Journal of Machine Learning Research 13 (2012) 1927-1966 Submitted 1/12; Revised 3/12; Published 6/12

Integrating a Partial Model into Model Free Reinforcement Learning

Aviv Tamar AVIVT@TX.TECHNION.AC.IL
Dotan Di Castro DOT@TX.TECHNION.AC.IL
Ron Meir RMEIR@EE.TECHNION.AC.IL
Department of Electrical Engineering
Technion
Haifa 32000, Israel

Editor: Peter Dayan

Abstract

In reinforcement learning an agent uses online feedback from the environment in order to adaptively select an effective policy. Model free approaches address this task by directly mapping environmental states to actions, while model based methods attempt to construct a model of the environment, followed by a selection of optimal actions based on that model. Given the complementary advantages of both approaches, we suggest a novel procedure which augments a model free algorithm with a partial model. The resulting *hybrid* algorithm switches between a model based and a model free mode, depending on the current state and the agent's knowledge. Our method relies on a novel definition for a partially known model, and an estimator that incorporates such knowledge in order to reduce uncertainty in stochastic approximation iterations. We prove that such an approach leads to improved policy evaluation whenever environmental knowledge is available, without compromising performance when such knowledge is absent. Numerical simulations demonstrate the effectiveness of the approach on policy gradient and Q-learning algorithms, and its usefulness in solving a call admission control problem.

Keywords: reinforcement learning, temporal difference, stochastic approximation, markov decision processes, hybrid model based model free algorithms

1. Introduction

In Reinforcement Learning (RL) an agent attempts to improve its performance over time at a given task, based on continual interaction with the (usually unknown) environment, (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998). This improvement takes place by modifying the action selection policy, based on feedback from the environment and prior knowledge available to the agent. Formally, RL is often phrased as the problem of finding a mapping, the so called *policy*, from the environment's states to the agent's actions that maximizes a given functional of a reward function.

Most RL algorithms can be classified into either *model based* (also termed indirect) or *model free* (direct) approaches (Sutton and Barto, 1998; Bertsekas and Tsitsiklis, 1996). In the former setting, taking its inspiration from the field of Adaptive Control (Kumar, 1985), the agent maintains an explicit model of the environmental dynamics, typically in the form of a *Markov Decision Process* (MDP), while interacting with it. Based on this model, a *planning* problem is solved where techniques from *Dynamic Programming* (Bertsekas, 2006) are applied in order to find the optimal policy function. On the other hand, within the model free setting, the agent does not try to build a model of the MDP, but rather attempts to find the optimal policy by directly mapping environmental

states to actions. In this sense, no model of the environmental dynamics is required. While it can be shown that both approaches, under mild conditions, asymptotically reach the same optimal policy on typical MDP's, it is known that each approach possesses distinct merits. Model based methods often make better use of a limited amount of experience and thus achieve a better policy with fewer environmental interactions. On the other hand, model free methods are simpler, require less computational resources, and are not affected by biases in the design (or estimation) of the model.

The view taken in this work is that this dichotomy between algorithmic approaches, although popular, is not necessarily desirable. As an example, consider a scenario where some parts of the environment are known in advance, but computational resources are limited, restricting the use of proper model based approaches. In this case, a *hybrid approach* may allow us to benefit from using parts of the model in the algorithm, without sacrificing its simplicity, thus striking a balance between the merits of each approach. Surprisingly, the concept of combining model free and model based algorithms has received very little attention in the RL literature, and theoretical guarantees to its advantages are lacking.

In this work we pursue such a hybrid approach applicable to cases where partial model information is available in a specific form which we term *partially known MDP*. We provide a method for integrating such information into RL algorithms of the Stochastic Approximation (SA) type (Kushner and Yin, 2003; Borkar, 2008). This class of online model free algorithms includes many standard RL approaches that have been used effectively in practice (e.g., Tesauro, 1995; Crites and Barto, 1996). The method we propose reduces uncertainty in the algorithm trajectory, thereby improving its performance. Our theoretical analysis focuses on a particular model free algorithm - the well known TD(0) policy evaluation algorithm, and we prove that our hybrid method leads to improved performance, as long as sufficiently accurate partial knowledge is available. The effectiveness and generality of our method is further demonstrated in two numerical simulations. In the first, we apply it to a policy gradient type algorithm, and investigate its performance in randomly generated MDPs. In the second, we consider a call admission control problem. As it turns out, our partially known MDP definition is a natural choice for describing plausible partial knowledge in such problems, and performance improvement is demonstrated for a Q-learning algorithm.

1.1 A High-Level Sketch of the Method

Online SA algorithms attempt to optimize some parameter of the system, using "noise corrupted" system measurements as a data stream for an iterative optimization process. These algorithms deal with noise by making only small changes to the parameters at each step, so that over many iterations the noise averages out, and the parameters asymptotically follow a *mean trajectory.* Intuitively, any prior knowledge about the system should reduce our uncertainty about its behavior and thus enable some noise reduction. In this work we propose a method that *reduces the noise at each step*. We do this by observing that the update at each step can be viewed as a simple estimate of the mean update. Using the partially known MDP we propose an improved estimator, thereby reducing the noise variance. A key property of our estimator is that it is unbiased, thus it *preserves the algorithm's mean trajectory.* This assures us that the overall *function* of the algorithm will remain intact, while the reduction in noise variance gives reason to expect an improvement in *performance*.

1.2 Related Work

One difficulty of RL is coping with the stochasticity inherent to RL algorithms. In this work we define a partially known MDP, and use this partial knowledge to improve the asymptotic performance of stochastic approximation type algorithms. Our method is novel, and specifically deals with this difficulty. We note that a different notion of a partially known MDP was used by Kearns and Singh (2002) and Brafman and Tennenholtz (2003) to tackle a different difficulty of RL - the 'exploration exploitation' tradeoff. Thus, the partial model which we use only to reduce stochastic fluctuations may further be used to explore or exploit more efficiently. On the other hand, the advantage of our approach is that it is general, and may be easily applied to a large class of model free algorithms.

When a full model of state transitions is available, applying our method to Q-learning results in an algorithm known as Real Time Dynamic Programming (RTDP) (Barto et al., 1995). Thus, the method presented in this work may be viewed as a bridge between the model free Q-Learning and the model based RTDP.

In Section 4 we analyze the asymptotic fluctuations in a fixed step TD(0) algorithm with a partial model. A similar analysis of TD(0) without partial model was given by Dayan and Sejnowski (1994) for a decreasing step size and without explicit convergence rate results, and by Konda (2002) using a similar technique to bound the convergence rate. Singh and Dayan (1998) provided update equations for the MSE of TD(0), which we use as a measure of convergence rate, though their equations were only solvable by simulation. This work presents explicit *values* of the asymptotic MSE.

On a slightly different note, an early approach towards a hybrid model based - model free RL algorithm is the Dyna architecture (Sutton, 1990), in which interactions with the environment are used both for a direct policy update, using a model free RL algorithm, and for an update of an environmental model. This model is then used to generate simulated trajectories which are fed to the same model free algorithm for further policy improvement. In a more recent work by Abbeel et al. (2006), a hybrid approach is proposed that combines policy search on an inaccurate model, with policy evaluations in the real environment. Finally, we note that the idea of combining model based and model free approaches has been proposed in the context of animal and human learning, suggesting an explanation for behavioral choice experiments (Daw et al., 2005). To the best of our knowledge, our work presents the first formal proof of the *advantage* of a hybrid algorithm over a standard model free algorithm.

1.3 Organization

This paper is organized as follows. In Section 2 we describe our estimator in the context of estimating the expectation of a random variable. This allows us to derive all its important properties without the notational burden of the SA setting. In Section 3 we describe the RL environment and introduce the partially known MDP. We then describe a method that integrates it in model free SA algorithms. Our main results are in Section 4, in which we analyze how our proposed method influences the algorithm's overall performance. Focusing on TD(0), we show that improvement in performance is achieved. In Section 5 we investigate the effects of inaccuracies in the partial model, and extend our results to inaccurate partially known MDP's. In Section 6 we demonstrate through simulation the applicability of our method to other model free RL algorithms. We conclude and discuss future work in Section 7.

2. Estimation of a Random Variable Mean with Partial Knowledge

Our method of using partial knowledge in an SA algorithm is based on constructing a better estimator for the mean update at each step. In this section we describe our estimator in the context of estimating the mean of a random variable. This allows us to derive all its important properties without the notational burden of the SA setting. The results we derive will then easily transfer to the more complicated SA setting.

Let X be a random variable over a finite and discrete set Ω and let $P(\omega) \triangleq \Pr(X=\omega)$ denote the probability distribution of X. Since $P(\omega)$ contains all the information about X, a natural definition for partial knowledge in this setting is information regarding some of its attributes. In particular, we assume that for an a-priori given subset of Ω, the ratios between the probability distribution values are provided. Denote by K this set for which the ratios of P are known,[1]

$$K \triangleq \left\{ \omega : \omega \in \Omega \text{ s.t. } \frac{P(\omega)}{\sum\limits_{\omega' \in K} P(\omega')} \text{ is known} \right\}. \tag{1}$$

We refer to K as the *partial knowledge* set. Suppose we are given one sample of X, denoted by x, and we wish to estimate (without bias) the expectation $\mu = \mathrm{E}[X] \triangleq \sum_{\omega \in \Omega} \omega P(\omega)$. Our estimator can be any function of x, and of values and probability ratios in the partial knowledge set K.

The Maximum Likelihood (ML) estimator of μ is derived by first using x to generate the ML estimate of the complete probability distribution $\hat{P}(\omega)$, and then calculating the expectation $\sum_\omega \omega \hat{P}(\omega)$. For a given known set K, let $\mathcal{P}_K(\omega)$ denote the set of all probability distributions $P(\omega)$ that satisfy the ratios in (1). If the observed sample x is not in K, then the ML estimate for the probability distribution $\hat{P}(\omega; x \notin K)$ is given by

$$\hat{P}(\omega; x \notin K) = \underset{P(\omega) \in \mathcal{P}_K(\omega)}{\operatorname{argmax}} P(x) = \delta_{x,\omega}, \tag{2}$$

where $\delta_{x,\omega}$ is Kronecker's delta. Conversely, if x is in K, the ML estimate $\hat{P}(\omega; x \in K)$ is

$$\hat{P}(\omega; x \in K) = \underset{P(\omega) \in \mathcal{P}_K(\omega)}{\operatorname{argmax}} P(x) = \frac{\mathbf{1}_\omega^K P(\omega)}{\sum\limits_{\omega' \in K} P(\omega')}, \tag{3}$$

where $\mathbf{1}_\omega^K$ denotes the indicator function that equals 1 if $\omega \in K$ and 0 otherwise. Letting $\bar{K}$ denote the complement of the set K, combining (2) and (3) gives the ML estimate for the probability distribution $\hat{P}(\omega)$

$$\hat{P}(\omega; x) = \mathbf{1}_x^K \frac{\mathbf{1}_\omega^K P(\omega)}{\sum\limits_{\omega' \in K} P(\omega')} + \mathbf{1}_x^{\bar{K}} \delta_{x,\omega}. \tag{4}$$

By taking an expectation of (4), we derive the ML estimate for μ given the partial knowledge, which we denote by $\hat{\mu}_K$

$$\hat{\mu}_K(x) = \mathbf{1}_x^K \cdot \frac{\mathrm{E}[X \cdot \mathbf{1}_X^K]}{\mathrm{E}[\mathbf{1}_X^K]} + \mathbf{1}_x^{\bar{K}} \cdot x. \tag{5}$$

1. Note that knowledge of the exact probability distribution values is a special case of this definition.

Note that (5) uses the partial knowledge in a very intuitive way. It 'replaces' samples in the known set with their weighted average, which by (1) is known. An important property of the estimator $\hat{\mu}_K$ is that it is unbiased, as expressed in the following Lemma.

Lemma 1 *The estimator $\hat{\mu}_K$ is unbiased, namely $\mathrm{E}[\hat{\mu}_K] = \mu$.*

Proof By direct calculation

$$\begin{aligned}\mathrm{E}[\hat{\mu}_K] &= \mathrm{E}[\mathbf{1}_X^K] \cdot \frac{\mathrm{E}[X \cdot \mathbf{1}_X^K]}{\mathrm{E}[\mathbf{1}_X^K]} + \mathrm{E}\left[\mathbf{1}_X^{\bar{K}} \cdot X\right] \\ &= \mathrm{E}\left[X\left(\mathbf{1}_X^K + \mathbf{1}_X^{\bar{K}}\right)\right] = \mathrm{E}[X].\end{aligned}$$

■

In the following Lemma the Mean Squared Error (MSE) of $\hat{\mu}_K$ is computed. Let $P(K) = \sum_{\omega \in K} P(\omega)$, and let $P_K(\omega)$ denote the probability measure over the known set K, namely

$$P_K(\omega) \triangleq \mathbf{1}_\omega^K P(\omega) / P(K).$$

Denote by $\mathrm{E}_K[\cdot]$ and $\mathrm{Var}_K[\cdot]$ the expectation and variance under the probability measure P_K.

Lemma 2 *The MSE of $\hat{\mu}_K$ is $\mathrm{E}\left[(\hat{\mu}_K - \mu)^2\right] = \mathrm{Var}[X] - P(K) \cdot \mathrm{Var}_K[X]$.*

Proof Observe that for any function $f(\cdot)$

$$\begin{aligned}\mathrm{E}_K\left[(f(X) - \mu)^2\right] &= \mathrm{E}_K\left[(f(X) - \mathrm{E}_K f(X) + \mathrm{E}_K f(X) - \mu)^2\right] \\ &= \mathrm{Var}_K f(X) + (\mathrm{E}_K[f(X)] - \mu)^2, \qquad (6)\end{aligned}$$

where the cross terms in the second equality vanish.
Next, we have

$$\begin{aligned}\mathrm{E}[\hat{\mu}_K(X) - \mu]^2 &= \mathrm{E}\left[\left(\mathbf{1}_X^K + \mathbf{1}_X^{\bar{K}}\right)(\hat{\mu}_K(X) - \mu)^2\right] \\ &= P(K)\,\mathrm{E}_K\left[(\hat{\mu}_K(X) - \mu)^2\right] + \mathrm{E}\left[\mathbf{1}_X^{\bar{K}}(X - \mu)^2\right] \\ &= P(K)(\mathrm{E}_K[X] - \mu)^2 + \mathrm{E}\left[\mathbf{1}_X^{\bar{K}}(X - \mu)^2\right] \\ &= P(K)\left(\mathrm{E}_K\left[(X - \mu)^2\right] - \mathrm{Var}_K[X]\right) + \mathrm{E}\left[\mathbf{1}_X^{\bar{K}}(X - \mu)^2\right] \\ &= \mathrm{E}\left[(X - \mu)^2\right] - P(K) \cdot \mathrm{Var}_K[X],\end{aligned}$$

where in the fourth equality we used (6) with $f(X) = X$. ■

One could disregard the partial knowledge altogether, and choose to use the sample x itself as an unbiased estimate for μ. Denote this estimator, which will be referred to as the sample estimator, by

$$\hat{\mu}(x) = x. \qquad (7)$$

When no partial information is available, $\hat{\mu}$ seems like the most reasonable choice (actually, it can be shown that $\hat{\mu}$ is the only unbiased estimator in that case). It is easy to see that the MSE of $\hat{\mu}$ is

$\mathrm{Var}[X]$, and from Lemma 2 we deduce that when the cardinality of the known set satisfies $|K| > 1$, and $P(K) > 0$, the MSE of $\hat{\mu}_K$ is smaller than that of $\hat{\mu}$. In parameter estimation parlance, we say that $\hat{\mu}_K$ *dominates* $\hat{\mu}$ (Schervish, 1995).

As will be shown in the next section, the update at each iteration of an SA algorithm can be seen as the estimation of an expected update direction. This estimation is based on one sample, obtained through observation of the system dynamics at that step, and the estimator used is just $\hat{\mu}$. When partial knowledge of these dynamics is available, we propose to use $\hat{\mu}_K$ instead, and benefit from its reduced variance.

An appropriate question at this point is whether a better estimator than $\hat{\mu}_K$ exists. We refer the interested reader to appendix A, where we show that $\hat{\mu}_K$ is admissible. For the following discussion however, the results of Lemmas 1 and 2 suffice.

3. A Stochastic Approximation Algorithm with Partial Model Knowledge

In this section we describe our method of endowing a model free RL algorithm with partial model knowledge. We start with some general definitions of the RL environment and SA algorithms. Then, we consider a situation where partial knowledge of the environment model is available. Based on the estimator developed in the previous section, we propose a general form of SA algorithms that incorporate such knowledge.

3.1 Preliminaries

We describe the notation used throughout the paper, the RL environment, and the stochastic approximation method.

3.1.1 NOTATION

Throughout the rest of the paper the following notation is used. All vectors are column vectors, and $(\cdot)^T$ denotes the transpose operator. The product $A \circ B$ denotes the element-wise product (Hadamard product) of A and B. $\mathrm{Tr}[\cdot]$ is the trace of a matrix. The cardinality of a set K is denoted by $|K|$, and its complement by $\bar{K}$. Unless noted otherwise, a subscript of a variable denotes time. The i-th element of a vector A is denoted by $[A]_i$ or $A(i)$, depending on the context. The (i,j) element of a matrix B is denoted by $[B]_{ij}$.

3.1.2 RL ENVIRONMENT

We consider an agent interacting with an unknown environment, modeled by an MDP in discrete time with a finite state set $\mathcal{X}$ and action set $\mathcal{U}$. Each selected action $u \in \mathcal{U}$ at a state $x \in \mathcal{X}$ determines a stochastic transition to the next state $y \in \mathcal{X}$ with a probability $P_u(y|x)$.

For each state x the agent receives a corresponding deterministic reward $r(x)$, which is bounded by $r_{\max}$, and depends only on the current state.[2] The agent maintains a *policy function,* $\mu_\theta(u|x)$, parametrized by a vector $\theta \in \mathbb{R}^L$, mapping a state x into a probability distribution over the actions $\mathcal{U}$. Under policy μ_θ, the environment and the agent induce a Markovian transition matrix, denoted by P_{μ_θ}, which we assume to be ergodic.[3] This Markovian transition matrix has a stationary distribution

2. Generalizing the results presented here to state-action rewards is straightforward. Generalization to stochastic rewards is also possible by considering mean rewards.
3. That is, aperiodic, recurrent, and irreducible.

over the state space $\mathcal{X}$, denoted by π_{μ_θ}. Let $\Pi_{\mu_\theta} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ be a diagonal matrix where its elements are $\Pi_{\mu_\theta} = \mathrm{diag}(\pi_{\mu_\theta})$. Our goal is to optimize θ with respect to some performance criteria. The tuning of θ is performed online in the following fashion. At time n, the current parameter value equals θ_n and the agent is in state x_n. It then chooses an action u_n according to $\mu_{\theta_n}(u|x_n)$, observes x_{n+1}, and updates θ_{n+1} according to some protocol.

3.1.3 STOCHASTIC APPROXIMATION

Stochastic approximation methods (Kushner and Yin, 2003; Borkar, 2008) are a class of iterative stochastic algorithms, to which many model free RL algorithms belong (Bertsekas and Tsitsiklis, 1996). Analysis of SA methods has received considerable attention over the past few decades, and many analysis techniques are available. In particular, the ODE approach introduced by Ljung (1977), is a widely used method for investigating the asymptotic behavior of SA iterates. The algorithms that we deal with in this paper are all cast in the following SA form,[4]

$$\theta_{n+1} = \theta_n + \varepsilon_n F(\theta_n, x_n, u_n, x_{n+1}), \tag{8}$$

where $\{\varepsilon_n\}$ are positive step sizes. The key idea of the technique is the following. Iterate (8) can be decomposed into a deterministic function of the current state, action and parameter, denoted by $g(\theta_n, x_n, u_n)$, and a martingale difference noise term δM_n,

$$\theta_{n+1} = \theta_n + \varepsilon_n (g(\theta_n, x_n, u_n) + \delta M_n), \tag{9}$$

where $g(\theta_n, x_n, u_n) \triangleq \mathrm{E}[F(\theta, x_n, u_n, x_{n+1})|\theta_n, x_n, u_n]$, $\delta M_n \triangleq F(\theta_n, x_n, u_n, x_{n+1}) - g(\theta_n, x_n, u_n)$, and the expectation is taken over the next state x_{n+1}.

Suppose that the effect of the martingale difference noise weakens due to repeated averaging, and further assume that there exists a continuous function $\bar{g}(\theta)$ such that $\frac{1}{m}\sum_{i=n}^{m+n-1} g(\theta, x_i, u_i) \to \bar{g}(\theta)$ w.p.1 as $m, n \to \infty$.[5] Consider the following ordinary differential equation (ODE)

$$d\theta/dt = \bar{g}(\theta). \tag{10}$$

Then, a typical result of the ODE method in the SA setup suggests that the asymptotic limits of (8) and (10) are identical. Another aspect of SA relates to the rate of convergence of such iterates (Kushner and Yin, 2003), an issue that will be elaborated on later.

3.1.4 A NOTE ON TYPES OF CONVERGENCE

The type of convergence to the asymptotic limit depends primarily on the step size used. Let θ^* denote an asymptotically stable fixed point of (10), and assume that it is unique. Then, for a suitably decreasing step size, convergence w.p. 1 of θ_n to θ^* can be established. For a constant step size, θ_n can be shown to converge weakly to a random variable centered on θ^*. In the following we use the term convergence ambiguously, and the precise definition should be inferred from the context. For a detailed and rigorous discussion of the types of convergence in SA the reader is referred to Kushner and Yin (2003).

4. This is not the most general SA form, but one that is cast to the RL setup.

5. Note that for stationary policies, the strong law of large numbers for Markov chains may be used to write $\bar{g}$ explicitly $\bar{g}(\theta) = \mathrm{E}[g(\theta, x, u)|\theta] = \sum_{x\in X} \pi_{\mu_\theta}(x) \sum_{u\in U} \mu_\theta(u|x) g(\theta, x, u)$.

3.2 Partial Model Based Algorithm

A key observation obtained from examining Equations (8-9), is that $F(\theta_n, x_n, u_n, x_{n+1})$ in the SA algorithm is just the sample estimator (7) of $g(\theta_n, x_n, u_n)$, the mean update at each step. The estimation variance in this case stems from the stochastic transition from x_n to x_{n+1}. In the following we assume that we have, prior to running the algorithm, some information about these transitions in the form of partial transition probability ratios. Similarly to Section 2, define the known set for state x and action u as

$$K_{x,u} \triangleq \left\{ y : y \in \mathcal{X} \text{ s.t. } \frac{P_u(y|x)}{\sum_{y' \in K_{x,u}} P_u(y'|x)} \text{ is known} \right\}. \tag{11}$$

We refer to the known sets for all states and actions as the *partially known MDP*.

It is clear that definition (11) is motivated by the theoretical results presented in Section 2, and at this point it may well be questioned whether such a definition has any use in practice. We refer the concerned reader to Section 6, where it is shown that in certain problems definition (11) arises as the *natural* representation of partial model knowledge.

Denote by $\mathbf{1}^K_{n+1}$ an indicator function that equals 1 if $\{x_{n+1}\}$ belongs to K_{x_n,u_n} and 0 otherwise. Based on the estimator introduced in Section 2, we propose the following update rule for the tunable parameter, denoted by θ^K, which we refer to as the *Integrated Partial Model* (IPM) iteration

$$\theta^K_{n+1} = \theta^K_n + \varepsilon_n \left(\mathbf{1}^K_{n+1} F^K_n + \mathbf{1}^{\bar{K}}_{n+1} F(\theta^K_n, x_n, u_n, x_{n+1}) \right), \tag{12}$$

where, abusing notation, $F^K_n = F^K_n(\theta^K_n, x_n, u_n)$, and

$$F^K_n \triangleq \frac{\sum_{y \in K_{x_n,u_n}} P_{u_n}(y|x_n) F(\theta^K_n, x_n, u_n, y)}{\sum_{y \in K_{x_n,u_n}} P_{u_n}(y|x_n)}. \tag{13}$$

Similarly to (9), iterate (12) can also be decomposed into a mean function $g^K(\theta^K_n, x_n, u_n)$ and a martingale difference noise δM^K_n

$$\theta^K_{n+1} = \theta^K_n + \varepsilon_n \left(g^K(\theta^K_n, x_n, u_n) + \delta M^K_n \right),$$

and by Lemma 1 we have $g^K(\theta, x, u) = g(\theta, x, u)$. Similarly, defining $\bar{g}^K(\theta) = \mathrm{E}\left[g^K(\theta, x, u) | \theta \right]$ we get that $\bar{g}^K(\theta) = \bar{g}(\theta)$, and we reach the following important conclusion, which is summarized as a theorem.

Theorem 3 *The IPM iteration defined in* (12) *leads to the same characteristic ODE* $d\theta/dt = \bar{g}(\theta)$ *as the regular SA iteration* (8).

Since the asymptotic behavior of the SA iterate (8) is governed by its ODE, Theorem 3 assures us that using the IPM iteration (12) does not change this behavior, and thus the *function* of the algorithm remains intact. If (8) can be shown to converge to some limit point, iterate (12) can be shown to converge *to the same limit*. Furthermore, from Lemma 2 we have that if the partially known MDP is not null, then on each iteration the variance of the noise term is reduced. This gives us reason to expect an improvement in the overall performance of the algorithm.

3.3 Step Size Considerations

As it turns out, the improvement in performance attained by the IPM iteration is heavily influenced by the step size used. This can be intuitively explained using the following example. Let $\{z_i\}$ be a sequence of i.i.d. bounded random variables, with mean μ_z and variance σ_z^2. Consider the following SA iteration

$$\theta_{n+1} = \theta_n + \varepsilon_n (z_{n+1} - \theta_n).$$

For a decreasing step size of the form $\varepsilon_n = 1/(n+1)$, the value of θ_n is simply the empirical average, which converges w.p. 1 to μ_z. As a performance measure, consider the MSE defined by $\mathrm{E}\|\theta_n - \mu_z\|^2$, which equals σ_z^2/n. Integration of partial knowledge based on (12) in this case is equivalent to averaging variables with the same mean but with a reduced variance, and the MSE still approaches zero at a rate $O(1/n)$. On the other hand, when the step size is constant, θ_n converges in mean to μ_z, but the MSE converges to a non-zero value which, intuitively, is proportional[6] to the variance σ_z^2. Any variance reduction in this case would thus prove valuable.

The use of a constant step size, though clearly undesirable in the preceding example, is quite common in RL applications, as it allows the iterates to quickly reach a neighborhood of the desired solution, and can cope with time varying environments. In the following discussion, we shall thus focus our analysis on algorithms with a constant step size.

4. TD(0) with Partial Model Knowledge

In this section we apply our IPM method of Equation (12) to the well known model free algorithm Temporal Difference (TD(0); Sutton and Barto, 1998). The simplicity of TD(0) allows us to mathematically characterize its performance in terms of *convergence rate*, and to quantify the impact of using the IPM method on it. The mathematical results we derive specifically for TD(0) are also characteristic of more complex algorithms, as will be shown in subsequent sections.

4.1 Definitions

Throughout this section, we assume that the agent's policy μ is deterministic and fixed, mapping a specific action to each state, denoted by $u(x)$.

4.1.1 VALUE FUNCTION ESTIMATION

Letting $0 < \gamma < 1$ denote a discount factor, define the value function for state x under policy μ as the expected discounted return when starting from state x and executing policy μ

$$V^{\mu}(x) \triangleq \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t) \,\middle|\, x_0 = x\right].$$

Since in this section the policy μ is constant, from now on we omit the superscript μ in $V^{\mu}(x)$, and the subscript μ in P_{μ} , π_{μ}, and Π_{μ}.

The value function is a vector of size $|\mathcal{X}|$. When the state space is large, Function Approximation (FA) is often used to find an approximation to the value function in a subspace of size $L < |X|$. Linear

6. A precise value is given in the next section.

FA is implemented as follows. Given a set of $|X|$ linearly independent basis vectors $\phi(x) \in \mathbb{R}^L$, the goal is to find an approximation to $V(x)$, denoted by $\hat{V}(x,\theta)$ and defined as

$$\hat{V}(x,\theta) = \phi(x)^T \theta.$$

Note that the tunable parameter θ in this case is a vector of L linear weights. In vector form we write $\hat{V}(\theta) = \Phi\theta$, where $\Phi \in \mathbb{R}^{|X|\times L}$ is a matrix composed of rows of basis vectors.

Define the *Temporal Difference* (TD) at time n as

$$d_n \triangleq r(x_n) + \gamma\phi(x_{n+1})^T\theta_n - \phi(x_n)^T\theta_n.$$

For some small step size ε, the fixed step TD(0) algorithm updates θ online in the following manner,

$$\theta_{n+1} = \theta_n + \varepsilon d_n \phi(x_n). \tag{14}$$

This is an SA algorithm as defined in (8), and its associated ODE is (Bertsekas and Tsitsiklis, 1996, Lemma 6.5)

$$\frac{d\theta}{dt} = b + A\theta. \tag{15}$$

where

$$\begin{aligned} A &\triangleq \Phi^T \Pi (\gamma P - I)\Phi, \\ b &\triangleq \Phi^T \Pi r. \end{aligned} \tag{16}$$

Equation (15) is linear and has a fixed point θ^* that satisfies

$$A\theta^* = -b.$$

Furthermore, the eigenvalues of A all have a negative real part (Bertsekas and Tsitsiklis, 1996, Lemma 6.6b), and therefore θ^* is a unique and stable fixed point.

4.1.2 INTEGRATED PARTIAL MODEL TD(0)

We now use the method developed in Section 3 to integrate a partial model into the TD(0) algorithm. Since the policy is deterministic we drop the u subscript in the known set definition. Using (12) and (13) we define IPM-TD(0)

$$\begin{aligned} \theta^K_{n+1} &= \theta^K_n + \varepsilon d^K_n \phi(x_n), \\ d^K_n &\triangleq r(x_n) + \gamma\left(\mathbf{1}^K_{n+1} F^K_n + \mathbf{1}^{\bar{K}}_{n+1}\phi(x_{n+1})^T\theta^K_n\right) - \phi(x_n)^T\theta^K_n, \\ F^K_n &\triangleq \frac{\sum_{y\in K_{x_n}} P(y|x_n)\phi(y)^T\theta^K_n}{\sum_{y\in K_{x_n}} P(y|x_n)}. \end{aligned} \tag{17}$$

Using Theorem 3 we conclude that the IPM-TD(0) iterates have the same characteristic ODE as the TD(0) iterates of Equation (14), and therefore converge to the same fixed point θ^* of (15).

After establishing that the asymptotic trajectory (or, in other words the algorithmic 'function') of the algorithm remains intact, we shall now investigate whether adding the partial knowledge can be guaranteed to improve *performance*.

4.2 Performance Improvement Proof

In this section we prove that the performance of the IPM-TD(0) iteration is superior in terms of asymptotic MSE to regular TD(0). The formal approach we follow here may be carried out for other SA algorithms as well, though the expressions involved may become more complicated.

Recall that the asymptotic limit point of both regular TD(0) and IPM-TD(0) is θ^*. A natural performance measure in this case is the asymptotic MSE defined by

$$\lim_{n\to\infty} \mathrm{E}\left\|\theta_n - \theta^*\right\|^2.$$

The remainder of this section is devoted to showing that integrating a partial model reduces the asymptotic MSE, namely

$$\lim_{n\to\infty} \mathrm{E}\left\|\theta_n^K - \theta^*\right\|^2 < \lim_{n\to\infty} \mathrm{E}\left\|\theta_n - \theta^*\right\|^2,$$

whenever the known set K is not null.

By Lemma 2, at each iteration step we are guaranteed (as long as our partial model is not null) a reduction in the noise variance. This clearly indicates that some improvement in the asymptotic MSE can be expected, but a precise quantification of this is more complicated. A powerful tool for this task is the rate of convergence theory for SA (or a 'limit theorem for fluctuations', as termed by Borkar, 2008). In their treatment of rate of convergence, Kushner and Yin (2003, p. 315) discuss the properties of the sequence

$$\rho_n \triangleq \left(\theta_n - \theta^*\right)/\sqrt{\varepsilon}. \tag{18}$$

Application of their Theorem 10.1.3 to the TD(0) iteration results in the following theorem.

Theorem 4 *The sequence ρ_n converges in distribution, as $\varepsilon \to 0$ and $n \to \infty$ such that $n\varepsilon \to \infty$,*[7] *to a normally distributed random variable, which is the stationary distribution of the stochastic differential equation*

$$dU = AU\,dt + dW. \tag{19}$$

A is defined in (16)*, and W is a Wiener process with covariance matrix $\Sigma = \Sigma_0 + \Sigma_1 + \Sigma_1^T$ where*

$$\Sigma_0 = \lim_{n\to\infty} E\left[\left(d_n\phi\left(x_n\right)\right)\left(d_n\phi\left(x_n\right)\right)^T \middle| \theta_n = \theta^*\right], \tag{20}$$

$$\Sigma_1 = \sum_{j=1}^{\infty} \lim_{n\to\infty} E\left[\left(d_n\phi\left(x_n\right)\right)\left(d_{n+j}\phi\left(x_{n+j}\right)\right)^T \middle| \theta_n = \theta_{n+j} = \theta^*\right].$$

For the IPM iteration (17) *we have Σ_0^K, Σ_1^K where d_n^K replaces d_n in* (20)*.*

The proof of Theorem 4 consists of verifying a lengthy set of technical assumptions required for Theorem 10.1.3 of Kushner and Yin (2003), and is fully described in Appendix E.

The stationary solution to (19) is normally distributed with zero mean and covariance R, which can be easily computed (Papoulis and Pillai, 2002, §9.2) by observing that (19) describes Gaussian white noise filtered through a linear system, leading to

$$R = \lim_{t\to\infty} e^{At}\left\{\int_0^t e^{-As}\Sigma\left(e^{-As}\right)^T ds\right\}\left(e^{At}\right)^T. \tag{21}$$

7. We retain this assumption on ε and n in the sequel.

Let $\{\lambda_i\}_{i=1}^L$ denote the eigenvalues of A, which all have a negative real part (Bertsekas and Tsitsiklis, 1996, Lemma 6.6b), and let Γ be its diagonalizing matrix, that is, $A = \Gamma\Lambda\Gamma^{-1}$ where Λ is diagonal. Also, define a matrix $\chi \in \mathbb{R}^{L\times L}$ such that $[\chi]_{ij} = -1/(\lambda_i + \lambda_j)$. The limit in (21) can be written as

$$R = \lim_{t\to\infty} \Gamma \left\{ \int_0^t e^{\Lambda(t-s)} \Gamma^{-1} \Sigma \left(\Gamma^{-1}\right)^T e^{\Lambda(t-s)} ds \right\} \Gamma^T. \tag{22}$$

Note that the term in the curly brackets in (22) is a matrix, with its (i,j)th component equal to

$$\begin{aligned} & \int_0^t e^{\lambda_i(t-s)} \left[\Gamma^{-1}\Sigma\left(\Gamma^{-1}\right)^T\right]_{i,j} e^{\lambda_j(t-s)} ds \\ = \quad & \left[\Gamma^{-1}\Sigma\left(\Gamma^{-1}\right)^T\right]_{i,j} (\lambda_i+\lambda_j)^{-1} \left(-1 + e^{(\lambda_i+\lambda_j)t}\right). \end{aligned}$$

Substituting in (22) and taking the limit gives

$$R = \Gamma\left(\chi \circ \left(\Gamma^{-1}\Sigma\left(\Gamma^{-1}\right)^T\right)\right)\Gamma^T,$$

and using (18) and Theorem 4, the limit of the MSE is

$$\mathrm{E}\,\|\theta_n - \theta^*\|^2 \to \varepsilon \mathrm{Tr}\left[\Gamma\left(\chi \circ \left(\Gamma^{-1}\Sigma\left(\Gamma^{-1}\right)^T\right)\right)\Gamma^T\right]. \tag{23}$$

The difference in MSE between the original iterate (14) and the IPM iterate (17) lies in the difference between Σ_0, Σ_0^K and Σ_1, Σ_1^K. We now derive explicit expressions for these matrices. In the following, for clarity we adopt the following notation. Let x' denote the state following x. Let $P(K_x) = \sum_{x'\in K_x} P(x'|x)$, and let $P_{K_x}(x')$ denote the probability measure over the known transitions from state x, namely

$$P_{K_x}(x') \triangleq \begin{cases} P(x'|x)/P(K_x) & if \quad x' \in K_x \\ 0 & if \quad x' \notin K_x \end{cases}.$$

Denote by $\mathrm{E}_K[f(x')|x]$, $\mathrm{Var}_K[f(x')|x]$, and $\mathrm{Cov}_K[f(x')|x]$ the expectation, variance, and covariance matrix of some function f of x' given that the current state is x, under the probability measure P_{K_x}.

Lemma 5 *We have* $\Sigma_0 = \Sigma_0^K + \gamma^2 \sum_x [\pi]_x \phi(x) \mathrm{Var}_K\left[\phi(x')^T\theta^* \middle| x\right] \phi(x)^T$.

Proof See Appendix B. ■

In order to simplify calculations, in the remainder of the analysis we deal with a table based algorithm.

Assumption 6 *The algorithm is table based, namely* $\Phi = I$.

Under the table based assumption, the temporal difference terms at subsequent times are not correlated, leading to the following result.

Proposition 7 *Under assumption 6 we have* $\Sigma_1 = \Sigma_1^K = 0$.

Proof For a table based case, θ^* satisfies Bellman's equation for a fixed policy (Bertsekas and Tsitsiklis, 1996)

$$\theta^*(x) = r(x) + \gamma \mathrm{E}\left[\theta^*\left(x'\right)\middle| x\right]. \tag{24}$$

Now, for every j we have

$$\begin{aligned}
& \mathrm{E}\left[\left(d_n \phi(x_n)\right)\left(d_{n+j}\phi(x_{n+j})\right)^T \middle| \theta_n = \theta_{n+j} = \theta^*\right] \\
= \; & \mathrm{E}\left[\left(r(x_n) + \gamma\theta^*(x_{n+1}) - \theta^*(x_n)\right)\left(r(x_{n+j}) + \gamma\theta^*(x_{n+j+1}) - \theta^*(x_{n+j})\right)\right] \\
= \; & \mathrm{E}\left[\mathrm{E}\left[\left(r(x_n) + \gamma\theta^*(x_{n+1}) - \theta^*(x_n)\right)\left(r(x_{n+j}) + \gamma\theta^*(x_{n+j+1}) - \theta^*(x_{n+j})\right)\middle| x_n, \ldots, x_{n+j}\right]\right] \\
= \; & 0,
\end{aligned}$$

where the last equation follows from (24). Thus, in the expression for Σ_1, every element in the sum is zero. For Σ_1^K we can use Lemma 1 to obtain the same result. ■

Generalizing these results to the FA case involves analysis of the correlations in Σ_1, Σ_1^K and is deferred to future work. Nevertheless, we provide numerical simulations with FA that demonstrate similar behavior to the table based case.

Let Δ_Σ denote the diagonal matrix defined by

$$\begin{aligned}
\Delta_\Sigma &\triangleq \Sigma_0 - \Sigma_0^K \\
&= \gamma^2 \sum_x [\pi]_x \phi(x) P(K_x) \mathrm{Var}_K\left[\phi(x')^T \theta^* \middle| x\right] \phi(x)^T.
\end{aligned}$$

Substituting $\Phi = I$ gives a simple expression for the diagonal elements of Δ_Σ

$$[\Delta_\Sigma]_{xx} = \gamma^2 [\pi]_x P(K_x) \mathrm{Var}_K\left[\theta^*\left(x'\right)\middle| x\right].$$

Note that Δ_Σ has no negative elements. We are interested in the difference in asymptotic MSE, which, based on (23) is given by

$$\begin{aligned}
\delta_{MSE} &= \mathrm{E}\|\theta_n - \theta^*\|^2 - \mathrm{E}\|\theta_n^K - \theta^*\|^2 \\
&\to \varepsilon \cdot \mathrm{Tr}\left[\Gamma\left(\chi \circ \left(\Gamma^{-1}\Delta_\Sigma\left(\Gamma^{-1}\right)^T\right)\right)\Gamma^T\right].
\end{aligned} \tag{25}$$

If the known set is not null, then δ_{MSE} is positive (it can be seen as the asymptotic MSE of an iterate with the same matrix A, but with Δ_Σ instead of Σ_0, which by definition is positive), and thus the algorithm's performance improves. We summarize this result in the following theorem.

Theorem 8 *Consider the table based online TD(0) iterate for* θ_n *described by* (14) *with* $\Phi = I$*, and the IPM-TD(0) iterate for* θ_n^K *described by* (17) *with the same requirement on* Φ*. Assuming that there is at least one state* $x \in \mathcal{X}$ *such that* $P(K_x)\mathrm{Var}_K\left[\theta^*(x')|x\right] > 0$*, then the asymptotic MSE of the iterates satisfy* $\lim_{n\to\infty} E\|\theta_n^K - \theta^*\|^2 = \lim_{n\to\infty} E\|\theta_n - \theta^*\|^2 - \delta_{MSE}$*, where* δ_{MSE} *is given in* (25)*, and* $\delta_{MSE} > 0$*.*

Theorem 8 therefore assures us that the reduction in noise variance at each step, guaranteed by Lemma 2, translates into a reduction in the overall error of the algorithm. Note that the simple dependence of the MSE on ε allows for a different interpretation of the performance in terms of *convergence rate* - for some desired MSE, the partial knowledge allows us to use a larger step size ε, and thus *converge faster*. This issue will also be demonstrated in simulation.

We comment on a decreasing step size. For a step size of the form $\varepsilon_n = 1/n^{\alpha}$, $0.5 < \alpha \leq 1$, a similar analysis can be performed with ρ_n defined as $\rho_n = n^{\alpha/2}(\theta_n - \theta^*)$. In this case, θ_n converges to θ^* w.p. 1, and the MSE decreases to zero at a rate $O\left(n^{-\alpha/2}\right)$. Integrating a partial model in this case will reduce fluctuations in the converging path of the system. The performance gain of integrating a partial model is therefore more pronounced when the step size is constant.

4.3 Numerical Simulations of IPM-TD(0)

We conclude this section with a demonstration of the performance of the IPM-TD(0) algorithm, and a comparison with the theory established above.[8]

Our simulations are on a set of abstract randomly constructed MDP's termed Generalized Average Reward Non-stationary Environment Test-bench or in short GARNET (Bhatnagar et al., 2007). GARNET MDP's comprise a class of randomly constructed finite MDP's serving as a test-bench for RL algorithms. A GARNET MDP is characterized in our case by four parameters and is denoted by GARNET$(|\mathcal{X}|, |\mathcal{U}|, B, \sigma)$. The parameter $|\mathcal{X}|$ is the number of states in the MDP, $|\mathcal{U}|$ is the number of actions, B is the branching factor of the MDP, that is, the number of uniformly distributed non-zero entries in each line of the MDP's transition matrices, and the reward in each state is normally distributed with variance σ. For each GARNET MDP we also construct a 'partially known' MDP characterized by a parameter p_K, $0 \leq p_K \leq 1$ such that each transition in the original MDP is known w.p. p_K. The value of p_K therefore indicates our level of knowledge about the MDP, ranging from no knowledge at all ($p_K = 0$) up to knowing the complete MDP ($p_K = 1$).

For a GARNET$(10,5,10,1)$ MDP, a random deterministic policy was chosen and its value function was evaluated using algorithm (17). The error $\|\theta_n^K - \theta^*\|^2$, averaged over 500 different runs with the same initial conditions, is plotted in Figure 1 (left) for different values of p_K. The asymptotic MSE was calculated using (23) and is shown for comparison. In Figure 1 (middle), the step size for an iteration with partial knowledge was set such that the asymptotic MSE would match that of the iteration without partial knowledge. As can be seen, this caused the IPM iteration to converge *faster*.

For the next simulation a linear FA was used, with basis vectors $\phi(x) \in \{0,1\}^L$, where the number of nonzero values in each $\phi(x)$ is l. The nonzero values were chosen uniformly at random, with any two states having different feature vectors. Figure 1 (right) shows the error $\|\theta_n^K - \theta^*\|^2$for a GARNET$(30,5,10,1)$ MDP, where we used linear FA with $L = 10$ and $l = 2$. As can be seen, the behavior observed in the tabular case is characteristic of the FA case as well.

5. Inaccuracy of the Partial Model

Until now, we have assumed that our partial model contained accurate probability ratios. Obviously, such a strong assumption is not realistic, and in any practical situation our partial knowledge would contain some degree of error. In this section we consider the effect of inaccuracies in the partial model on the performance of the IPM-TD(0) method. Specifically, our goal is to show that if the inaccuracy in the partially known model is small enough, then an improvement in performance over regular TD(0) can still be guaranteed, and we seek bounds on the error in the algorithm induced by the inaccuracy in the model. Before we go into mathematical detail we first describe our conceptual approach.

8. The code for generating the results presented here can be found at the author's web site.

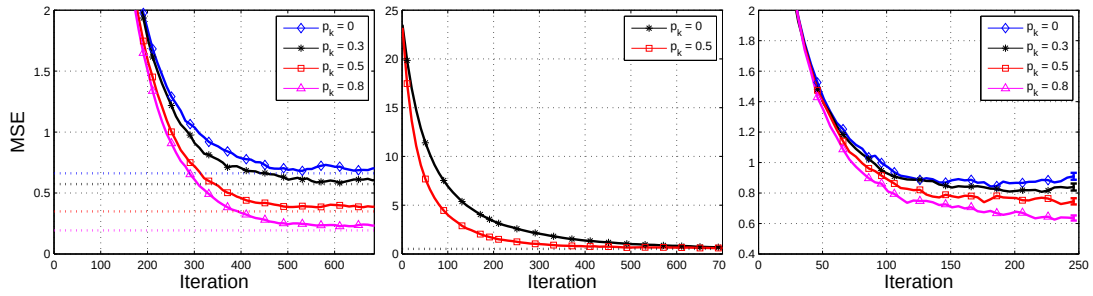

Figure 1: TD(0) with a Partial Model. *Left* : MSE of Table Based IPM-TD(0) on a GARNET(10,5,10,1) MDP with a deterministic random policy, for different values of p_K. Step size is $\varepsilon = 0.2$. Dashed lines show the asymptotic MSE calculated by (23). *Middle* : MSE of Table Based IPM-TD(0) on a GARNET(10,5,10,1) MDP with a deterministic random policy. For $p_K = 0$ (black-solid) a step size $\varepsilon = 0.15$ was used, and the asymptotic MSE was calculated using (23) (black-dashed). For $p_K = 0.5$ (red-solid) a step size was calculated (using (23)) such that its asymptotic MSE would equal that of $p_K = 0$. *Right* : MSE of linear FA IPM-TD(0) on a GARNET(30,5,10,1) MDP with a deterministic random policy, for different values of p_K. Step size is $\varepsilon = 0.15$. The linear FA parameters are $L = 10$ and $l = 2$. A discount factor of $\gamma = 0.7$ was used in all simulations. All results are averaged over 500 different runs with the same initial conditions. Error bars display the standard error of the mean; for clarity of presentation the bars are displayed only for the last iteration.

A key point in the analysis of IPM-TD(0) in Section 4 was that since the estimator $\hat{\mu}_K$ in (5) is unbiased, then the ODE of the stochastic approximation does not change, and asymptotically the algorithm concentrates around its fixed point which is the true value function. This is no longer valid when the partial model is not accurate, as the inaccuracy induces a bias in $\hat{\mu}_K$. Since we use the estimator at every time step, this bias may accumulate, and the crucial question here is how it affects the algorithm asymptotically, and whether it can be guaranteed that small model errors do not cause a large deviation from the true value function.

The improvement in performance of IPM-TD(0) relied on the variance reduction property of $\hat{\mu}_K$. We shall see that if the inaccuracy in the partial model is small enough, then this property can still be guaranteed.

Thus, our analysis consists of investigating the *bias* and the *variance* of IPM-TD(0) with an inaccurate model. As we have done earlier, we first describe some results in the context of estimating the mean of a random variable, and later extend the results to the MDP setting.

5.1 Estimation of a Random Variable Mean

Consider the definitions of Section (2), and let $\left\{\hat{P}(\omega)\right\}_{\omega\in\Omega}$ denote inaccurate probabilities, obtained by some means. For some $\varepsilon > 0$ we define an $\varepsilon-$known set K^ε by

$$K^\varepsilon \triangleq \left\{\omega : \omega \in \Omega \text{ s.t. } \left|\hat{P}_{K^\varepsilon}(\omega) - P_{K^\varepsilon}(\omega)\right| < \varepsilon\right\}, \tag{26}$$

where the probability measures $\hat{P}_{K^\varepsilon}$ and P_{K^ε} are defined by $\hat{P}_{K^\varepsilon}(\omega) \triangleq \hat{P}(\omega) / \sum_{\omega' \in K^\varepsilon} \hat{P}(\omega')$ and $P_{K^\varepsilon}(\omega) \triangleq P(\omega) / \sum_{\omega' \in K^\varepsilon} P(\omega')$, respectively. Also denote by $\hat{\mathrm{E}}_{K^\varepsilon}[\cdot]$ and $\hat{\mathrm{Var}}_{K^\varepsilon}[\cdot]$ the expectation and variance under the measure $\hat{P}_{K^\varepsilon}$, and by $\mathrm{E}_{K^\varepsilon}[\cdot]$ and $\mathrm{Var}_{K^\varepsilon}[\cdot]$ the expectation and variance under the measure P_{K^ε}.

We motivate the definition of the $\varepsilon-$known set with an example. Let $\{x^i\}_{i=1}^n$ denote i.i.d. samples of X. For some set $K \subset \Omega$ let $\hat{P}_K(\omega)$ denote the count ratios in K

$$\hat{P}_K(\omega) = \begin{cases} \frac{\sum_{i=1}^n \mathbf{1}(x^i=\omega)}{\sum_{i=1}^n \mathbf{1}(x^i \in K)} & \text{for } \omega \in K \\ 0 & \text{else} \end{cases},$$

where $\mathbf{1}$ is the indicator function. It can be shown that $\hat{P}_K(\omega)$ is an unbiased estimate of $P_K(\omega)$, and by the law of large numbers we have that for large n, the difference $\left|\hat{P}_K(\omega) - P_K(\omega)\right|$ is small. Furthermore, for a finite n, Chernoff type bounds can be used to bound this difference with high probability by some small ε, motivating definition (26).

An estimator for μ that uses the $\varepsilon-$known set is derived by plugging K^ε instead of K in (5)

$$\hat{\mu}_{K^\varepsilon}(x) = \mathbf{1}_x^{K^\varepsilon} \hat{\mathrm{E}}_{K^\varepsilon}[X] + \mathbf{1}_x^{\bar{K}^\varepsilon} x. \tag{27}$$

Note that since the known set is not accurate, the estimator (27) is no longer unbiased. The following theorem, which we prove in Appendix C, bounds the bias and variance of $\hat{\mu}_{K^\varepsilon}(x)$.

Theorem 9 *The bias of $\hat{\mu}_{K^\varepsilon}(x)$ satisfies*

$$\left|\mathrm{E}\left[\hat{\mu}_{K^\varepsilon}(X)\right] - \mathrm{E}[X]\right| \leq \varepsilon P(K^\varepsilon) \sum_{x \in K^\varepsilon} |x|.$$

The variance of $\hat{\mu}_{K^\varepsilon}(x)$ satisfies

$$\begin{aligned} Var\left[\hat{\mu}_{K^\varepsilon}(X)\right] \leq\ & Var[X] - P(K^\varepsilon) \cdot Var_{K^\varepsilon}[X] \\ & + \varepsilon P(K^\varepsilon) \left(\varepsilon \left(\sum_{x \in K^\varepsilon} |x| \right)^2 + 2 \left(\sum_{x \in K^\varepsilon} |x| \right) \left| E_{K^\varepsilon}[X] - E[X] \right| \right). \end{aligned} \tag{28}$$

5.2 Error Bound for IPM-TD(0)

We now derive asymptotic error bounds for IPM-TD(0) with a constant stepsize $\tilde{\varepsilon}$, when the partial model is inaccurate. We treat only the table based algorithm

$$\begin{aligned} \theta_{n+1}(x_n) &= \theta_n(x_n) + \tilde{\varepsilon} d_n^K, \\ d_n^K &\triangleq r(x_n) + \gamma \left(\mathbf{1}_{n+1}^{K^\varepsilon} \hat{\mathrm{E}}_{K^\varepsilon_{x_n}}\left[\theta_n(x_{n+1}) | x_n\right] + \mathbf{1}_{n+1}^{\bar{K}^\varepsilon} \theta_n(x_{n+1}) \right) - \theta_n(x_n), \end{aligned} \tag{29}$$

where $\hat{\mathrm{E}}_{K^\varepsilon_{x_n}}\left[\theta_n(x_{n+1}) | x_n\right]$ denotes expectation under the probability measure in the $\varepsilon-$known set $K^\varepsilon_{x_n}$

$$K_x^\varepsilon \triangleq \left\{ y : y \in \mathcal{X} \text{ s.t. } \left| \frac{\hat{P}(y|x)}{\sum_{y' \in K_x^\varepsilon} \hat{P}(y'|x)} - \frac{P(y|x)}{\sum_{y' \in K_x^\varepsilon} P(y'|x)} \right| < \varepsilon \right\}. \tag{30}$$

The ODE for (29) can be written as

$$\frac{d\theta}{dt} = \Pi\left(r + \left(\gamma(P+\delta P) - I\right)\theta\right), \tag{31}$$

where

$$\delta P_{ij} = \begin{cases} \left(\sum_{k\in K_i^\varepsilon} P(k|i)\right)\left(\frac{\hat{P}(j|i)}{\sum_{k\in K_i^\varepsilon}\hat{P}(k|i)} - \frac{P(j|i)}{\sum_{k\in K_i^\varepsilon}P(k|i)}\right), & j \in K_i^\varepsilon, \\ 0, & j \notin K_i^\varepsilon. \end{cases} \tag{32}$$

Recalling that the true value function satisfies $\theta^* = (I-\gamma P)^{-1} r$, the asymptotic limit point of the ODE (31) is denoted by $\theta^* + \delta\theta$, and satisfies

$$\theta^* + \delta\theta = (I - \gamma(P+\delta P))^{-1} r. \tag{33}$$

In the next subsection we show how to bound the error term $\delta\theta$.

5.2.1 A BOUND ON THE BIAS

We would like to bound the term $\delta\theta$, which is the error in the value function, and can be seen as the total *bias* induced by the IPM method with the inaccurate model. Note that (33) describes a perturbed linear system (Horn and Johnson, 1985, §5). Using tools for dealing with such systems, we can bound the error as presented in the following theorem.

Theorem 10 *Let $K_{\max}^\varepsilon$ denote the cardinality of the largest $\varepsilon-$known set K_x^ε,*

$$K_{\max}^\varepsilon = \max_x |K_x^\varepsilon|, \tag{34}$$

and let ε satisfy $\varepsilon < \frac{1-\gamma}{\gamma K_{\max}^\varepsilon}$. Then the maximal error in the ODE limit is bounded by

$$\frac{\|\delta\theta\|_\infty}{\|\theta^*\|_\infty} \le \frac{\kappa}{1-\kappa\cdot\left(\frac{K_{\max}^\varepsilon\varepsilon}{1-\gamma}\right)}\left(\frac{K_{\max}^\varepsilon\varepsilon}{1-\gamma}\right), \tag{35}$$

where κ satisfies

$$\kappa \le \frac{1+\gamma}{1-\gamma}. \tag{36}$$

Proof For a matrix norm $\|\cdot\|_p$, if $\left\|(I-\gamma P)^{-1}\right\|_p \|\gamma\delta P\|_p < 1$ we have (Horn and Johnson, 1985, 5.8.8)

$$\frac{\|\delta\theta\|_p}{\|\theta^*\|_p} \le \frac{\kappa(I-\gamma P)}{1-\kappa(I-\gamma P)\left(\|\gamma\delta P\|_p / \|I-\gamma P\|_p\right)}\frac{\|\gamma\delta P\|_p}{\|I-\gamma P\|_p}, \tag{37}$$

where κ is the matrix condition number $\kappa(A) = \left\|A^{-1}\right\|_p \|A\|_p$. We now bound each of the terms on the right hand side of (37). We use the norm $\|\cdot\|_\infty$ which is induced by the max vector norm, and can be alternatively defined (Horn and Johnson, 1985, 5.6.5) as the *maximum row sum* matrix norm

$$\|\delta P\|_\infty \triangleq \max_i \sum_j \left|P_{ij}\right|. \tag{38}$$

From this definition, (32), (34), and (30), it is clear that

$$\|\gamma\delta P\|_\infty < \gamma K^\varepsilon_{\max}\varepsilon. \tag{39}$$

Theorem 5.6.9 in Horn and Johnson (1985) asserts that for any matrix norm we have

$$\|A\| \geq \rho(A),$$

where $\rho(A)$ is the spectral radius of A

$$\rho(A) \triangleq \max\{|\lambda| : \lambda \text{ is an eigenvalue of } A\}.$$

For the matrix $I-\gamma P$ we have

$$\rho(I-\gamma P) > 1-\gamma,$$

since P is stochastic and thus its largest eigenvalue is 1. Using (39) we therefore have

$$\frac{\|\gamma\delta P\|_\infty}{\|I-\gamma P\|_\infty} \leq \frac{\gamma K^\varepsilon_{\max}\varepsilon}{1-\gamma}.$$

We now bound $\kappa(I-\gamma P) = \|I-\gamma P\|_\infty \left\|(I-\gamma P)^{-1}\right\|_\infty$. First, by the triangle equality we have

$$\|I-\gamma P\|_\infty \leq \|I\|_\infty + \gamma\|P\|_\infty = 1+\gamma, \tag{40}$$

since P is a stochastic matrix and by definition (38) we have $\|P\|_\infty = 1$. Next we have by definition of the induced norm

$$\left\|(I-\gamma P)^{-1}\right\|_\infty = \max_{\|r\|_\infty=1} \left\|(I-\gamma P)^{-1} r\right\|_\infty \leq \frac{1}{1-\gamma}, \tag{41}$$

since $(I-\gamma P)^{-1} r$ can be seen as the value function associated with a reward vector r, which can have a maximum value of $r_{\max}/(1-\gamma)$. From (40) and (41) we have

$$\kappa(I-\gamma P) \leq \frac{1+\gamma}{1-\gamma}.$$

All is left is to verify that $\left\|(I-\gamma P)^{-1}\right\|_\infty \|\gamma\delta P\|_\infty < 1$. Using (39) and (41) this is satisfied if

$$\varepsilon < \frac{1-\gamma}{\gamma K^\varepsilon_{\max}}.$$

■

The bound in (35) can be simplified when ε is small, as described in the following corollary.

Corollary 11 *For small enough* ε *we have*

$$\frac{\|\delta\theta\|_\infty}{\|\theta^*\|_\infty} \leq \frac{K^\varepsilon_{\max}\varepsilon(1+\gamma)}{(1-\gamma)^2} + O\left(\varepsilon^2\right).$$

Proof Substitute (36) in (35), where for small ε we have $\varepsilon/\left(1-\frac{\kappa K^\varepsilon_{\max}\varepsilon}{1-\gamma}\right) = \varepsilon + O\left(\varepsilon^2\right)$. ■

The bounds in Theorem 10 and Corollary 11 show that the accumulated bias induced by the model inaccuracies is linear in ε, but also in $K^\varepsilon_{\max}$, thus it is preferred to keep the $\varepsilon-$known set for each state relatively small, as each element in the set contributes to an accumulated error. Note that the term $1/(1-\gamma)^2$ is a consequence of the fact that the estimation bias for each state now accumulates over a whole trajectory.

5.2.2 A BOUND ON THE VARIANCE

We shall now bound the variance of IPM-TD(0). We follow directly the analysis of Section 4 for IPM-TD(0) with an accurate model, and note that the only difference[9] is in the calculation for the term $\Sigma_0^{K^\varepsilon}$, which, using (33), becomes

$$\begin{aligned}\left[\Sigma_0^{K^\varepsilon}\right]_{xx} &= \mathrm{E}\left[(d_n^K)^2 \middle| \theta = \theta^* + \delta\theta, x_n = x\right] \\ &= \gamma^2 \mathrm{Var}\left[\left(\mathbf{1}_{n+1}^{K^\varepsilon} \hat{\mathrm{E}}_{K^\varepsilon_{x_n}}\left[\left[\theta^* + \delta\theta\right](x_{n+1}) | x_n\right] + \mathbf{1}_{n+1}^{\bar{K}^\varepsilon}\left[\theta^* + \delta\theta\right](x_{n+1})\right) \middle| x_n = x\right].\end{aligned}$$

In the following we focus on bounding the term on the right hand side, and show sufficient conditions under which $\left[\Sigma_0^{K^\varepsilon}\right]_{xx} < [\Sigma_0]_{xx}$ for every $x \in |\mathcal{X}|$. For notational simplicity, we drop the dependence on x, and treat a single random variable taking values in $\{[\theta^* + \delta\theta]_i\}_{i=1}^{|\mathcal{X}|}$, with the appropriate probabilities $P(x'|x)$. Thus, we need to bound

$$\gamma^{-2}\left[\Sigma_0^{K^\varepsilon}\right]_{xx} = \mathrm{Var}\left[\mathbf{1}^{K^\varepsilon}\hat{\mathrm{E}}_{K^\varepsilon}\left[\theta^* + \delta\theta\right] + \mathbf{1}^{\bar{K}^\varepsilon}\left(\theta^* + \delta\theta\right)\right], \tag{42}$$

and compare to $[\Sigma_0]_{xx}$. The following theorem, which is proved in Appendix D, bounds $\gamma^{-2}\left[\Sigma_0^{K^\varepsilon}\right]_{xx}$.

Theorem 12 *Let* $b_x \triangleq \sum_{x \in K_x^\varepsilon} |[\theta^*]_x|$, *and* $c_x \triangleq \left|E_{K_x^\varepsilon}[\theta^*] - E[\theta^*]\right|$, *and assume that* $\max\{\delta\theta\} \le \eta$. *Then the elements of the diagonal matrix* Σ_0^K *satisfy*

$$\begin{aligned}\gamma^{-2}\left[\Sigma_0^{K^\varepsilon}\right]_{xx} \le\ & \gamma^{-2}[\Sigma_0]_{xx} - P(K_x^\varepsilon)\cdot \mathrm{Var}_{K_x^\varepsilon}[\theta^*] + \eta^2 + 2\eta\gamma^{-1}\sqrt{[\Sigma_0]_{xx}} \\ & + \varepsilon P(K_x^\varepsilon)\left(\varepsilon b_x^2\left(1 - P(K_x^\varepsilon)\right) + 2b_x c_x\right).\end{aligned} \tag{43}$$

Using our previous bound on the bias we have the following corollary.

Corollary 13 *For small enough* ε *we have*

$$\begin{aligned}\gamma^{-2}\left([\Sigma_0]_{xx} - \left[\Sigma_0^{K^\varepsilon}\right]_{xx}\right) \ge\ & P(K_x^\varepsilon)\cdot \mathrm{Var}_{K_x^\varepsilon}[\theta^*] \\ & - \varepsilon\frac{K_{\max}^\varepsilon(1+\gamma)\|\theta^*\|_\infty}{(1-\gamma)^2}\left(\frac{K_{\max}^\varepsilon\varepsilon(1+\gamma)\|\theta^*\|_\infty}{(1-\gamma)^2} + \frac{2}{\gamma}\sqrt{[\Sigma_0]_{xx}}\right) \\ & - \varepsilon P(K_x^\varepsilon)\left(\varepsilon b_x^2\left(1 - P(K_x^\varepsilon)\right) + 2b_x c_x\right).\end{aligned}$$

Proof Apply Corollary 11 to bound η in Theorem 12. ■

The following corollary translates the previously established bounds to a performance improvement guarantee.

Corollary 14 *For a small enough* ε *an improvement in the asymptotic MSE of IPM-TD(0) can be guaranteed.*

9. Note that the $\Sigma_1^{K^\varepsilon}$ term is still zero, for the same reasons described in Section 4 .

Proof By Corollary 13 ε can be chosen such that for every x we have

$$\left[\Sigma_0^{K^\varepsilon}\right]_{xx} < \left[\Sigma_0\right]_{xx},$$

thus we can follow the development of the performance improvement proof in Section 4 and conclude that the sequence $(\theta_n^K - \theta^*)/\sqrt{\tilde{\varepsilon}}$ converges in distribution to a Gaussian centered on $\delta\theta$ and with a covariance $\hat{R}$, which is smaller than R (as defined in Section 4). Since we can use Theorem 10 to also bound the bias $\delta\theta$, we can choose ε small enough such that asymptotically $\mathrm{E}\|\theta_n^K - \theta^*\|^2 < \mathrm{E}\|\theta_n - \theta^*\|^2$. ■

To conclude, from Theorem 10 and Theorem 12 we see that for small enough ε, we get a small bias and a reduction in the variance. The specific terms in the bounds can be used to find a suitable ε that guarantees a performance improvement. If the probabilities $\hat{P}$ are obtained empirically from a trajectory of the MDP, then a Chernoff type bound can be used to further bound the number of observations required for a desired ε. This issue is beyond the scope of this paper.

6. Numerical Experiments

In this section, the performance improvement obtained by the IPM method is demonstrated on two different model free RL algorithms,[10] and two different problems. Our goal is to demonstrate both the *generality* of the method, and its *usefulness* in practical applications. Generality is demonstrated by application of the method to two very different RL algorithms - policy gradient and Q-learning. The only common feature to these two algorithms is their representation as a stochastic approximation. Together with the theory presented in previous sections, these results suggest that the IPM method can be applied succesfully to a wide variety of RL algorithms. The usefulness of the method is demonstrated in the solution of a call admission control problem. In this problem, it is shown that values of the partially known MDP (11) capture meaningful physical quantities of the problem, thus, (11) may be seen as the natural representation for partial knowledge in such problems. The performance improvement obtained by the IPM method suggests that it may be used succesfully in practice.

6.1 IPM Q-Learning for Admission Control

In this section we consider the call admission control problem for a single link, which arises when a telecommunication provider wants to sell the limited bandwidth of the link to different types of customers so as to maximize expected long term revenue. In this scenario, the customers differ in their bandwidth demand, the price they pay for its usage, and the frequency of their requests.

When the link is empty, it is reasonable that every customer request should be accepted, as it generates some revenue. On the other hand, when the link is almost full, a clever policy might decide to save the available bandwidth for the more profitable requests, at the expense of rejecting the less profitable ones. Thus, it is clear that a good policy should take into account both the bandwidth demand and profit of each request type, and its arrival frequency. When some of these quantities are not known in advance, a *learning* policy may be employed. Specifically, in the following we consider a case where these quantities are known only for *some* of the request types, which will naturally lead to the use of a partial model in the learning algorithm. Such a scenario can occur

10. The code for generating the results presented here can be found at the author's web site.

when, for example, new customers are added to a system, or when some of the request types have features that change in time.[11]

One approach to designing an admission policy is to formulate the problem as an MDP, for which an optimal policy is well defined, and solve it using RL approaches, as has been done by Marbach et al. (1998) and Marbach and Tsitsiklis (1998). In the following we present this approach, and show that in this problem our partially known MDP definition emerges as a very natural representation for partial model knowledge.

6.1.1 PROBLEM FORMULATION

Consider a service provider with a bandwidth of B units, which supports a finite set $\{1,2,...,M\}$ of different service types. Each service type is characterized by its bandwidth demand $b(m)$, its call arrival rate $\alpha(m)$, and its average holding time $1/\beta(m)$, where we assume that the calls arrive according to independent Poisson processes, and that the holding times are exponentially and independently distributed. Whenever a call of type m arrives, the controller can decide whether to accept or reject it, and if it is accepted and enough bandwidth is available, an immediate reward $c(m)$ is recieved. The objective is to find an admission controller (policy) that maximizes the average return. This problem can be represented by an MDP as follows.

6.1.2 STATE SPACE, CONTROLS, AND REWARD

The configuration of the link is denoted by $s = (s(1),\ldots,s(M))$, where $s(m) \in \{0,1,2,\ldots\}$ denotes the number of customers of type m currently using the link. Transitions between different configurations are triggered by events which we denote by $\omega = \{\omega(1),\ldots,\omega(M)\}$, where $\omega(m) = 1$ if a new customer of type m requests service, $\omega(m) = -1$ if a customer of type m departs from the system, and $\omega(m) = 0$ otherwise. The state x of the system consists of the link configuration together with the event which triggered a transition,

$$x = (s,\omega),$$

and the complete state space is given by

$$\mathcal{X} = \left\{ x = (s,\omega) | \sum s(m)\, b(m) \leq B,\ \sum |\omega(m)| \leq 1 \text{ and } \omega(m) < 0 \text{ only if } s(m) > 0 \right\}.$$

The possible controls are to accept or reject a call, denoted by $u \in \{u_a, u_r\}$, respectively, and the immediate reward is

$$r(x,u) = \begin{cases} c(m) & \text{if } u = u_a,\ \omega(m) = 1,\ (s+\omega,\omega) \in \mathcal{X}, \\ 0 & \text{else.} \end{cases}$$

The goal is to find the optimal policy with respect to the the *average reward*

$$\eta = \mathrm{E}[r(x)].$$

11. We note that a learning policy may be required even when the model is fully known, as finding the optimal policy is often an intractable problem. This 'fully known' scenario may be seen as a special case of the following presentation.

6.1.3 TRANSITION PROBABILITIES

In order to transform the continuous time process into discrete transitions between events, a uniformization technique (Gallager, 1995, §6.4) is used. Define z to be the maximal transition rate, given by

$$z = \max_{x \in \mathcal{X}} \left\{ \sum_{m=1}^{M} \left(\alpha(m) + s(m) \beta(m) \right) \right\}. \tag{44}$$

For a state $x = (s, \omega)$, the probability that the next event is an arrival of a call of type m is equal to $\alpha(m)/z$. The probability for a departure of a call of type m is $\beta(m) s(m)/z$. By normalization, the probability that in the next event nothing happens is $1 - \sum_{m=1}^{M} \left(\alpha(m) + s(m) \beta(m) \right)/z$.

6.1.4 PARTIALLY KNOWN MDP

For this problem, a natural definition for partial model knowledge is through the arrival and departure rates α, β, namely

$$M_K \triangleq \{ m : m \in 1, \ldots, M \text{ s.t. } \alpha(m), \beta(m) \text{ are known} \}.$$

As an example where such partial knowledge arises in practice, consider a case where new jobs (with unknown rates) are added to an existing system (with previously known rates). Note that generally, the values in M_K do not suffice for calculating z in (44), hence the transition probabilities of the MDP are not known. Nevertheless, the key point here is that in the *ratios* between transition probabilities, the z terms cancel out, therefore the partial MDP definition (11) can be satisfied. In particular, letting $i \in M_K$, we have

$$\frac{P(\text{arrival of type } i)}{\sum_{j \in M_K} P(\text{arrival of type } j)} = \frac{\alpha(i)}{\sum_{j \in M_K} \alpha(j)},$$

and similar expressions hold for probabilities of departures.

6.1.5 IPM Q-LEARNING

The model free RL algorithm we use for this problem is a variant of the popular Q-Learning algorithm for average return.[12] For each state-action pair, a Q value is maintained, and updated according to

$$Q_{n+1}(x_n, u_n) = Q_n(x_n, u_n) + \varepsilon_n \left(r(x_n, u_n) + \max_{u'} Q_n\left(x_{n+1}, u'\right) - Q_n(x_n, u_n) - \frac{1}{|\mathcal{X}||\mathcal{U}|} \sum_{x,u} Q_n(x, u) \right). \tag{45}$$

The greedy deterministic policy $u(x)$ w.r.t. the Q values at time n is

$$u(x) = \operatorname*{argmax}_{u'} Q_n\left(x, u'\right).$$

Update (45) is an SA, and was shown to converge (Abounadi et al., 2001) under suitable step sizes ε_n to a fixed point Q^*, such that the greedy policy w.r.t. Q^* is optimal. Applying the IPM method in

12. This is also known as relative value iteration.

Call Type m	1	2	3	4
$\alpha(m)$	1.8	1.4	1.6	1.4
$\beta(m)$	0.4	0.7	0.5	0.4
$b(m)$	1	1	1	1
$c(m)$	1.4	1	1.6	1

Table 1: Call Types

this case simply amounts to replacing $\max_{u'} Q_n(x_{n+1}, u')$ in (45) with

$$\mathbf{1}_{n+1}^{K} \cdot \frac{\sum_{y \in K_{x_n,u_n}} P_{u_n}(y|x_n) \max_{u'} Q_n(y,u')}{\sum_{y \in K_{x_n,u_n}} P_{u_n}(y|x_n)} + \mathbf{1}_{n+1}^{\bar{K}} \cdot \max_{u'} Q_n\left(x_{n+1}, u'\right). \tag{46}$$

We now report on the results of using IPM Q-learning for optimizing a call admission control policy.

6.1.6 RESULTS

In our experiments, we consider a link with a bandwidth of 7 units, and 4 call types. The parameters for each call are summarized in Table 1, and the size of the state space in this configuration is $|\mathcal{X}| = 2490$. IPM Q-Learning was run with initial values $Q_0(x,u) = r(x,u)$ and a step size $\varepsilon_n = \gamma_0 / (\gamma_1 + v_n(x_n,u_n))$, where $v_n(x,u)$ denotes the number of visits to the state action pair (x,u) up to time n. The values of γ_0, γ_1 were manually tuned for optimal performance, and set to $\gamma_0 = \gamma_1 = 40$. The action selection policy while learning was $\varepsilon-$greedy, with $\varepsilon = 0.1$. The partial model for each experiment is represented by a single parameter k, such that the arrival and departure rates of all calls of type $m \leq k$ are known. Figure 2 shows the average reward η as a function of iteration. As can be seen, incorporation of partial model knowledge by the IPM method resulted in a significant performance improvement.

6.2 IPM Policy Gradient

In this experiment simulations were performed on randomly generated MDP's, as described in Section 4.3. In the experiments, the agent maintains a stochastic policy function parametrized by $\theta \in \mathbb{R}^{L \cdot |\mathcal{U}|}$, and given by

$$\mu_\theta(u|x) = e^{\theta^T \xi(x,u)} / \sum_{u'} e^{\theta^T \xi(x,u')},$$

where the state-action feature vectors $\xi(x,u) \in \{0,1\}^{L \cdot |\mathcal{U}|}$ are constructed from the state features $\phi(x)$ defined in Section 4.3 as follows

$$\xi(x,u) \triangleq (0, \ldots (L \times (u-1) \text{ zeros}), \phi(x), 0, \ldots (L \times (|\mathcal{U}| - u) \text{ zeros}))^T.$$

The agent's goal is to find the parameter θ which maximizes the *average reward* $\eta = \mathrm{E}[r(x)]$. Policy Gradient algorithms achieve this goal by estimating the gradient w.r.t. θ of the average reward, $\nabla_\theta \eta$, and performing a stochastic gradient ascent on the parameters to reach a local maximum. One such algorithm was proposed by Marbach and Tsitsiklis (1998). At time n we update the parameter vector θ and a scalar λ which is an estimate of η,

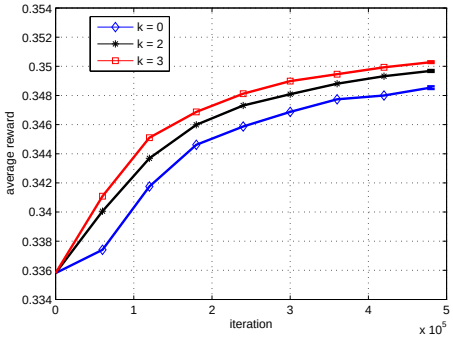

Figure 2: IPM Q-Learning for Admission control. Implementation of IPM Q-Learning, (45) and (46), for the call admission control problem of Table 1. Average reward of the greedy policy is plotted vs. iteration number for different values of k. Results are averaged over 100 different runs with the same initial conditions. Error bars display the standard error of the mean; for clarity of presentation the bars are displayed only for the last iteration.

$$\theta_{n+1} = \theta_n + \varepsilon\left(r\left(x_n\right) - \lambda_n\right) z_n, \tag{47}$$

$$\lambda_{n+1} = \lambda_n + \varepsilon'\left(r\left(x_n\right) - \lambda_n\right), \tag{48}$$

where ε and ε' are step sizes, and $\varepsilon' < \varepsilon$. We then simulate a transition to the next state, and update the vector z by

$$z_{n+1} = z_n + L_{x_n,u_n}\left(\theta_n\right),$$

where $L_{x_n,u_n}\left(\theta_n\right)$ is the likelihood ratio $L_{x,u}\left(\theta\right) = \nabla_\theta \log \mu_\theta(u|x)$. Every time a predefined recurrent state of the MDP is visited, z_{n+1} is reset to zero.

Denote by $\mathbf{1}_n^K$ an indicator function that equals 1 if x_n belongs to $K_{x_{n-1},u_{n-1}}$ and 0 otherwise. Incorporating partial knowledge into the algorithm using (12) simply amounts to replacing $r\left(x_n\right)$ in (47-48) with

$$\mathbf{1}_n^K \cdot \frac{\sum_{y \in K_{x_{n-1},u_{n-1}}} P_{u_{n-1}}(y|x_{n-1}) r(y)}{\sum_{y \in K_{x_{n-1},u_{n-1}}} P_{u_{n-1}}(y|x_{n-1})} + \mathbf{1}_n^{\bar{K}} \cdot r\left(x_n\right).$$

We simulated the policy gradient algorithm on a GARNET$(30,5,10,1)$ MDP. The state features were constructed as described in Section 4.3 with $L = 10$, $l = 2$. Figure 3 shows the average reward η as a function of iteration. These results indicate that the variance reduction in each iteration (guaranteed by Lemma 2) resulted, on average, in a better estimation of the gradient $\nabla_\theta \eta$, and therefore a better policy at each step.

7. Discussion and Future Work

Generally, when devising a solution to a difficult problem, one should incorporate into it all reliably available information. Model free RL algorithms typically operate without explicit knowledge of

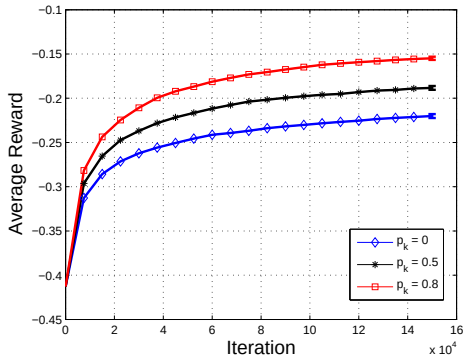

Figure 3: Policy Gradient with a Partial Model. Implementation of the algorithm described in Section 6.2 on a GARNET(30,5,10,1) MDP, with step size parameters $\varepsilon = 0.03$ and $\varepsilon' = 0.003$. The linear FA parameters are $L = 10$ and $l = 2$. Average reward is plotted vs. iteration number for different values of p_K. Results are averaged over 500 different runs with the same initial conditions. Error bars display the standard error of the mean; for clarity of presentation the bars are displayed only for the last iteration.

the underlying environment, and therefore, when such knowledge is available, using these algorithms 'out of the box' is clearly suboptimal. In this work we have presented a general method of integrating partial environmental knowledge into a large class of model free algorithms. Our method improves the asymptotic behavior of the algorithm, and at each iteration reduces the estimation variance due to the uncertainty in the environment. We have proved mathematically (for TD(0)) and demonstrated in simulation (for Policy Gradient and Q-learning) an improvement in the algorithm's overall performance.

From a more conceptual point of view, we have shown that two distinct approaches to RL, the model free and the model based approaches, can be combined in such a way that gains from their respective merits. From this perspective, this work is just a first step towards a theoretical understanding of the combination of different RL approaches.

A few issues are in need of further investigation.

In this work we have not addressed the question of how the partially known model can be acquired. A number of possibilities come to mind. In a transfer learning or tutor learning settings, the partial model can come from an expert who has exact knowledge of a model that is partially similar. In a multi-agent setting with communication, information about different parts of the model can be gathered independently by each of the agents, and combined to create a partial model of the environment.

An interesting possibility is to simultaneously gather information while adapting the policy using some model free algorithm. Using the SA algorithm (8), at the time of the n'th update of θ, we have already encountered a state-action trajectory of size n. Can we use this trajectory to construct an *estimated* partial MDP model, use it as in algorithm (12), and guarantee an improvement in the algorithm's performance? This should be done with caution, since using the same trajectory for updating the parameter and the estimated model may cause overfitting. To see this consider the following example. Let $\{x_i\}$ be a sequence of normally distributed i.i.d. random variables with

mean m, and assume that our goal is to estimate m. A natural approach is to use the empirical mean given by $\theta_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, which can also be calculated recursively using the following SA iterate

$$\begin{aligned} \theta_{n+1} &= \theta_n + \frac{1}{n+1}(x_{n+1} - \theta_n), \\ \theta_0 &= 0. \end{aligned} \tag{49}$$

One may hope, that by the time of the n'th update of θ we could use the $n-1$ values of x_i already observed to build a partial model for x_n, and similarly to (12), use it to manipulate (49) in such a way that guarantees a performance improvement (in the estimation of m). However, it is known that for a normal distribution, the empirical mean is also the minimum variance unbiased estimator for m (Schervish, 1995). Our manipulation of (49) would therefore either add bias or increase the variance. Though this issue deserves careful analysis, we note that when a constant step size is used, the major influences on the current value of the parameter are the most recent measurements, thus older samples can be safely used to construct a partial model, mitigating the severity of this problem.

Finally, we note that the IPM method adds to the algorithm a computational cost of $O(K_{\max})$ evaluations of $F(\theta_n, x_n, u_n, x_{n+1})$ at each iteration. In our experiments, this cost proved to be negligible in comparison to the computational cost of the simulator. However, if the computation of $F(\theta_n, x_n, u_n, x_{n+1})$ is demanding, one may face a tradeoff between the performance of the resulting policy and the computational cost of obtaining it.

Acknowledgments

The authors would like to thank Nahum Shimkin for helpful discussions.

Appendix A. Admissibility of $\hat{\mu}_K$

In this section, based on the definitions of Section 2, we address the following issue. Can a better estimator than $\hat{\mu}_K(x)$ be found?

Since the MSE of any estimator, within a non-Bayesian setting, depends on the unknown μ, comparison of different estimators is a difficult task. A popular comparison framework is that of *admissible* estimators (Schervish, 1995). For a given known set K, an estimator is said to be admissible if there is no other estimator that achieves a smaller MSE for every distribution in $\mathcal{P}_K(\omega)$. Clearly, admissibility is a desirable property for an estimator, since an inadmissible estimator is guaranteed to be sub-optimal. The next theorem states that $\hat{\mu}_K$ is admissible.

Theorem 15 *The estimator $\hat{\mu}_K$ of* (5) *is admissible.*

Proof Let $\tilde{P}(\omega) \in \mathcal{P}_K(\omega)$ be defined as

$$\tilde{P}(\omega) = \frac{\mathbf{1}_\omega^K P(\omega)}{\sum_{\omega' \in K} P(\omega')}.$$

For $X \sim \tilde{P}(\omega)$ it is clear that $\hat{\mu}_K(x) = \mathrm{E}[X]$ for all x, therefore $\mathrm{E}[\hat{\mu}_K(X) - \mu]^2 = 0$, and no other estimator achieves a smaller MSE in this case. ■

Appendix B. Proof of Lemma 5

Proof By the ergodicity of the Markov chain the joint probability for subsequent states is

$$\lim_{n\to\infty} P(x_n, x_{n+1}) \quad = \quad P(x_{n+1}|x_n)[\pi]_{x_n}.$$

Now, observe that

$$\begin{aligned}
&\mathrm{E}\left[(d_n\phi(x_n))(d_n\phi(x_n))^T \middle| \theta_n = \theta^*, x_n\right] \\
&= \mathrm{Cov}\left[d_n\phi(x_n)|\theta_n = \theta^*, x_n\right] + \mathrm{E}\left[(d_n\phi(x_n))|\theta_n = \theta^*, x_n\right]\mathrm{E}\left[(d_n\phi(x_n))|\theta_n = \theta^*, x_n\right]^T \\
&= \gamma^2\phi(x_n)\,\mathrm{Cov}\left[\phi(x_{n+1})^T\theta^* \middle| x_n\right]\phi(x_n)^T \\
&\quad + \mathrm{E}\left[d_n\phi(x_n)|\theta_n = \theta^*, x_n\right]\mathrm{E}\left[d_n\phi(x_n)|\theta_n = \theta^*, x_n\right]^T,
\end{aligned}$$

where the second equality follows from

$$\mathrm{Cov}\left[d_n\phi(x_n)|\theta_n, x_n\right] = \mathrm{Cov}\left[d_n\phi(x_n) - r(x_n) + \phi(x_n)^T\theta_n \middle| \theta_n, x_n\right],$$

since adding constants does not change the covariance. Using Lemma 1 and Lemma 2 we derive an expression for the IPM iteration

$$\begin{aligned}
&\mathrm{E}\left[(d_n^K\phi(x_n))(d_n^K\phi(x_n))^T \middle| \theta_n = \theta^*, x_n\right] \\
&= \gamma^2\phi(x_n)\left(\mathrm{Cov}\left[\phi(x_{n+1})^T\theta^* \middle| x_n\right] - P(K_x)\,\mathrm{Cov}_K\left[\phi(x_{n+1})^T\theta^* \middle| x_n\right]\right)\phi(x_n)^T \\
&\quad + \mathrm{E}\left[d_n^K\phi(x_n)|\theta_n = \theta^*, x_n\right]\mathrm{E}\left[d_n^K\phi(x_n)|\theta_n = \theta^*, x_n\right]^T \\
&= \mathrm{E}\left[(d_n\phi(x_n))(d_n\phi(x_n))^T \middle| \theta_n = \theta^*, x_n\right] - \gamma^2\phi(x_n)P(K_x)\,\mathrm{Cov}_K\left[\left(\phi(x_{n+1})^T\theta^*\right) \middle| x_n\right]\phi(x_n)^T.
\end{aligned}$$

We therefore have that

$$\begin{aligned}
\Sigma_0 &= \lim_{n\to\infty}\mathrm{E}\left[(d_n\phi(x_n))(d_n\phi(x_n))^T \middle| \theta_n = \theta^*\right] \\
&= \lim_{n\to\infty}\mathrm{E}\left[\mathrm{E}\left[(d_n\phi(x_n))(d_n\phi(x_n))^T \middle| \theta_n = \theta^*, x_n\right]\right] \\
&= \sum_x [\pi]_x\,\mathrm{E}\left[(d_n\phi(x_n))(d_n\phi(x_n))^T \middle| \theta_n = \theta^*, x_n\right] \\
&= \Sigma_0^K + \gamma^2\sum_x [\pi]_x\phi(x)P(K_x)\,\mathrm{Cov}_K\left[\phi(x')^T\theta^* \middle| x\right]\phi(x)^T \\
&= \Sigma_0^K + \gamma^2\sum_x [\pi]_x\phi(x)P(K_x)\,\mathrm{Var}_K\left[\phi(x')^T\theta^* \middle| x\right]\phi(x)^T.
\end{aligned}$$

■

Appendix C. Proof of Theorem 9

Proof First, observe that

$$\left|\hat{\mathrm{E}}_{K^\varepsilon}[X] - \mathrm{E}_{K^\varepsilon}[X]\right| = \left|\sum_{x\in K^\varepsilon} x\left(\hat{P}_{K^\varepsilon}(x) - P_{K^\varepsilon}(x)\right)\right| \le \varepsilon\sum_{x\in K^\varepsilon}|x|. \tag{50}$$

Using (50) a simple bound on the bias of $\hat{\mu}_{K^\varepsilon}$ is derived

$$\begin{aligned}\left|\mathrm{E}\left[\hat{\mu}_{K^\varepsilon}(X)\right]-\mathrm{E}[X]\right| &= \left|\sum_{x\in K^\varepsilon} P(x)\left(\hat{\mathrm{E}}_{K^\varepsilon}[X]-x\right)\right| \\ &= P(K^\varepsilon)\left|\hat{\mathrm{E}}_{K^\varepsilon}[X]-\mathrm{E}_{K^\varepsilon}[X]\right| \leq \varepsilon P(K^\varepsilon)\sum_{x\in K^\varepsilon}|x|.\end{aligned}$$

We now derive a bound on the MSE of $\hat{\mu}_{K^\varepsilon}$.

$$\begin{aligned}\mathrm{E}\left[\left(\hat{\mu}_{K^\varepsilon}(X)-\mathrm{E}[X]\right)^2\right] &= \sum_{x\in K^\varepsilon} P(x)\left(\hat{\mathrm{E}}_{K^\varepsilon}[X]-\mathrm{E}[X]\right)^2+\sum_{x\in\bar{K}^\varepsilon} P(x)\left(x-\mathrm{E}[X]\right)^2 \\ &= P(K^\varepsilon)\left(\hat{\mathrm{E}}_{K^\varepsilon}[X]-\mathrm{E}_{K^\varepsilon}[X]+\mathrm{E}_{K^\varepsilon}[X]-\mathrm{E}[X]\right)^2 \\ &\quad +\sum_{x\in\bar{K}^\varepsilon} P(x)\left(x-\mathrm{E}[X]\right)^2 \\ &\leq \mathrm{Var}[X]-P(K^\varepsilon)\cdot\mathrm{Var}_{K^\varepsilon}[X] \\ &\quad +P(K^\varepsilon)\left(\left(\varepsilon\sum_{x\in K^\varepsilon}|x|\right)^2+2\varepsilon\left(\sum_{x\in K^\varepsilon}|x|\right)\left|\mathrm{E}_{K^\varepsilon}[X]-\mathrm{E}[X]\right|\right)\end{aligned}$$

where in the inequality in the third line we used (50) and Lemma 2. We see that we have an improvement in MSE terms if $\left(\varepsilon\sum_{x\in K^\varepsilon}|x|\right)^2+2\varepsilon\left(\sum_{x\in K^\varepsilon}|x|\right)\left|\mathrm{E}_{K^\varepsilon}[X]-\mathrm{E}[X]\right| < \mathrm{Var}_{K^\varepsilon}[X]$, which is always satisfied as $\varepsilon\to 0$. Similarly, for the variance we have

$$\begin{aligned}\mathrm{E}\left[\left(\hat{\mu}_{K^\varepsilon}(X)-\mathrm{E}[\hat{\mu}_{K^\varepsilon}(X)]\right)^2\right] &= \mathrm{E}\left[\left(\hat{\mu}_{K^\varepsilon}(X)-\mathrm{E}[X]\right)^2\right]-\left(\mathrm{E}[\hat{\mu}_{K^\varepsilon}(X)]-\mathrm{E}[X]\right)^2 \\ &\leq \mathrm{Var}[X]-P(K^\varepsilon)\cdot\mathrm{Var}_{K^\varepsilon}[X] \\ &\quad +P(K^\varepsilon)\left(\left(\varepsilon\sum_{x\in K^\varepsilon}|x|\right)^2+2\varepsilon\left(\sum_{x\in K^\varepsilon}|x|\right)\left|\mathrm{E}_{K^\varepsilon}[X]-\mathrm{E}[X]\right|\right).\end{aligned}$$

■

Appendix D. Proof of Theorem 12

Note that without the $\delta\theta$ terms in (42), the bound on the variance for a random variable (28) could be used to compare $[\Sigma_0^K]_{xx}$ with $[\Sigma_0]_{xx}$. For small $\delta\theta$, the difference in the variance should be small, as is proved in the following Lemma, which we first motivate with a simple example.

Let $X\in\{1,2\}$ and $X'\in\{1+\eta,2+\eta\}$ be two random variables satisfying $P(X=1)=P(X'=1+\eta)$ and $P(X=2)=P(X'=2+\eta)$. We expect that for small η, the difference between $\mathrm{Var}[X]$ and $\mathrm{Var}[X']$ should also be small.

Lemma 16 *Let $X\sim P(\cdot)$ be a random variable over a finite set $\{\Omega_i\}_{i=1}^N$, where $\Omega_i\in\mathbb{R}$. Let $X'\sim P(\cdot)$ be a random variable over a finite set $\{\Omega_i'\}_{i=1}^N$, such that $|\Omega_i-\Omega_i'|<\eta$, $i=1,\ldots,N$. The variance of X' satisfies*

$$\mathrm{Var}\left[X'\right]\leq\mathrm{Var}[X]+\eta^2+2\eta\sqrt{\mathrm{Var}[X]}.$$

Proof First we have

$$\mathrm{Var}\left[X'\right] = \mathrm{E}\left[\left(X' - \mathrm{E}\left[X'\right]\right)^2\right] \leq \mathrm{E}\left[\left(X' - \mathrm{E}\left[X\right]\right)^2\right],$$

since

$$\mathrm{E}\left[\left(X' - \mathrm{E}\left[X\right]\right)^2\right] = \mathrm{Var}\left[X'\right] + \left(\mathrm{E}\left[\left(X' - \mathrm{E}\left[X\right]\right)\right]\right)^2.$$

Next we have

$$\mathrm{E}\left[\left(X' - \mathrm{E}\left[X\right]\right)^2\right] = \sum_{x'} P\left(x'\right)\left(x' - \mathrm{E}\left[X\right]\right)^2,$$

but since $|\Omega_i - \Omega_i'| < \eta,\ \forall i$ then by the triangle equality we have $|x' - \mathrm{E}[X]| \leq |x - \mathrm{E}[X]| + \eta$, so we have

$$\begin{aligned}
\mathrm{E}\left[\left(X' - \mathrm{E}\left[X\right]\right)^2\right] &\leq \sum_x P\left(x\right)\left(|x - \mathrm{E}\left[X\right]| + \eta\right)^2 \\
&= \mathrm{Var}\left[X\right] + \eta^2 + 2\eta \sum_x P\left(x\right)|x - \mathrm{E}\left[X\right]| \\
&\leq \mathrm{Var}\left[X\right] + \eta^2 + 2\eta\sqrt{\mathrm{Var}\left[X\right]},
\end{aligned}$$

where in the last inequality we used Cauchy–Schwartz under the expectation inner product:

$$\left\langle |x - \mathrm{E}\left[X\right]|, 1\right\rangle^2 \leq \left\langle |x - \mathrm{E}\left[X\right]|, |x - \mathrm{E}\left[X\right]|\right\rangle \left\langle 1, 1\right\rangle = \mathrm{Var}\left[X\right].$$

■

We now combine the result of Lemma 16 and the bound on the variance developed for the random variable (28) to prove Theorem 12.

Proof (of Theorem 12) First, we use Lemma 16 to bound (42)

$$\begin{aligned}
\mathrm{Var}\left[\mathbf{1}^{\kappa^\varepsilon}\hat{\mathrm{E}}_{\kappa^\varepsilon}\left[\theta^* + \delta\theta\right] + \mathbf{1}^{\bar{\kappa}^\varepsilon}\left(\theta^* + \delta\theta\right)\right] &\leq \mathrm{Var}\left[\mathbf{1}^{\kappa^\varepsilon}\hat{\mathrm{E}}_{\kappa^\varepsilon}\left[\theta^*\right] + \mathbf{1}^{\bar{\kappa}^\varepsilon}\left(\theta^*\right)\right] \\
&\quad + \eta^2 + \frac{2\eta}{\gamma}\sqrt{\mathrm{Var}\left[\theta^*\right]},
\end{aligned}$$

and substitute $[\Sigma_0]_{xx} = \gamma^2 \mathrm{Var}[\theta^*]$. We now use (28) to bound $\mathrm{Var}\left[\mathbf{1}^{\kappa^\varepsilon}\hat{\mathrm{E}}_{\kappa^\varepsilon}\left[\theta^*\right] + \mathbf{1}^{\bar{\kappa}^\varepsilon}\left(\theta^*\right)\right]$, which results in (43). ■

Appendix E. Proof of Theorem 4

As stated before, Theorem 4 is a consequence of Theorem 10.1.3 in Kushner and Yin (2003), for which a long set of assumptions is required. For the sake of clarity, this section is organized as follows. We first present a constrained version of the IPM-TD(0) algorithm and show that it converges weakly to a unique point. We then present some definitions needed for Theorem 10.1.3 in Kushner and Yin (2003), followed by an explicit statement of the theorem, without the required assumptions. Finally, we present the assumptions one by one, and prove that they are indeed fulfilled.

E.1 Constrained Algorithm

An important issue in the analysis of an SA algorithm (8) is the boundedness of the iterates. For many convergence results, a required condition is that the sequence θ_n be bounded almost surely. This condition is not trivial, and there are several approaches to satisfying it. One simple approach is to truncate the iterate θ_n when it leaves some prespecified constraint set denoted by H (Kushner and Yin, 2003). This will be done by introducing a 'correction' term Z_n

$$\theta_{n+1} = \theta_n + \varepsilon F(\theta_n, x_n, u_n, x_{n+1}) + \varepsilon Z_n, \tag{51}$$

where εZ_n is the vector of shortest Euclidean length needed to take $\theta_n + \varepsilon F(\theta_n, x_n, u_n, x_{n+1})$ back to the constraint set H if it is not in H. Respectively, the correction term needs to be added to the associated ODE

$$\frac{d\theta}{dt} = \bar{g}(\theta) + z_t, \tag{52}$$

where z_t maintains θ in H. Recall the unconstrained ODE for TD(0) (15), and its fixed point θ^*. Since in TD(0) θ^* is bounded by the maximal value of the MDP $r_{\max}/(1-\gamma)$, we can choose H to be large enough such that $\theta^* \in H$. The following Lemma guarantees that in this case, the additional z_t term in (52) does not change the ODE's unique fixed point.

Lemma 17 *Assuming $\theta^* \in H$, the constrained ODE for IPM-TD(0) with linear function approximation $d\theta/dt = b + A\theta + z_t$, with b, A, z_t defined in Section 4.1, has a unique and asymptotically stable fixed point θ^*, which satisfies $A\theta^* = -b$.*

Proof Consider as a Lyapunov function the Euclidean distance to θ^*, $\mathcal{V}(\theta) = (\theta - \theta^*)^T (\theta - \theta^*)$. For the unconstrained ODE (10), we have[13] $\dot{\mathcal{V}}(\theta) = \theta^T \left(A + A^T\right) \theta$, and since A is negative definite we have $\dot{\mathcal{V}}(\theta) < 0$, and $\mathcal{V}(\theta)$ is a valid Lyapunov function. Let $\dot{\mathcal{V}}_H(\theta)$ correspond to the constrained ODE. Since θ^* is in H, by the geometric definition of the correction terms, we have $\dot{\mathcal{V}}_H(\theta) \leq \dot{\mathcal{V}}(\theta) < 0$, therefore $\mathcal{V}$ is also valid for the constrained ODE (52). ■

We now state a convergence theorem that relates between the limit point of the ODE (52) and the asymptotic behavior of algorithm (51). The assumptions needed for this theorem are satisfied by default, by the definition of our RL environment and algorithm, and are thus omitted.

Let $\theta(t)$ denote the piece-wise constant continuous time interpolation θ_n, where $\theta(t) = \theta_n$ on the time interval $[n\varepsilon, n\varepsilon + \varepsilon)$ for $t \geq 0$ and $\theta(t) = \theta_0$ for $t < 0$. Similarly define $Z(t)$ as the piece-wise constant continuous time interpolation of Z_n.

Theorem 18 *(Theorem 8.2.2 in Kushner and Yin, 2003) Consider algorithm* (51)*. For any non decreasing sequence of integers q_ε, for each sub-sequence of $\{\theta(\varepsilon q_\varepsilon + \cdot), Z(\varepsilon q_\varepsilon + \cdot)\}, \varepsilon > 0$, there exist a further sub-sequence and a process $(\theta(\cdot), Z(\cdot))$ such that*

$$(\theta(\varepsilon q_\varepsilon + \cdot), Z(\varepsilon q_\varepsilon + \cdot)) \Rightarrow (\theta(\cdot), Z(\cdot))$$

as $\varepsilon \to 0$ through the convergent sub-sequence, where

$$\theta(t) = \theta(0) + \int_0^t \bar{g}(\theta(s))\, ds + Z(t). \tag{53}$$

Let $\varepsilon q_\varepsilon \to \infty$ as $\varepsilon \to 0$. Then, for almost all ω, the path $\theta(\omega, \cdot)$ lies in a limit set of (53)*.*

13. See derivation in the proof of **20.3.3.**

E.2 Definitions

The following technical definitions are required for the convergence result.

Let $D^L(-\infty,\infty)$ (and $D^L[0,\infty)$, respectively) denote the $L-$fold product space of real valued functions on the interval $(-\infty,\infty)$ (resp. on $[0,\infty)$) that are right continuous and have left-hand limits, with the Skorohod topology used.[14]

Let $\{q_\varepsilon\}$ be a sequence of non-negative integers. In order to investigate the asymptotic behavior we will examine $\theta(\varepsilon q_\varepsilon + \cdot)$, where $\varepsilon q_\varepsilon \to \infty$. We also demand $\varepsilon(q_\varepsilon - p_\varepsilon) \to \infty$ where p_ε are non decreasing and non-negative integers used in assumption 20.3.

Define the normalized error process

$$U_n = (\theta_{q_\varepsilon+n} - \theta^*)/\sqrt{\varepsilon},$$

and let $U^\varepsilon(\cdot)$ denote its piecewise constant right continuous interpolation, with interpolation intervals ε, on $[0,\infty)$. Define $W^\varepsilon(\cdot)$ on $(-\infty,\infty)$ by

$$W^\varepsilon(t) = \begin{cases} \sqrt{\varepsilon}\sum_{i=q_\varepsilon}^{q_\varepsilon+t/\varepsilon-1} F(\theta^*, x_i, u_i, x_{i+1}), & t \geq 0 \\ -\sqrt{\varepsilon}\sum_{i=q_\varepsilon}^{q_\varepsilon+t/\varepsilon-1} F(\theta^*, x_i, u_i, x_{i+1}), & t < 0 \end{cases} \tag{54}$$

E.3 A Theorem on Fluctuations in SA

Theorem 19 *(10.1.3 in Kushner and Yin, 2003) Consider algorithm* (51) *and let assumption 20 hold. Then the sequence* $\{U^\varepsilon(\cdot), W^\varepsilon(\cdot)\}$ *converges weakly in* $D^L[0,\infty) \times D^L(-\infty,\infty)$ *to a limit denoted by* $\{U(\cdot), W(\cdot)\}$*, and*

$$dU = AU\,dt + dW,$$

where the matrix A is defined in 20.8, $W(\cdot)$ *is a Wiener process with covariance matrix* Σ *described in 20.5, and* $U(\cdot)$ *is stationary.*

Theorem 4 is a direct consequence of Theorem 19, with $n = \omega(1/\varepsilon)$ satisfying the requirement on q_ε.

E.4 Assumptions for Theorem 19

The set of assumptions 20 which we describe in the following is designed to fit a wide variety of algorithms, and are thus quite complicated. The IPM-TD(0) algorithm with which we are concerned is a very simple case of this theorem, as it is linear, bounded, and stationary, and the Markovian state transitions are ergodic, and defined over a finite state space.[15] Moreover, many of the assumptions that follow are used in order to reduce a more complicated algorithm to these simpler settings, and to show that the residual that remains is small in some sense. Thus, many complicated terms in the assumptions just vanish, and some assumptions are true by default.

14. See Kushner and Yin (2003, p. 228, 238) for more details on D^L.
15. In Borkar (2008) a simpler result regarding fluctuations with a fixed step size is given, albeit for a martingale difference noise scenario. The Markovian state dependent noise in our case requires the more complicated approach of Kushner and Yin (2003).

Assumption 20 *The following holds*[16]

1. $\left\{F\left(\theta_n, x_n, u_n, x_{n+1}\right) I_{\{|\theta_n - \theta^*| \leq \rho\}}\right\}$ is uniformly integrable for small $\rho > 0$.

   Proof $F\left(\theta_n, x_n, u_n, x_{n+1}\right)$ is uniformly integrable since on every sample path θ_n is bounded (by the constraint), $r\left(x_n\right)$ is bounded by r_{max} and $\phi\left(x_n\right)$ is also bounded by definition. Since this is true for every sample path, $F\left(\theta_n, x_n, u_n, x_{n+1}\right) I_{\{|\theta_n - \theta^*| \leq \rho\}}$ is uniformly integrable for all ρ. ■

2. There is a sequence of non-negative and non decreasing integers N_ε such that $\theta\left(\varepsilon N_\varepsilon + \cdot\right)$ converges weakly to the process with constant value θ^* strictly inside the constraint set.

   Proof By the weak convergence Theorem 18, choosing N_ε such that $\varepsilon N_\varepsilon \to \infty$, and by Lemma 17, we have that IPM-TD(0) converges weakly to the process with constant value θ^* strictly inside the constraint set. ■

3. There are non decreasing and non-negative integers p_ε (that can be taken to be greater than N_ε) such that

   $$\left\{\left(\theta_{p_\varepsilon + n} - \theta^*\right) / \sqrt{\varepsilon}; \varepsilon > 0, n \geq 0\right\}$$

   is tight.

   Proof For the proof of this assumption we use Theorem 10.5.2 in Kushner and Yin (2003), which we now state.

   Theorem 21 *(10.5.2 in Kushner and Yin, 2003) Assume the constrained algorithm 51 with constraint set H, where θ^* is in the interior of H. Assume that 20.2 and 20.3.1-20.3.7 hold in H. Then there are $p_\varepsilon < \infty$ such that $\left\{\left(\theta_{p_\varepsilon + n} - \theta^*\right) / \sqrt{\varepsilon}; \varepsilon > 0, n \geq 0\right\}$ is tight.*

   ■

3.1 θ^* is a globally asymptotically stable (in the sense of Lyapunov) point of the ODE $d\theta/dt = \bar{g}\left(\theta\right) + z_t$.

   Proof This is satisfied by Lemma 17. ■

3.2 The non-negative and continuously differentiable function $V\left(\cdot\right)$ is a Lyapunov function for the ODE. The second order partial derivatives are bounded and $\left|\nabla_\theta V\left(\theta\right)\right|^2 \leq K_1\left(V\left(\theta\right) + 1\right)$, where K_1 is an arbitrary positive number.

   Proof Choose V to be of the form $V\left(\theta\right) = \left(\theta - \theta^*\right)^T \left(\theta - \theta^*\right)$. As was in the proof of Lemma 17, V is a valid Lyapunov function. The second order partial derivatives are zero, and

   $$\nabla_\theta V\left(\theta\right) = 2\left(\theta - \theta^*\right)$$

16. We exclude assumptions which are true by definition of our RL settings.

$$|\nabla_\theta V(\theta)|^2 = 4|\theta-\theta^*|^2 = 4V(\theta).$$

■

3.3 There is a $\lambda > 0$ such that $V_\theta^T(\theta)\bar{g}(\theta) \le -\lambda V(\theta)$

Proof Recalling from (15) and (16) that $\bar{g}(\theta) = b + A\theta$, we have

$$\begin{aligned}
\frac{1}{2}\left((\nabla_\theta V(\theta))^T \bar{g}(\theta) + \left((\nabla_\theta V(\theta))^T \bar{g}(\theta)\right)^T\right) &= (\theta-\theta^*)^T(b+A\theta) + (b+A\theta)^T(\theta-\theta^*) \\
&= (\theta-\theta^*)^T b + b^T(\theta-\theta^*) \\
&\quad + (\theta-\theta^*)^T A\theta + \theta^T A^T(\theta-\theta^*) \\
&= (\theta-\theta^*)^T b + b^T(\theta-\theta^*) \\
&\quad + (\theta-\theta^*)^T\left(A+A^T\right)(\theta-\theta^*) \\
&\quad + (\theta-\theta^*)^T A\theta^* + \theta^{*T} A^T(\theta-\theta^*) \\
&= (\theta-\theta^*)^T(b+A\theta^*) + (b+A\theta^*)^T(\theta-\theta^*) \\
&\quad + (\theta-\theta^*)^T\left(A+A^T\right)(\theta-\theta^*) \\
&= (\theta-\theta^*)^T\left(A+A^T\right)(\theta-\theta^*).
\end{aligned}$$

Let λ' denote the largest eigenvalue of $A+A^T$. We have that $(\theta-\theta^*)^T\left(A+A^T\right)(\theta-\theta^*) \le \lambda'(\theta-\theta^*)^T(\theta-\theta^*)$, and

$$(\nabla_\theta V(\theta))^T \bar{g}(\theta) = (\theta-\theta^*)^T\left(A+A^T\right)(\theta-\theta^*) \le \lambda' V(\theta).$$

■

3.4 For each $K > 0$, $\sup_n \mathrm{E}|F(\theta_n, x_n, u_n, x_{n+1})|^2 I_{\{|\theta_n-\theta^*|\le K\}} \le K_1 \mathrm{E}[V(\theta_n)+1]$, where K_1 does not depend on K.

Proof

Satisfying this requirement is immediate, since $F(\theta_n, x_n, u_n, x_{n+1})$ is bounded on every sample path. This follows from the fact that on every sample path θ_n is bounded (by the constraint), $r(x_n)$ is bounded by r_{max} and $\phi(x_n)$ is also bounded by definition. ■

3.5 The sum $\Gamma_n(\theta) = \varepsilon\sum_{i=n}^{\infty}(1-\varepsilon)^{i-n}\mathrm{E}_n[g(\theta, x_i, u_i) - \bar{g}(\theta)]$, where E_n denotes expectation conditioned on the history up to time n, is well defined in that the sum of the norms of the summands is integrable for each θ, and $\mathrm{E}|\Gamma_n(\theta_n)|^2 = O(\varepsilon^2)$.

Proof From Lemma 6.7 in Bertsekas and Tsitsiklis (1996) (which relies on the exponential mixing time of Markov chains) we have that $|\mathrm{E}_n[g(\theta, x_i, u_i) - \bar{g}(\theta)]| \le c\rho^{i-n}|\theta|$ for some

$c > 0$ and $\rho < 1$. This gives

$$\begin{aligned}\left|\sum_{i=n}^{\infty}(1-\varepsilon)^{i-n}\mathrm{E}_n\left[g\left(\theta,x_i,u_i\right)-\bar{g}\left(\theta\right)\right]\right| &\leq \sum_{i=n}^{\infty}\left|\mathrm{E}_n\left[g\left(\theta,x_i,u_i\right)-\bar{g}\left(\theta\right)\right]\right| \\ &\leq \sum_{i=n}^{\infty}c\rho^{i-n}\left|\theta\right| \\ &= \frac{c\left|\theta\right|}{1-\rho},\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}\left|\Gamma_n\left(\theta_n\right)\right|^2 &\leq \varepsilon^2\mathrm{E}\left|\frac{c\left|\theta_n\right|}{1-\rho}\right|^2 \\ &= \frac{\varepsilon^2 c}{1-\rho}\left|\theta^*\right|^2 \\ &= O\left(\varepsilon^2\right).\end{aligned}$$

■

3.6 $\mathrm{E}\left|\Gamma_{n+1}\left(\theta_{n+1}\right)-\Gamma_{n+1}\left(\theta_n\right)\right|^2 = O\left(\varepsilon^2\right)$.

Proof

We have

$$\begin{aligned}&\Gamma_{n+1}\left(\theta_{n+1}\right)-\Gamma_{n+1}\left(\theta_n\right) \\ =\ &\varepsilon\sum_{i=n+1}^{\infty}(1-\varepsilon)^{i-n-1}\mathrm{E}_{n+1}\left[g\left(\theta_{n+1},x_i,u_i\right)-\bar{g}\left(\theta_{n+1}\right)\right] \\ &-\varepsilon\sum_{i=n+1}^{\infty}(1-\varepsilon)^{i-n-1}\mathrm{E}_{n+1}\left[g\left(\theta_n,x_i,u_i\right)-\bar{g}\left(\theta_n\right)\right] \\ =\ &\varepsilon\sum_{i=n+1}^{\infty}(1-\varepsilon)^{i-n-1}\mathrm{E}_{n+1}\left[g\left(\theta_{n+1},x_i,u_i\right)-g\left(\theta_n,x_i,u_i\right)-\left(\bar{g}\left(\theta_{n+1}\right)-\bar{g}\left(\theta_n\right)\right)\right].\end{aligned}$$

Using the triangle inequality

$$\begin{aligned}&\left|\Gamma_{n+1}(\theta_{n+1})-\Gamma_{n+1}(\theta_n)\right| \leq \\ &\varepsilon\sum_{i=n+1}^{\infty}(1-\varepsilon)^{i-n-1}\left|\mathrm{E}_{n+1}[g\left(\theta_{n+1},x_i,u_i\right)-g\left(\theta_n,x_i,u_i\right)-\left(\bar{g}\left(\theta_{n+1}\right)-\bar{g}\left(\theta_n\right)\right)]\right|\end{aligned}$$

By the linearity of $g,\bar{g}$, and since θ_n is bounded, we have that$\left|g\left(\theta_{n+1},x_i,u_i\right)-g\left(\theta_n,x_i,u_i\right)\right| < k\varepsilon$ for some k, and $\left|\bar{g}\left(\theta_{n+1}\right)-\bar{g}\left(\theta_n\right)\right| < \bar{k}\varepsilon$ for some $\bar{k}$. We therefore have :

$$\left|\mathrm{E}_{n+1}\left[g\left(\theta_{n+1},x_i,u_i\right)-g\left(\theta_n,x_i,u_i\right)\right]\right| \leq \mathrm{E}_{n+1}\left[\left|g\left(\theta_{n+1},x_i,u_i\right)-g\left(\theta_n,x_i,u_i\right)\right|\right] < k\varepsilon,$$

and similarly

$$\left|\mathrm{E}_{n+1}\left[\bar{g}\left(\theta_{n+1}\right)-\bar{g}\left(\theta_{n}\right)\right]\right|<\bar{k}\varepsilon.$$

Now

$$\begin{aligned}\left|\Gamma_{n+1}\left(\theta_{n+1}\right)-\Gamma_{n+1}\left(\theta_{n}\right)\right| &\leq \varepsilon^{2}\left(k+\bar{k}\right)\sum_{i=n+1}^{\infty}\left(1-\varepsilon\right)^{i-n-1}\\ &= \varepsilon\left(k+\bar{k}\right),\end{aligned}$$

and

$$\begin{aligned}\left|\Gamma_{n+1}\left(\theta_{n+1}\right)-\Gamma_{n+1}\left(\theta_{n}\right)\right|^{2} &\leq \varepsilon^{2}\left|k+\bar{k}\right|^{2}\\ &= O\left(\varepsilon^{2}\right).\end{aligned}$$

■

3.7 Let θ^{H} denote the projection of θ onto H. Then for all θ, $V\left(\theta^{H}\right)\leq V\left(\theta\right)$.

Proof This assumption was shown to hold in the proof of Lemma 17. ■

4. For a small $\rho>0$, and any sequence $\varepsilon\to\infty$ and $n\to\infty$ such that $\theta_{n}\to\theta^{*}$ in probability,

$$\mathrm{E}\left[\left|\delta M_{n}-\delta M_{n}\left(\theta^{*}\right)\right|^{2}I_{\{|\theta_{n}-\theta^{*}|\leq\rho\}}\right]\to 0.$$

Proof Recall that we have

$$\begin{aligned}\delta M_{n} &= \mathrm{E}\left[d_{n}^{K}\phi\left(x_{n}\right)|x_{n},u_{n}\right]-d_{n}^{K}\phi\left(x_{n}\right)\\ &= \gamma\sum_{y}P_{u_{n}}\left(y|x_{n}\right)\phi(y)^{T}\theta_{n}\phi\left(x_{n}\right)\\ &\quad -\gamma\left(\mathbf{1}_{n+1}^{K}\frac{\sum_{y\in K_{x_{n},u_{n}}}P_{u_{n}}\left(y|x_{n}\right)\phi(y)^{T}}{\sum_{y\in K_{x_{n},u_{n}}}P_{u_{n}}\left(y|x_{n}\right)}+\mathbf{1}_{n+1}^{\bar{K}}\phi(x_{n+1})^{T}\right)\theta_{n}\phi\left(x_{n}\right)\\ &= \gamma\left(\sum_{y}P_{u_{n}}\left(y|x_{n}\right)\phi(y)^{T}-\mathbf{1}_{n+1}^{K}\frac{\sum_{y\in K_{x_{n},u_{n}}}P_{u_{n}}\left(y|x_{n}\right)\phi(y)^{T}}{\sum_{y\in K_{x_{n},u_{n}}}P_{u_{n}}\left(y|x_{n}\right)}-\mathbf{1}_{n+1}^{\bar{K}}\phi(x_{n+1})^{T}\right)\theta_{n}\phi\left(x_{n}\right).\end{aligned}$$

The difference $\delta M_{n}-\delta M_{n}\left(\theta^{*}\right)$ can therefore be written as

$$\delta M_{n}-\delta M_{n}\left(\theta^{*}\right)=a\left(x_{n}\right)^{T}\left(\theta_{n}-\theta^{*}\right)b\left(x_{n}\right),$$

where a and b are vector valued functions of x_{n}. By the Cauchy–Schwarz inequality, for every x_{n}

$$\left|\delta M_{n}-\delta M_{n}\left(\theta^{*}\right)\right|\leq\left|a\left(x_{n}\right)\right|\left|\theta_{n}-\theta^{*}\right|\left|b\left(x_{n}\right)\right|,$$

and since the state space is finite, a and b are bounded, therefore there exists some constant k such that for every x_n

$$|a(x_n)||b(x_n)| \leq k,$$

and we have that

$$\mathrm{E}\left[|\delta M_n - \delta M_n(\theta^*)|^2 I_{\{|\theta_n-\theta^*|\leq\rho\}}\right] \leq k^2|\theta_n-\theta^*|^2 \to 0.$$

■

5. The sequence of processes $W(\cdot)$ defined on $(-\infty,\infty)$ by (54) converges weakly in $D^L(-\infty,\infty)$ to a Wiener process $W(\cdot)$, with covariance matrix Σ.

Proof

For the proof of this assumption we use Theorem 10.6.2 in Kushner and Yin (2003),which we now state.

Theorem 22 *(10.6.2 in Kushner and Yin, 2003) Assume 20.5.1-20.5.4. Then* $\{W(\cdot)\}$ *defined in* (54) *converges weakly to a Wiener process with covariance matrix* $\Sigma = \Sigma_0 + \Sigma_1 + \Sigma_1^T$.

■

5.1 The following equations hold:

$$\begin{aligned}\lim_{N\to\infty}\sup_n \mathrm{E}\left|\sum_{j=n+N}^{\infty}\mathrm{E}\left(F(\theta^*,x_j,u_j,x_{j+1})\middle|x_n,u_n\right)\right| &= 0,\\ \lim_{N\to\infty}\sup_n \mathrm{E}\left|\sum_{i=n+N}^{\infty}\mathrm{E}\left(F(\theta^*,x_n,u_n,x_{n+1})F(\theta^*,x_i,u_i,x_{i+1})|x_n,u_n\right)^T\right| &= 0.\end{aligned}$$

Proof

Since the transition probabilities at step j converge to the steady state transition probabilities (a property of ergodic Markov chains) exponentially fast in j, and since at the steady state $\mathrm{E}(F(\theta^*,x,u,x')) \equiv \bar{g}(\theta^*) = 0$, we have that for some $\rho < 1$ and some vector c

$$\left|\mathrm{E}\left(F(\theta^*,x_j,u_j,x_{j+1})\middle|x_n,u_n\right)\right| < c\rho^{j-n},$$

therefore for every n

$$\begin{aligned}\lim_{N\to\infty}\left|\sum_{j=n+N}^{\infty}\mathrm{E}\left(F(\theta^*,x_j,u_j,x_{j+1})\middle|x_n,u_n\right)\right| &\leq \lim_{N\to\infty} c\sum_{j=N}^{\infty}\rho^j\\ &= \lim_{N\to\infty}\frac{c\rho^N}{1-\rho}\\ &= 0.\end{aligned}$$

The same goes for the covariance, since there exists some $\rho' < 1$ and some matrix c' such that

$$\mathrm{E}\left(F\left(\theta^*, x_n, u_n, x_{n+1}\right) F\left(\theta^*, x_i, u_i, x_{i+1}\right) \middle| x_n, u_n\right)^T < c' \rho'^{i-n}.$$

■

5.2 The sets $\left\{\left|F\left(\theta^*, x_n, u_n, x_{n+1}\right)\right|^2\right\}$ and $\left\{\left|\sum_{j=i}^{\infty} \mathrm{E}\left(F\left(\theta^*, x_j, u_j, x_{j+1}\right) \middle| x_i, u_i\right)\right|^2\right\}$ are uniformly integrable.

Proof As was shown before, $F\left(\theta^*, x_n, u_n, x_{n+1}\right)$ is bounded, and therefore $\left\{\left|F\left(\theta^*, x_n, u_n, x_{n+1}\right)\right|^2\right\}$ is uniformly integrable. Also, as was shown in the proof of A5.1, for every i

$$\left|\sum_{j=i}^{\infty} \mathrm{E}\left(F\left(\theta^*, x_j, u_j, x_{j+1}\right) \middle| x_i, u_i\right)\right| \leq \frac{c}{1-\rho},$$

which is bounded, and therefore $\left\{\left|\sum_{j=i}^{\infty} \mathrm{E}\left(F\left(\theta^*, x_j, u_j, x_{j+1}\right) \middle| x_i, u_i\right)\right|^2\right\}$ is uniformly integrable.

■

5.3 There is a matrix Σ_0 such that

$$\frac{1}{m} \sum_{j=n}^{n+m-1} \mathrm{E}\left[F\left(\theta^*, x_j, u_j, x_{j+1}\right) F\left(\theta^*, x_j, u_j, x_{j+1}\right)^T \middle| x_n, u_n\right] - \Sigma_0 \to 0$$

in probability as $n, m \to \infty$.

Proof

Since the Markov chain is ergodic, by the law of large numbers this is satisfied by defining

$$\Sigma_0 = \lim_{n \to \infty} \mathrm{E}\left[F\left(\theta^*, x_n, u_n, x_{n+1}\right) F\left(\theta^*, x_n, u_n, x_{n+1}\right)^T\right].$$

■

5.4 There is a matrix Σ_1 such that

$$\frac{1}{m} \sum_{j=n}^{n+m-1} \sum_{k=j+1}^{\infty} \mathrm{E}\left[F\left(\theta^*, x_j, u_j, x_{j+1}\right) F\left(\theta^*, x_k, u_k, x_{k+1}\right)^T \middle| x_n, u_n\right] - \Sigma_1 \to 0$$

in probability as $n, m \to \infty$.

Proof

Since the Markov chain is ergodic, by the law of large numbers this is satisfied by defining

$$\Sigma_1 = \sum_{j=1}^{\infty} \lim_{n \to \infty} \mathrm{E}\left[F\left(\theta^*, x_n, u_n, x_{n+1}\right) F\left(\theta^*, x_{n+j}, u_{n+j}, x_{n+j+1}\right)^T\right].$$

■

6. $g(\cdot,x,u)$ is continuously differentiable for each x,u, and can be expanded as

$$g(\theta,x,u) = g(\theta^*,x,u) + [\nabla_\theta g(\theta^*,x,u)]^T(\theta-\theta^*) + [y(\theta,x,u)]^T(\theta-\theta^*),$$

where

$$[y(\theta,x,u)]^T(\theta-\theta^*) = \int_0^1 [\nabla_\theta g(\theta^*+s(\theta-\theta^*),x,u) - \nabla_\theta g(\theta^*,x,u)]\,ds, \tag{55}$$

and if $\delta \to 0$ as $\varepsilon \to 0$ and $n \to \infty$, then

$$\mathrm{E}\,|y(\theta_n,x_n,u_n)|\,I_{\{|\theta_n-\theta^*|\le\delta\}} \to 0$$

as $\varepsilon \to 0$ and $n \to \infty$.

Proof Recall that for IPM-TD(0)

$$g(\theta,x,u) = \left(r(x) + \left(\gamma\sum_y P_u(y|x)\,\phi(y)^T - \phi(x)^T\right)\theta\right)\phi(x),$$

which is linear in θ and thus can be expanded as

$$\begin{aligned} g(\theta,x,u) &= \left(r(x) + \left(\gamma\sum_y P_u(y|x)\,\phi(y)^T - \phi(x)^T\right)(\theta-\theta^*+\theta^*)\right)\phi(x) \\ &= g(\theta^*,x,u) + \left(\gamma\sum_y P_u(y|x)\,\phi(y)^T - \phi(x)^T\right)(\theta-\theta^*)\,\phi(x) \\ &= g(\theta^*,x,u) + [\nabla_\theta g(\theta^*,x,u)]^T(\theta-\theta^*). \end{aligned}$$

Since $\nabla_\theta g(\theta,x,u)$ does not depend on θ, the integral in (55) is zero and $y(\theta,x,u)=0$, thus the assumption is satisfied. ■

7. The set $\{\nabla_\theta g(\theta^*,x_n,u_n)\}$ is uniformly integrable.

Proof As was shown above, $\nabla_\theta g(\theta^*,x_n,u_n)$ is clearly bounded, and therefore uniformly integrable. ■

8. There is a Hurwitz matrix A such that

$$\frac{1}{m}\sum_{j=n}^{n+m+1}\left[\mathrm{E}\left[\nabla_\theta g^T(\theta^*,x_j,u_j)\,\middle|\,x_n,u_n\right] - A\right] \to 0$$

in probability as $\varepsilon \to 0$ and $n,m \to \infty$.

Proof We have,

$$\nabla_\theta g^T(\theta^*, x, u) = \phi(x)\left(\gamma \sum_y P_u(y|x)\,\phi(y) - \phi(x)\right)^T.$$

Recall our definition of A

$$A \triangleq \Phi^T \Pi_\mu (\gamma P_\mu - I)\,\Phi.$$

Then, by the law of large numbers, we have

$$\frac{1}{m}\sum_{j=n}^{n+m+1}\left[\mathrm{E}\left[\nabla_\theta g^T(\theta^*, x_i, u_i)\middle| x_n, u_n\right] - A\right] \to 0.$$

As was stated before, it can be shown (Bertsekas and Tsitsiklis, 1996, Lemma 6.6b) that the eigenvalues of A all have a negative real part, therefore A is Hurwitz. ■

References

P. Abbeel, M. Quigley, and A.Y. Ng. Using inaccurate models in reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 1–8. ACM, 2006.

J. Abounadi, D. Bertsekas, and V.S. Borkar. Learning algorithms for markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698, 2001.

A.G. Barto, S.J. Bradtke, and S.P. Singh. Learning to act using real-time dynamic programming. *Artificial Intelligence*, 72(1-2):81–138, 1995.

D.P. Bertsekas. *Dynamic Programming and Optimal Control, Vol I & II*. Athena Scientific, third edition, 2006.

D.P. Bertsekas and J. Tsitsiklis. *Neuro-dynamic Programming*. Athena Scientific, 1996.

S. Bhatnagar, R. Sutton, M. Ghavamzadeh, and M. Lee. Natural actor–critic algorithms. Technical Report TR09-10, Univ. of Alberta, 2007.

V.S. Borkar. *Stochastic Approximation: a Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

R.I. Brafman and M. Tennenholtz. R-max a general polynomial time algorithm for near-optimal reinforcement learning. *The Journal of Machine Learning Research*, 3:213–231, 2003.

R. Crites and A. Barto. Improving elevator performance using reinforcement learning. In *Advances in Neural Information Processing Systems 8*, pages 1017–1023. MIT Press, 1996.

N.D. Daw, Y. Niv, and P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711, 2005.

P. Dayan and T.J. Sejnowski. TD(λ) converges with probability 1. *Machine Learning*, 14:295–301, 1994.

R.G. Gallager. *Discrete Stochastic Processes*. Kluwer Academic Publishers, 1995.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2):209–232, 2002.

V. Konda. *Actor-Critic Algorithms*. PhD thesis, Massachusetts Institute of Technology, 2002.

P.R. Kumar. A survey of some results in stochastic adaptive control. *SIAM Journal on Control and Optimization*, 23:329–380, 1985.

H.J. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer Verlag, 2003.

L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22 (4):551 – 575, 1977.

P. Marbach and J. Tsitsiklis. Simulation-based optimization of markov reward processes. *IEEE Transactions on Automatic Control*, 46(2):191–209, 1998.

P. Marbach, O. Mihatsch, M. Schulte, and J.N. Tsitsiklis. Reinforcement learning for call admission control and routing in integrated service networks. In *Advances in Neural Information Processing Systems 10*, pages 922–928. MIT Press, 1998.

A. Papoulis and S.U. Pillai. *Probability, Random Variables, and Stochastic Processes*. McGraw Hill, fourth edition, 2002.

M.J. Schervish. *Theory of Statistics*. Springer, 1995.

S. Singh and P. Dayan. Analytical mean squared error curves for temporal difference learning. *Machine Learning*, 32:5–40, 1998.

R.S. Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*, pages 216–224. Morgan Kaufmann, 1990.

R.S. Sutton and A.G. Barto. *Reinforcement Learning*. MIT Press, 1998.

G. Tesauro. Temporal difference learning and td-gammon. *Commun. ACM*, 38, March 1995.