th International Conference on Machine Learning.*

Kokotovic, P. V. 1984. Applications of singular perturbation techniques to control problems. *SIAM Review*, **26**(4).

Konda, V. R. 1997. *Learning algorithms for Markov decision processes.* M.Phil. thesis, Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India.

Konda, V. R., & Borkar, V. S. 1999. Actor-Critic like learning algorithms for Markov decision processes. *SIAM Journal on Control and Optimization*, **38**(1), 94–123.

Konda, V. R., & Tsitsiklis, J. N. 2000a. Actor-Critic Algorithms. *Pages 1008–1014 of: Advances in Neural Information Processing Systems.* MIT Press.

Konda, V. R., & Tsitsiklis, J. N. 2000b (February). *Actor-Critic Algorithms.* Submitted to the *SIAM Journal on Control and Optimization.*

Kushner, H. J., & Clark, D. S. 1978. *Stochastic Approximation for Constrained and Unconstrained Systems.* New York: Springer-Verlag.

Kushner, H. J., & Huang, H. 1979. Rates of convergence for stochastic approximation type algorithms. *SIAM Journal on Control and Optimization*, **17**, 607–617.

Kushner, H. J., & Yin, G. G. 1997. *Stochastic approximation algorithms and applications.* Springer-Verlag.

Kushner, H.J. 1984. Asymptotic behavior of stochastic approximation and large deviations. *IEEE Transactions on Automatic Control*, **29**, 984–90.

Kushner, H.J., & Yang, J. 1993. Stochastic approximation with averaging of the iterates: Optimal asymptotic rates of convergence for general processes. *SIAM Journal on Control and Optimization*, **31**, 1045–1062.

Marbach, P. 1998. *Simulation based optimization of Markov reward processes.* Ph.D. thesis, Massachusetts Institute of Technology.

Marbach, P., & Tsitsiklis, J. N. 2001. Simulation-Based Optimization of Markov Reward Processes. *IEEE Transactions on Automatic Control*, **46**(2), 191–209.

Meyn, S. P., & Tweedie, R. L. 1993. *Markov Chains and Stochastic Stability.* Springer-Verlag.

Michie, D., & Chambers, R. A. 1968. *BOXES: An experiment in adaptive control.* Edinburgh: Oliver and Boyd. Pages 137–152.

Nevel'son, M. B., & Has'minskii, R. Z. 1973. *Stochastic approcimation and recursive estimation.* American Mathematical Society.

Nummelin, E. 1978. A splitting technique for Harris recurrent chains. *Z. Wahrscheinlichkeitstheorie and Verw. Geb.*, **43**, 119–143.

Ormoneit, D., & Glynn, P. 2001. Kernel-based reinforcement learning in average-cost problems: An application to optimal portfolio choice. *In: Advances in Neural Information Processing Systems 13.* The MIT Press.

Ormoneit, D., & Sen, S. 2000. Kernel-based reinforcement learning. *Machine Learning.*

Polyak, B. T. 1976. Convergence and convergence rate of iterative stochastic algorithms I. *Automat. Remote Control*, **12**, 1858–1868.

Polyak, B. T. 1990. New method of stochastic approximation type. *Automat. Remote Control*, **51**, 937–946.

Polyak, B. T., & Juditsky, A. B. 1992. Acceleration of stochastic approximation by averaging. *SIAM Journal of Control and Optimization*, **30**(838–855).

Puterman, M. 1994. *Markov Decision Processes.* New York: John Wiley.

Singh, S., & Dayan, P. 1998. Analytical Mean Squared Error Curves in Temporal Difference Learning. *Machine Learning*, **32**, 5–40.

Singh, S. P., & Sutton, R. S. 1996. Reinforcement Learning with Replacing Eligibility Traces. *Machine Learning*, 123–158.

Sutton, R., & Barto, A. 1998. *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press.

Sutton, R. S. 1984. *Temporal Credit Assignment in Reinforcement Learning.*

Sutton, R. S. 1988. Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, **3**, 9–44.

Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. 2000. Policy gradient methods for reinforcement learning with function approximation. *Pages 1057–1063 of: Advances in Neural Information Processing Systems*, vol. 12.

Tsitsiklis, J. N. 1994. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, **16**, 185–202.

Tsitsiklis, J. N., & Van Roy, B. 1996. Feature-based methods for large scale dynamic programming. *Machine Learning*, **22**, 59–94.

Tsitsiklis, J. N., & Van Roy, B. 1997. An analysis of Temporal-Difference Learning with Function approximation. *IEEE Transactions on Automatic Control*, **42**(5), 674–690.

Tsitsiklis, J. N., & Van Roy, B. 1999a. Average cost temporal-difference learning. *Automatica*, **35**(11), 1799–1808.

Tsitsiklis, J. N., & Van Roy, B. 1999b. Optimal Stopping of Markov Processes: Hilbert Space Theory, Approximation Algorithms, and an Application to Pricing Financial Derivatives. *IEEE Transactions on Automatic Control*, **44**(10), 1840–1851.

Van Roy, B. 1998. *Learning and Value Function Approximation in Complex Decision Processes.* Ph.D. thesis, Massachusetts Institute of Technology.

Watkins, C. 1989. *Learning from delayed rewards.* Ph.D. thesis, Cambridge Univ., Cambridge, England.

Watkins, C., & Dayan, P. 1992. Q-learning. *Machine Learning*, **8**, 279–292.

Widrow, B., McCool, J., Larimore, M. G., & Johnson, C. R. 1976. Stationary and non-stationary learning characterstics of the LMS adaptive filter. *Proc. IEEE*, **64**(8), 1151–1161.

Williams, R. 1992. Simple Statistical Gradient Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*, **8**, 229–256.

Williams, R., & Baird, L. 1990. A mathematical analysis of actor-critic architectures for learning optimal controls through incremental dynamic programming. *Pages 96–101 of: Sixth Yale Workshop on Adaptive and Learning Systems.*

Witten, I. H. 1977. An adaptive optimal controller for discrete-time Markov environments. *Information and Control,*, **34**, 286–295.