LIDS-P-2322
March 6, 1996

An Analysis of Temporal-Difference Learning with Function Approximation[1]

John N. Tsitsiklis and Benjamin Van Roy

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139
e-mail: jnt@mit.edu, bvr@mit.edu

[1] This research was supported by the NSF under grant ECS 9216531, EPRI under contract 8030-10, and ARO under grant DAAL-03-92-G-0115.

ABSTRACT

We discuss the temporal-difference learning algorithm, as applied to approximating the cost-to-go function of an infinite-horizon discounted Markov chain, using a function approximator involving linear combinations of fixed basis functions. The algorithm we analyze performs on-line updating of a parameter vector during a single endless trajectory of an ergodic Markov chain with a finite or infinite state space. We present a proof of convergence (with probability 1), a characterization of the limit of convergence, and a bound on the resulting approximation error. In addition to proving new and stronger results than those previously available, our analysis is based on a new line of reasoning that provides new intuition about the dynamics of temporal-difference learning. Finally, we prove that on-line updates, based on entire trajectories of the Markov chain, are in a certain sense necessary for convergence. This fact reconciles positive and negative results that have been discussed in the literature, regarding the soundness of temporal-difference learning.

1 Introduction

The problem of predicting the expected long-term future cost (or reward) of a stochastic dynamic system manifests itself in both time-series prediction and control. An example in time-series prediction is that of estimating the net present value of a corporation, as a discounted sum of its future cash flows, based on the current state of its operations. In control, the ability to predict long-term future cost as a function of state enables the ranking of alternative states in order to guide decision-making. Indeed, such predictions constitute the *cost-to-go function* that is central to dynamic programming and optimal control (Bertsekas, 1995).

Temporal-difference learning, originally proposed by Sutton (1988), is a method for approximating long-term future cost as a function of current state. The algorithm is recursive, efficient, and simple to implement. Linear combinations of fixed basis functions are used to approximate the mapping from state to future cost.[2] The weights of the linear combination are updated upon each observation of a state transition and the associated cost. The objective is to improve approximations of long-term future cost as more and more state transitions are observed. The trajectory of states and costs can be generated either by a physical system or a simulated model. In either case, we view the system as a Markov chain. Adopting terminology from dynamic programming, we will refer to the function mapping states of the Markov chain to expected long-term cost as the cost-to-go function.

Though temporal-difference learning is simple and elegant, a rigorous analysis of its behavior requires significant sophistication. Several previous papers have presented positive results about the algorithm. These include (Sutton, 1988), (Watkins and Dayan, 1992), (Tsitsiklis, 1994), (Jaakola et al., 1994), (Dayan and Sejnowski, 1994), and (Gurvits et al., 1995), all of which only deal with cases where the number of tunable parameters is the same as the cardinality of the state space. Such cases are not practical when state spaces are large or infinite. The more general case, involving the use of function approximation, is addressed by results in (Dayan, 1992), (Tsitsiklis and Van Roy, 1994), and (Gordon, 1995). Tsitsiklis and Van Roy (1994) and Gordon (1995) establish convergence with probability 1. However, their results only apply to a very limited class of function approximators and involve variants of a constrained version of temporal difference learning, known as TD(0). Dayan (1992) establishes convergence in the mean for the general class of function approximators involving linear combinations of fixed basis functions. However, this form of convergence is rather weak, and the analysis used in the paper does not directly lead to approximation error bounds or interpretable characterizations of the limit of convergence.

In addition to the positive results, counter-examples to variants of the algorithm have been offered in several papers. These include (Boyan and Moore, 1995), (Tsitsiklis and Van Roy, 1994), and (Gordon, 1995). The key feature that distinguishes these negative results from their positive counterparts is that the variants of temporal-difference learning used do not employ on-line state sampling. In particular, sampling is done by a mechanism that samples states with frequencies independent from the dynamics of the underlying system. Our results shed light on these counter-examples by showing that, for the general class of linearly-parameterized function approximators, convergence is guaranteed if and only if states are sampled according to the steady-state probabilities of the Markov chain of

[2] Actually, nonlinearly parameterized functions such as neural networks can also be used, though we do not address this case in the paper.

interest. Given that the steady-state probabilities are usually unknown, the only viable approach to generating the required samples is to perform on-line sampling. By this we mean that the samples should consist of an actual sequence of visited states obtained either through simulation of a Markov chain or observation of a physical system.

In this paper, we focus on the application of temporal-difference learning to infinite-horizon discounted Markov chains with finite or infinite state spaces. Though finite state absorbing Markov chains have been the dominant setting for past analyses, we find the infinite-horizon framework to be the most natural and elegant setting for temporal difference learning. Furthermore, the ideas used in our analysis can easily be applied to prove similar results in the context of absorbing Markov chains. Though this extension is omitted from this paper, it can be found in (Bertsekas and Tsitsiklis, 1996).

The contributions in this paper are as follows:

1. Convergence (with probability 1) is established for the case where approximations are generated by linear combinations of (possibly unbounded) basis functions over a (possibly infinite) state space. This is the first such result that handles the case of "compact representations" of the cost-to-go function, in which there are fewer parameters than states. (In fact, convergence of on-line algorithms in the absence of an absorbing state, had not been established even for the case of a lookup table representation.)

2. The limit of convergence is characterized as the solution to a set of interpretable linear equations, and a bound is placed on the resulting approximation error.

3. We reconcile positive and negative results concerning temporal-difference learning by proving a theorem that identifies the importance of on-line sampling.

4. Our methodology leads to an interpretation of the limit of convergence and hence new intuition on temporal-difference learning and the dynamics of weight updating.

This paper is organized as follows. In Section 2, we provide a precise definition of the algorithm that we will be studying. In Section 3, we recast temporal-difference learning in a way that sheds light into its mathematical structure. Section 4 contains our main convergence result together with our assumptions. We develop some mathematical machinery in Section 5, which captures the fundamental ideas involved in the analysis. Section 6 presents a proof of the convergence result, which consists primarily of the technicalities required to integrate the machinery supplied by section 5. Our analysis is valid for general state spaces, subject to certain technical assumptions. In Section 7, we show that these technical assumptions are automatically valid whenever the state space is finite. In Section 8, we argue that the class of infinite state Markov chains that satisfy our assumptions is broad enough to be of practical interest. Section 9 contains our converse convergence result, which establishes the importance of on-line sampling. Finally, Section 10 contains some concluding remarks.

2 Definition of Temporal-Difference Learning

In this section, we define precisely the nature of temporal-difference learning, as applied to approximation of the cost-to-go function for an infinite-horizon discounted Markov chain. While the method as well as our subsequent results are applicable to Markov chains with a fairly general state space, we restrict our attention to the case where the state space is countable. This allows us to work with relatively simple notation; for example, the Markov chain can be defined in terms of an (infinite) transition probability matrix as opposed to a transition probability kernel. The extension to the case of general state spaces requires the translation of the matrix notation into operator notation, but is otherwise straightforward.

We consider an ergodic Markov chain whose states lie in a finite or countably infinite subset of $\Re^N$. By indexing the states with positive integers, we can view the state space as the set $S = \{1, \ldots, n\}$, where n is possibly infinite. We denote the vector of coordinates at which a state $i \in S$ is located by $\sigma(i) \in \Re^N$. Note that we could alternatively index states using the set $\{\sigma(i) \mid i \in S\}$. However, this would somewhat complicate notation in our analysis. The sequence of states visited by the Markov chain is denoted by $\{i_t \mid t = 0, 1, \ldots\}$. The Markov chain is described by a (finite or infinite) transition probability matrix P whose (i,j)th entry, denoted by p_{ij}, is the probability that $i_{t+1} = j$ given that $i_t = i$. For any pair (i,j), we are given a scalar $g(i,j)$ that represents the cost of a transition from i to j. (Extensions to the case where the one-stage costs are random is discussed in our conclusions section.) Finally, we let $\alpha \in (0,1)$ be a discount factor.

The cost-to-go function $J^* : S \mapsto \Re$ associated with this Markov chain is defined by

$$J^*(i) \stackrel{\triangle}{=} E\left[\sum_{t=0}^{\infty} \alpha^t g(i_t, i_{t+1}) \mid i_0 = i\right],$$

assuming that this expectation is well-defined. It is convenient to view J^* as a vector instead of a function (its dimension is infinite if S is infinite).

We consider approximations of J^* using a function of the form

$$\tilde{J}(i,r) = \sum_{k=1}^{K} r(k)\phi_k(i).$$

Here, $r = (r(1), \ldots, r(K))$ is a parameter vector and each ϕ_k is a fixed scalar function defined on the state space S. The functions ϕ_k can be viewed as basis functions (or as vectors of dimension $|S|$), while each $r(k)$ can be viewed as the associated weight. To approximate the cost-to-go function, one usually tries to choose the parameter vector r so as to minimize some error metric between the functions $\tilde{J}(\cdot, r)$ and $J^*(\cdot)$.

It is convenient to define a vector-valued function $\phi : S \mapsto \Re^K$, by letting $\phi(i) = (\phi_1(i), \ldots, \phi_K(i))$. With this notation, the approximation can also be written in the form

$$\tilde{J}(i,r) = r'\phi(i),$$

or

$$\tilde{J}(r) = \Phi' r,$$

where Φ is viewed as a $K \times |S|$ matrix whose ith column is equal to $\phi(i)$; that is,

$$\Phi = \begin{bmatrix} | & & | \\ \phi(1) & \cdots & \phi(n) \\ | & & | \end{bmatrix}.$$

Note that

$$\nabla \tilde{J}(i,r) = \phi(i),$$

where the gradient is the vector of partial derivatives with respect to the components of r, and we have

$$\nabla \tilde{J}(r) = \Phi,$$

where $\nabla \tilde{J}(r)$ is the Jacobian matrix whose ith column is equal to $\nabla \tilde{J}(i, r)$.

Suppose that we observe a sequence of states i_t generated according to the transition probability matrix P and that at time t the parameter vector r has been set to some value r_t. We define the temporal difference d_t corresponding to the transition from i_t to i_{t+1} by

$$d_t = g(i_t, i_{t+1}) + \alpha \tilde{J}(i_{t+1}, r_t) - \tilde{J}(i_t, r_t).$$

Then, for $t = 0, 1, \ldots$, the temporal-difference learning method updates r_t according to the formula

$$\begin{aligned} r_{t+1} &= r_t + \gamma_t d_t \sum_{k=0}^{t} (\alpha\lambda)^{t-k} \nabla \tilde{J}(i_k, r_t) \\ &= r_t + \gamma_t d_t \sum_{k=0}^{t} (\alpha\lambda)^{t-k} \phi(i_k), \end{aligned}$$

where r_0 is initialized to some arbitrary vector, γ_t is a sequence of scalar step sizes, and λ is a parameter in $[0, 1]$. Since temporal-difference learning is actually a continuum of algorithms, parameterized by λ, it is often referred to as TD(λ).

A more convenient representation of TD(λ) is obtained if we define a sequence of *eligibility vectors* z_t (of dimension K) by

$$z_t = \sum_{k=0}^{t} (\alpha\lambda)^{t-k} \phi(i_k).$$

With this new notation, the TD(λ) updates are given by

$$r_{t+1} = r_t + \gamma_t d_t z_t,$$

and the eligibility vectors can be updated according to

$$z_{t+1} = \alpha\lambda z_t + \phi(i_{t+1}),$$

initialized with $z_{-1} = 0$.

3 Understanding Temporal-Difference Learning

Temporal-difference learning originated in the field of reinforcement learning. A view commonly adopted in the original setting is that the algorithm involves "looking back in time and correcting previous predictions." In this context, the eligibility vector keeps track of how the parameter vector should be adjusted in order to appropriately modify prior predictions when a temporal-difference is observed. In this paper, we take a different view which involves examining the "steady-state" behavior of the algorithm and arguing that this characterizes the long-term evolution of the parameter vector. In the remainder of this section, we introduce this view of TD(λ) and provide an overview of the analysis that it leads to.

Our goal in this section is to convey some intuition about how the algorithm works, and in this spirit, we maintain the discussion at an informal level, omitting technical assumptions and other details required to formally prove the statements we make. These technicalities will be addressed in subsequent sections, where formal proofs are presented.

We begin by introducing some notation that will make our discussion here, as well as the analysis later in the paper, more concise. Let $\pi(1), \ldots, \pi(n)$ denote the steady-state probabilities for the process i_t. We assume that $\pi(i) > 0$ for all $i \in S$. We define an $n \times n$ diagonal matrix D with diagonal entries $\pi(1), \ldots, \pi(n)$. It is easy to see that $\langle x, y \rangle_D \overset{\triangle}{=} x'Dy$ satisfies the requirements for an inner product on $\Re^n$. We denote the norm on this inner product space by $\|\cdot\|_D = \sqrt{\langle \cdot, \cdot \rangle_D}$, and the set of vectors $\{J \in \Re^n \mid \|J\|_D < \infty\}$ by $L_2(S, D)$. As we will later prove, J^* lies in $L_2(S, D)$, and it is in this inner product space that the approximations $\tilde{J}(r_t) = \Phi' r_t$ evolve. Regarding notation, we will also keep using $\|\cdot\|$, without a subscript, to denote the Euclidean norm on finite-dimensional vectors or the Euclidean-induced norm on finite matrices. (That is, for any matrix A, we have $\|A\| = \max_{\|x\|=1} \|Ax\|$.)

We define a "projection matrix" (more precisely, projection operator) Π that projects onto the subspace $\{\Phi' r \mid r \in \Re^K\}$, with respect to the inner product $\langle \cdot, \cdot \rangle_D$. Assuming that the basis functions ϕ_k, $k = 1, \ldots, K$, are linearly independent, the projection matrix is given by

$$\Pi = \Phi'(\Phi D \Phi')^{-1} \Phi D. \tag{1}$$

(Note that $\Phi D \Phi'$ is a $K \times K$ matrix.) By definition of a projection matrix, we have for any vector $J \in L_2(S, D)$,

$$\Pi J = \min_r \|J - \Phi' r\|_D.$$

Note that ΠJ^* is a natural approximation to J^*, given the fixed set of basis functions. In fact, ΠJ^* is the solution to the weighted linear least-squares problem of minimizing

$$\sum_{i \in S} \pi(i) (J^*(i) - \tilde{J}(r, i))^2$$

with respect to r. Note that the error associated with each state is weighed by the frequency with which the state is visited. (If the state space was continuous instead of countable, this sum would be replaced by an integral.)

We define an operator $T^{(\lambda)} : \Re^n \mapsto \Re^n$, indexed by a parameter $\lambda \in [0, 1)$ by

$$(T^{(\lambda)} J)(i) = (1 - \lambda) \sum_{m=0}^{\infty} \lambda^m E \left[\sum_{t=0}^{m} \alpha^t g(i_t, i_{t+1}) + \alpha^{m+1} J(i_{m+1}) \mid i_0 = i \right].$$

For $\lambda = 1$ we define

$$(T^{(1)} J)(i) = E \left[\sum_{t=0}^{\infty} \alpha^t g(i_t, i_{t+1}) \mid i_0 = i \right] = J^*(i),$$

so that $\lim_{\lambda \uparrow 1} (T^{(\lambda)} J)(i) = (T^{(1)} J)(i)$ (under some technical conditions). To interpret this operator in a meaningful manner, note that, for each m, the term

$$E \left[\sum_{t=0}^{m} \alpha^t g(i_t, i_{t+1}) + \alpha^{m+1} J(i_{m+1}) \mid i_0 = i \right]$$

is the expected cost to be incurred over m transitions plus an approximation to the remaining cost to be incurred, based on J. This sum is sometimes called the "m–stage truncated cost-to-go." Intuitively, if J is an approximation to the cost-to-go function, the m–stage truncated cost-to-go can be viewed as an improved approximation. Since $T^{(\lambda)}J$ is a weighted average over the m–stage truncated cost-to-go values, $T^{(\lambda)}J$ can also be viewed as an improved approximation to J^*. In fact, we will prove later that $T^{(\lambda)}$ is a contraction on $L_2(S, D)$, whose fixed point is J^*. Hence, $T^{(\lambda)}J$ is always closer to J^* than J is, in the sense of the norm $\|\cdot\|_D$.

To clarify the fundamental structure of TD(λ), we construct a process $X_t = (i_t, i_{t+1}, z_t)$. It is easy to see that X_t is a Markov process. In particular, z_{t+1} and i_{t+1} are deterministic functions of X_t and the distribution of i_{t+2} only depends on i_{t+1}. Note that at each time t, the random vector X_t, together with the current parameter vector r_t, provides all necessary information for computing r_{t+1}. By defining a function s with

$$s(r, X) = (g(i,j) + \alpha\tilde{J}(j,r) - \tilde{J}(i,r))z,$$

where $X = (i, j, z)$, we can rewrite the TD(λ) algorithm as

$$r_{t+1} = r_t + \gamma_t s(r_t, X_t).$$

As we will show later, for any r, $s(r, X_t)$ has a well defined "steady-state" expectation, which we denote by $E_0[s(r, X_t)]$. Intuitively, once X_t reaches steady-state, the TD(λ) algorithm, in an "average" sense, behaves like the following deterministic algorithm:

$$\bar{r}_{\tau+1} = \bar{r}_\tau + \gamma_\tau E_0[s(\bar{r}_\tau, X_t)].$$

Under some technical assumptions, the convergence of this deterministic algorithm implies convergence of TD(λ) and both algorithms share the same limit of convergence. Our study centers on an analysis of this deterministic algorithm.

It turns out that

$$E_0[s(r, X_t)] = \Phi D\Big(T^{(\lambda)}(\Phi' r) - \Phi' r\Big),$$

and thus, the deterministic algorithm takes on the form

$$\bar{r}_{t+1} = \bar{r}_t + \gamma_t \Phi D\Big(T^{(\lambda)}(\Phi'\bar{r}_t) - \Phi'\bar{r}_t\Big).$$

As a side note, observe that the execution of this deterministic algorithm would require knowledge of transition probabilities and the transition costs between all pairs of states, and, when the state space is large or infinite, this is not an implementable algorithm. Indeed, stochastic approximation algorithms like TD(λ) are motivated by the need to alleviate such stringent information and computational requirements. We introduce the deterministic algorithm solely for conceptual purposes, and not as a feasible alternative for practical use.

To gain some additional insight about the evolution of $\bar{r}_t$, we rewrite the deterministic algorithm in the following form

$$\bar{r}_{t+1} = \bar{r}_t + \gamma_t \nabla\tilde{J}(\bar{r}_t) D\Big(T^{(\lambda)}(\Phi'\bar{r}_t) - \Phi'\bar{r}_t\Big). \qquad (2)$$

Note that in the case of $\lambda = 1$, this becomes

$$\bar{r}_{t+1} = \bar{r}_t - \frac{\gamma_t}{2}\nabla\|J^* - \Phi'\bar{r}_t\|_D^2,$$

which is the iteration for a steepest descent method that minimizes

$$\sum_{i\in S} \pi(i)\Big(J^*(i) - \tilde{J}(r,i)\Big)^2$$

with respect to r. It is easy to show that, if the step sizes are appropriately chosen, $\Phi'\bar{r}_t$ will converge to ΠJ^*.

In the case of $\lambda < 1$, we can think of each iteration of the deterministic algorithm as that of a steepest descent method for minimizing

$$\sum_{i\in S} \pi(i)\Big((T^{(\lambda)}(\Phi'\bar{r}_t))(i) - \tilde{J}(r,i)\Big)^2$$

with respect to r, given a fixed r_t. Note, however, that the error function changes from one time step to the next, and therefore, it is not a true steepest descent method, which would involve a fixed error function. Nevertheless, if we think of $T^{(\lambda)}(\Phi'\bar{r}_t)$ as an approximation to J^*, the algorithm makes some intuitive sense. However, some subtleties are involved here.

To illustrate this, consider a probability distribution $q(\cdot)$ over the state space S, that is different from the steady-state distribution $\pi(\cdot)$. Define a diagonal matrix Q with diagonal entries $q(1), \dots, q(n)$. If we replace the matrix D in the deterministic variant of TD(1) with the matrix Q, we obtain

$$\bar{r}_{t+1} = \bar{r}_t - \frac{\gamma_t}{2}\nabla\|J^* - \Phi'\bar{r}_t\|_Q^2,$$

which is a steepest descent method that minimizes

$$\sum_{i\in S} q(i)\Big(J^*(i) - \tilde{J}(r,i)\Big)^2$$

with respect to r. If step sizes are appropriately chosen, $\Phi'\bar{r}_t$ will converge to $\Pi_Q J^*$, where Π_Q is the projection matrix with respect to the inner product $\langle\cdot,\cdot\rangle_Q$. On the other hand, if we replace D with Q in the TD(λ) algorithm for $\lambda < 1$, the algorithm might not converge at all! We will formally illustrate this phenomenon in Section 9.

To get a better grasp on the fundamental issues involved here, let us consider a more general algorithm that takes on the form

$$\bar{r}_{t+1} = \bar{r}_t + \gamma_t \nabla\tilde{J}(\bar{r}_t)Q\Big(F(\Phi'\bar{r}_t) - \Phi'\bar{r}_t\Big), \tag{3}$$

where F is a contraction with respect to $\|\cdot\|_D$. Note that by letting $F = T^{(\lambda)}$, we recover the deterministic variant of TD(λ). Like TD(λ), each iteration given by Equation (3) can be thought of as a steepest descent iteration on an error function given by

$$\sum_{i\in S} q(i)\Big((F(\Phi'\bar{r}_t))(i) - \tilde{J}(r,i)\Big)^2.$$

(The variable being optimized is r, while r_t remains fixed.) Note that the minimum of this (time-varying) error function at time t is given by $\Pi_Q F(\Phi'\bar{r}_t)$. Hence, letting $J_t = \Phi'\bar{r}_t$, we might think of $\Pi_Q F(J_t)$ as a "target vector," given a current vector J_t. We can define an algorithm of the form

$$J_{t+1} = \Pi_Q F(J_t), \tag{4}$$

which moves directly to the target, given a current vector J_t.

Intuitively, the iteration of Equation (3) can be thought of as an incremental form of Equation (4). Hence, one might expect the two algorithms to have similar convergence properties. In fact, they do. Concerning convergence of the algorithm given by Equation (4), note that if F is a contraction of the norm $\|\cdot\|_Q$, then the composition $\Pi_Q F(\cdot)$ is also a contraction of the norm $\|\cdot\|_Q$, since the projection Π_Q is a nonexpansion of that norm. However, there is no reason to believe that the projection Π_Q will be a nonexpansion of the norm $\|\cdot\|_D$ if $D \neq Q$. In this case, $\Pi_Q F(\cdot)$ may not be a contraction, and might even be an expansion. Hence, convergence can be guaranteed by the algorithms of Equations (4) and (3) if and only if the contraction and projection are with respect to the same norm. This idea captures exactly the issue that arises with variants of TD(λ) that sample states with frequencies independent of the dynamics of the Markov process. In particular, the state sampling frequencies are reflected in the matrix Q, while the dynamics of the Markov process make $T^{(\lambda)}$ a contraction with respect to $\|\cdot\|_D$. When states are sampled on-line, we have $Q = D$, while there is no such promise when states are sampled by an independent mechanism.

For another perspective on TD(λ), note that the deterministic variant, as given by Equation (2), can be rewritten in the form

$$\bar{r}_{t+1} = \bar{r}_t + \gamma_t(A\bar{r}_t + b),$$

for some matrix A and vector b. As we will show later, the contraction property of $T^{(\lambda)}$ and the fact that Π is a projection with respect to the same norm imply that the matrix A is negative definite. From this fact, it is easy to see that the iteration converges, given appropriate step size constraints. However, it is difficult to draw an intuitive understanding from the matrix A, as we did for the operators $T^{(\lambda)}$ and Π. Nevertheless, for simplicity of proof, we use the representation in terms of A and b when we establish that TD(λ) has the properties required for application of the available machinery from stochastic approximation. This machinery is what allows us to deduce convergence of the actual (stochastic) algorithm from that of the deterministic counterpart.

4 Convergence Result

In this section we present the main result of this paper, which establishes convergence and characterizes the limit of convergence of temporal-difference learning. We begin by stating the required assumptions.

The first assumption places constraints on the underlying infinite-horizon discounted Markov chain. Essentially, we assume that the Markov chain is ergodic, the steady-state variance of transition costs is finite, and that the cost-to-go from any state is well defined and finite. The formal statement follows:

Assumption 1 *(a) The Markov chain i_t has a unique invariant distribution π that satisfies*

$$\pi' P = \pi',$$

with $\pi(i) > 0$ for all i; here, π is a finite or infinite vector, depending on the cardinality of S. Let $E_0[\,\cdot\,]$ stand for expectation with respect to this distribution.

(b) The transition costs $g(i_t, i_{t+1})$ satisfy

$$E_0[g^2(i_t, i_{t+1})] < \infty.$$

(c) For every i, the expectation in the definition of the cost-to-go $J^*(i)$ is well-defined and finite.

In fact, we will show in Lemma 2 that Assumption 1(c) is a consequence of parts (a) and (b), and is therefore unnecessary.

Our second assumption ensures that the basis functions used for approximation are linearly independent and do not grow too fast.

Assumption 2 *(a) The matrix Φ has full row rank; that is, the basis functions $\{\phi_k \mid k = 1, \ldots, K\}$ are linearly independent.*
(b) For every k, the basis function ϕ_k satisfies

$$E_0[\phi_k^2(i)] < \infty.$$

The next assumption essentially requires that the Markov chain has a certain "degree of stability" and that the functions $\phi(\cdot)$ and $g(\cdot, \cdot)$ do not grow too fast. As will be shown in Section 6, this assumption is always satisfied when the state space S is finite. It is also satisfied in many situations of practical interest when the set S is infinite. Further discussion is given in Section 7.

Assumption 3 *(a) For any $q > 1$, there exists a constant μ_q such that for all i, t,*

$$E[\|\sigma(i_t)\|^q | i_0] \leq \mu_q(1 + \|\sigma(i_0)\|^q).$$

(b) There exist positive constants C_1, q_1 such that

$$\|\phi(i)\| \leq C_1(1 + \|\sigma(i)\|^{q_1}),$$

and

$$|g(i,j)| \leq C_1(1 + \|\sigma(i)\|^{q_1} + \|\sigma(j)\|^{q_1}).$$

(c) There exist positive constants C_2, q_2 such that, for all i_0 and $m \geq 0$,

$$\sum_{\tau=0}^{\infty} \Big\| E[\phi(i_\tau)\phi'(i_{\tau+m}) | i_0] - E_0[\phi(i_t)\phi'(i_{t+m})] \Big\| \leq C_2(1 + \|\sigma(i_0)\|^{q_2}),$$

and

$$\sum_{\tau=0}^{\infty} \Big\| E[\phi(i_\tau) g(i_{\tau+m}, i_{\tau+m+1}) | i_0] - E_0[\phi(i_t) g(i_{\tau+m}, i_{\tau+m+1})] \Big\| \leq C_2(1 + \|\sigma(i_0)\|^{q_2}).$$

Part of the above assumption is that the expectations in part (c) are all well-defined and finite. However, it will be seen later that this is actually a consequence of our earlier assumptions.

Our final assumption places the usual constraints on the sequence of step sizes.

Assumption 4 *The step sizes γ_t are nonnegative, nonincreasing, and predetermined (chosen prior to execution of the algorithm). Furthermore, they satisfy*

$$\sum_{t=0}^{\infty} \gamma_t = \infty,$$

and

$$\sum_{t=0}^{\infty} \gamma_t^2 < \infty.$$

The main result of this paper follows.

Theorem 1 *Under Assumptions 1, 2, 3, and 4, the following hold:*
(a) The cost-to-go function J^* is in $L_2(S, D)$.
(b) For any $\lambda \in [0,1]$, the TD(λ) algorithm, as defined in Section 2, converges with probability 1.
(c) The limit of convergence r^* is the unique solution of the equation

$$\Pi T^{(\lambda)}(\Phi' r^*) = \Phi' r^*.$$

(d) Furthermore, r^* satisfies

$$\|\Phi' r^* - J^*\|_D \leq \frac{\|\Pi J^* - J^*\|_D}{1 - \alpha(1-\lambda)/(1-\lambda\alpha)}.$$

In order to place Theorem 1 in perspective, let us discuss its relation with respect to available results. If one lets $\phi(i)$ be the ith unit vector for each i, and if we assume that S is finite, we are dealing with a lookup table representation of the cost-to-go function. In that case, we recover a result similar to those in (Jaakola et al., 1994) (actually, that paper dealt with the on-line TD(λ) algorithm only for Markov chains involving a termination state). We note that with a lookup table representation, the operator $T^{(\lambda)}$ is easily shown to be a maximum norm contraction, and general results on stochastic approximation methods based on contraction mappings (Jaakola et al., 1994; Tsitsiklis, 1994) become applicable. However, this contraction property is lost once function approximation is introduced and this approach does not extend.

Closer to our results is the work of Dayan (1992) who considered TD(λ) for the case of linearly parameterized compact representations and established convergence in the mean. However, convergence in the mean is a much weaker convergence property than the one we establish here. Finally, the work of Dayan and Sejnowski (1992) contains a sketch of a proof of convergence with probability 1. However, it is restricted to the case where the vectors $\phi(i)$ are linearly independent, which is essentially equivalent to having a lookup table representation. (A more formal proof, for this restricted case has been developed in (Gurvits et al., 1994).) Some of the ideas in our method of proof originate in the work of Sutton (1988) and Dayan (1992). Our analysis also leads to an interpretation of the limit of convergence. In particular, Theorem 1 offers an illuminating fixed–point equation, as well as a graceful bound on the approximation error. Previous works lack interpretable results of this kind.

5 Preliminaries

In this section we present a series of lemmas that contain all of the essential ideas behind Theorem 1. We also state a theorem concerning stochastic approximation that will be used to establish convergence.

We begin with a lemma that deals with a general property of ergodic Markov chains. This lemma is central to our analysis and will be often used in the sequel.

Lemma 1 *Under Assumption 1(a), for any $J \in L_2(S, D)$, we have $\|PJ\|_D \leq \|J\|_D$.*

Proof: The proof involves Jensen's inequality, the property $\pi' P = \pi'$, and some simple algebra:

$$\begin{aligned}
\|PJ\|_D^2 &= J'P'DPJ \\
&= \sum_{i=1}^n \pi(i) \left(\sum_{j=1}^n p_{ij} J(j) \right)^2 \\
&\leq \sum_{i=1}^n \pi(i) \sum_{j=1}^n p_{ij} J^2(j) \\
&= \sum_{j=1}^n \sum_{i=1}^n \pi(i) p_{ij} J^2(j) \\
&= \sum_{j=1}^n \pi(j) J^2(j) \\
&= \|J\|_D^2.
\end{aligned}$$

q.e.d.

Lemma 1 is useful in showing that J^* is in $L_2(S, D)$. In particular, we have the following result, where we use the notation $\bar{g}$ to denote the vector of dimension $|S|$ whose ith component is equal to $E[g(i_t, i_{t+1}) | i_t = i]$.

Lemma 2 *Under Assumptions 1(a)-(b), $J^*(i)$ is well-defined and finite for every $i \in S$. Furthermore, J^* is in $L_2(S, D)$, and*

$$J^* = \sum_{t=0}^{\infty} (\alpha P)^t \bar{g}.$$

Proof: If the Markov chain starts in steady state, it remains in steady-state, and therefore,

$$E_0 \left[\sum_{t=0}^{\infty} \alpha^t g^2(i_t, i_{t+1}) \right] = \frac{1}{1-\alpha} E_0[g^2(i_0, i_1)] < \infty,$$

where we are using the monotone convergence theorem to interchange the expectation and the summation, as well as Assumption 1(b). Since $|g(i_t, i_{t+1})| \leq 1 + g^2(i_t, i_{t+1})$, it follows that

$$\sum_{\imath \in S} \pi(i) E \left[\sum_{t=0}^{\infty} \alpha^t |g(i_t, i_{t+1})| \Big| i_0 = i \right] = E_0 \left[\sum_{t=0}^{\infty} \alpha^t |g(i_t, i_{t+1})| \right] < \infty.$$

Since $\pi(i) > 0$ for all i, the expectation defining $J^*(i)$ is well-defined and finite.

Using Fubini's Theorem to switch the order of expectation and summation in the definition of J^*, we obtain

$$\begin{aligned} J^*(i) &\triangleq E\Big[\sum_{t=0}^{\infty} \alpha^t g(i_t, i_{t+1}) | i_0 = i\Big] \\ &= \sum_{t=0}^{\infty} \alpha^t E[g(i_t, i_{t+1}) | i_0 = i] \\ &= \sum_{t=0}^{\infty} \alpha^t E[\bar{g}(i_t) | i_0 = i], \end{aligned}$$

and it follows that

$$J^* = \sum_{t=0}^{\infty} (\alpha P)^t \bar{g}.$$

To show that J^* is in $L_2(S, D)$, we have

$$\begin{aligned} \|J^*\|_D &\leq \sum_{t=0}^{\infty} \|(\alpha P)^t \bar{g}\|_D \\ &\leq \sum_{t=0}^{\infty} \alpha^t \|\bar{g}\|_D \\ &= \frac{\|\bar{g}\|_D}{1-\alpha}, \end{aligned}$$

where the second inequality follows from Lemma 1. Note that we have

$$\begin{aligned} \|\bar{g}\|_D^2 &= \sum_{i \in S} \pi(i) \Big(\sum_{j \in S} p_{ij} g(i,j) \Big)^2 \\ &\leq \sum_{i \in S} \pi(i) \sum_{j \in S} p_{ij} g^2(i,j) \\ &= E_0[g^2(i_t, i_{t+1})] \\ &< \infty, \end{aligned}$$

by Assumption 1(b). It follows that J^* is in $L_2(S, D)$. **q.e.d.**

The next lemma states that that the operator $T^{(\lambda)}$ maps $L_2(S, D)$ into itself and provides a formula for evaluating $T^{(\lambda)} J$.

Lemma 3 *Under Assumption 1, for any $J \in L_2(S, D)$, $T^{(\lambda)}(J)$ is in $L_2(S, D)$, and for $\lambda \in [0, 1)$, we have*

$$T^{(\lambda)} J = (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \left(\sum_{t=0}^{m} (\alpha P)^t \bar{g} + (\alpha P)^{m+1} J \right).$$

Proof:

$$\begin{aligned} (T^{(\lambda)} J)(i) &= (1-\lambda) \sum_{m=0}^{\infty} \lambda^m E\left[\sum_{t=0}^{m} \alpha^t g(i_t, i_{t+1}) + \alpha^{m+1} J(i_{m+1}) \mid i_0 = i \right] \\ &= (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \left(\sum_{t=0}^{m} \alpha^t E[\bar{g}(i_t) \mid i_0 = i] + \alpha^{m+1} E[J(i_{m+1}) \mid i_0 = i] \right), \end{aligned}$$

and the formula in the statement of the lemma follows.

We have shown in Lemma 2 that $\|\bar{g}\|_D < \infty$. Thus, for $\lambda < 1$, we can use Lemma 1 to obtain

$$\|(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\sum_{t=0}^{m}(\alpha P)^t\bar{g}\|_D \le (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\sum_{t=0}^{m}\alpha^t\|\bar{g}\|_D < \infty.$$

Similarly,

$$\begin{aligned}\|(1-\lambda)\sum_{m=0}^{\infty}\lambda^m(\alpha P)^{m+1}J\|_D &\le (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\alpha^{m+1}\|J\|_D \\ &\le \alpha\|J\|_D,\end{aligned}$$

for any $J \in L_2(S, D)$, by Lemma 1. This completes the proof. **q.e.d.**

Lemma 1 can also be used to show that $T^{(\lambda)}$ is a contraction on $L_2(S, D)$. This fact, which is captured by the next lemma, will be useful for establishing error bounds.

Lemma 4 *Under Assumption 1(a), for any $J, \bar{J} \in L_2(S, D)$, we have*

$$\|T^{(\lambda)}J - T^{(\lambda)}\bar{J}\|_D \le \frac{\alpha(1-\lambda)}{1-\alpha\lambda}\|J-\bar{J}\|_D \le \alpha\|J-\bar{J}\|_D.$$

Proof: The case of $\lambda = 1$ is trivial. For $\lambda < 1$, the result follows from Lemmas 3 and 1. In particular, we have

$$\begin{aligned}\|T^{(\lambda)}J - T^{(\lambda)}\bar{J}\|_D &= \left\|(1-\lambda)\sum_{m=0}^{\infty}\lambda^m(\alpha P)^{m+1}(J-\bar{J})\right\|_D \\ &\le (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\alpha^{m+1}\|J-\bar{J}\|_D \\ &= \frac{\alpha(1-\lambda)}{1-\alpha\lambda}\|J-\bar{J}\|_D.\end{aligned}$$

q.e.d.

The next lemma characterizes the fixed point of the composition $\Pi T^{(\lambda)}$. This fixed point must lie in the range of Π, which is the space $\{\Phi' r | r \in \Re^K\}$ (note that this is a subspace of $L_2(S, D)$, because of Assumption 2(b)). The lemma establishes existence and uniqueness of this fixed point, which we will denote by r^*. We will show in the next section that this r^* is also the limit of convergence of r_t.

Lemma 5 *Under Assumptions 1 and 2, $\Pi T^{(\lambda)}(\cdot)$ is a contraction and has a unique fixed point which is of the form $\Phi' r^*$ for a unique choice of r^*. Furthermore, r^* satisfies the following bound:*

$$\|\Phi' r^* - J^*\|_D \le \frac{\|\Pi J^* - J^*\|_D}{1 - \alpha(1-\lambda)/(1-\alpha\lambda)}.$$

Proof: Lemma 4 ensures that $T^{(\lambda)}$ is a contraction from $L_2(S, D)$ into itself, and it is easily seen that J^* is a fixed point. Since projections are nonexpansive, the composition $\Pi T^{(\lambda)}(\cdot)$ is also a contraction. It follows that $\Pi T^{(\lambda)}(\cdot)$ has a unique fixed point of the form $\Phi' r^*$, for some r^*. Because the functions $\phi_k(\cdot)$ are assumed linearly independent, it follows that the choice of r^* is unique.

Using the fact that J^* is in $L_2(S,D)$ (Lemma 2), we establish the desired bound. We have

$$\begin{aligned} \|\Phi' r^* - J^*\|_D &\leq \|\Phi' r^* - \Pi J^*\|_D + \|\Pi J^* - J^*\|_D \\ &= \|\Pi T^{(\lambda)}(\Phi' r^*) - \Pi J^*\|_D + \|\Pi J^* - J^*\|_D \\ &\leq \frac{\alpha(1-\lambda)}{1-\alpha\lambda}\|\Phi' r^* - J^*\|_D + \|\Pi J^* - J^*\|_D, \end{aligned}$$

and it follows that

$$\|\Phi' r^* - J^*\|_D \leq \frac{\|\Pi J^* - J^*\|_D}{1 - \alpha(1-\lambda)/(1-\alpha\lambda)}.$$

q.e.d.

We next set out to characterize the expected behavior of the steps taken by the TD(λ) algorithm in "steady-state." In particular, we will get a handle on $E_0[s(r, X_t)]$ for any given r. While this expression can be viewed as a limit of $E[s(r,X_t)|X_0]$ as t goes to infinity, it is simpler to view it as an expectation referring to a process that is already in steady-state. We therefore make a short digression to construct a stationary process X_t.

We proceed as follows. Let $\{i_t\}$ be a Markov chain that evolves according to the transition probability matrix P and which is already in steady-state, in the sense that $\Pr(i_t = i) = \pi(i)$ for all i and all t. Given any sample path of this Markov chain, we define

$$z_t = \sum_{\tau=-\infty}^{t} (\alpha\lambda)^{t-\tau}\phi(i_\tau). \tag{5}$$

Note that z_t is constructed by taking the stationary process $\phi(i_t)$, whose variance is finite (Assumption 2), and passing it through an exponentially stable linear time invariant system. It is then well known that the output z_t of this filter is finite with probability 1, and has also finite variance. With z_t so constructed, we let $X_t = (i_t, i_{t+1}, z_t)$ and note that this is a Markov process with the same transition probabilities as the Markov process X_t that was constructed in the middle of Section 3 (the evolution equation is the same). The only difference is that the process X_t of Section 3 was initialized with $z_{-1} = 0$, whereas here we have a stationary process X_t whose statistics are time invariant. We can now identify $E_0[\,\cdot\,]$ with the expectation with respect to this invariant distribution.

Prior to studying $E_0[s(r,X_t)]$, let us establish a few preliminary relations in the next lemma.

Lemma 6 *Under Assumptions 1, 2, and 3, the following relations hold:*
(a) $E_0[\phi(i_t)\phi'(i_{t+m})] = \Phi D P^m \Phi'$,
(b) There exists a finite constant G such that $\|E_0[\phi(i_t)\phi'(i_{t+m})]\| \leq G$, *for all* m.
(c) $E_0[z_t\phi'(i_t)] = \sum_{m=0}^{\infty}(\alpha\lambda)^m \Phi D P^m \Phi'$,
(d) $E_0[z_t\phi'(i_{t+1})] = \sum_{m=0}^{\infty}(\alpha\lambda)^m \Phi D P^{m+1} \Phi'$,
(e) $E_0[z_t g(i_t, i_{t+1})] = \sum_{m=0}^{\infty}(\alpha\lambda)^m \Phi D P^m \bar{g}$.
Furthermore, each of the above expressions is well defined and finite.

Proof: We first observe that for any $J, \bar{J} \in L_2(S,D)$, we have

$$E_0[J(i_t)\bar{J}(i_{t+m})] = \sum_{i\in S}\pi(i)\sum_{j\in S}\Pr(i_{t+m} = j \mid i_t = i)J(i)\bar{J}(j)$$

$$
\begin{aligned}
&= \sum_{i\in S} \pi(i)J(i)[P^m \bar{J}](i) \\
&= J'DP^m\bar{J}.
\end{aligned}
$$

(Note that $P^m\bar{J} \in L_2(S,D)$, by Lemma 1, and using the Cauchy-Schwartz inequality, $J'DP^m\bar{J}$ is finite.) By specializing to the case where we are dealing with vectors of the form $J = \Phi' r$ and $\bar{J} = \Phi'\bar{r}$ (these vectors are in $L_2(S,D)$ as a consequence of Assumption 2), we obtain

$$E_0[r'\phi(i_t)\phi'(i_{t+m})\bar{r}] = r'\Phi DP^m\Phi'\bar{r}.$$

Since the vectors r and $\bar{r}$ are arbitrary, it follows that

$$E_0[\phi(i_t)\phi'(i_{t+m})] = \Phi DP^m\Phi'.$$

We place a bound on the Euclidean induced matrix norm $\|\Phi DP^m\Phi'\|$ as follows. We have

$$
\begin{aligned}
\|\Phi DP^m\Phi'\| &\le K^2 \max_{k,j} |\phi_k DP^m\phi_j'| \\
&= K^2 \max_{k,j} |\phi_k D^{\frac{1}{2}} D^{\frac{1}{2}} P^m \phi_j'| \\
&\le K^2 \max_{k,j} \|\phi_k'\|_D \|P^m\phi_j'\|_D \\
&\le K^2 \max_{k} \|\phi_k'\|_D^2 \\
&= K^2 \max_{k} E_0[\phi_k^2(i)],
\end{aligned}
$$

which is a finite constant G, by Assumption 2(b). We have used here the notation ϕ_k to indicate the kth row of the matrix Φ, with entries $\phi_k(1), \ldots, \phi_k(n)$; note that this is a row vector.

We have so far verified parts (a) and (b) of the lemma. We now begin with the analysis for part (c). Note that $E_0[z_t\phi'(i_t)]$ is the same for all t and it suffices to prove the result for the case $t = 0$. We have

$$
\begin{aligned}
E_0[z_0\phi'(i_0)] &= E_0\left[\sum_{\tau=-\infty}^{0} (\alpha\lambda)^{-\tau}\phi(i_\tau)\phi'(i_0)\right] \\
&= \sum_{\tau=-\infty}^{0} (\alpha\lambda)^{-\tau} E_0[\phi(i_\tau)\phi'(i_0)]
\end{aligned}
$$

where the interchange of summation and expectation is justified by the dominated convergence theorem. The desired result follows by using the result of part (a).

The results of parts (d) and (e) are proved by entirely similar arguments, which we omit. **q.e.d.**

With the previous technical lemma at hand, we are ready to characterize $E_0[s(r, X_t)]$. This is done in the following lemma.

Lemma 7 *Under Assumptions 1 and 2, we have*

$$E_0[s(r,X_t)] = \Phi D\Big(T^{(\lambda)}(\Phi' r) - \Phi' r\Big),$$

which is well defined and finite for any finite r.

Proof: By applying Lemma 6, we have

$$\begin{aligned}
E_0[s(r, X_t)] &= E_0[z_t g(i_t, i_{t+1}) + \alpha z_t \phi'(i_{t+1})r - z_t \phi'(i_t)r] \\
&= \Phi D \sum_{m=0}^{\infty} (\alpha\lambda P)^m (\bar{g} + \alpha P \Phi' r - \Phi' r) \\
&= \Phi D \Big(\sum_{m=0}^{\infty} (\alpha\lambda P)^m \bar{g} + \sum_{m=0}^{\infty} (\alpha\lambda P)^m (\alpha P - I)\Phi' r \Big).
\end{aligned}$$

It follows that, for $\lambda = 1$,

$$E_0[s(r, X_t)] = \Phi D(J^* - \Phi' r),$$

and for $\lambda < 1$, after a little bit of algebra,

$$\begin{aligned}
E_0[s(r, X_t)] &= \Phi D \left((1-\lambda) \sum_{m=0}^{\infty} \lambda^m \sum_{t=0}^{m} (\alpha P)^t \bar{g} + \Big((1-\lambda) \sum_{m=0}^{\infty} \lambda^m (\alpha P)^{m+1} - I \Big) \Phi' r \right) \\
&= \Phi D \Big(T^{(\lambda)}(\Phi' r) - \Phi' r \Big),
\end{aligned}$$

by Lemma 3. Each expression is finite and well defined by Lemma 6. **q.e.d.**

The next lemma shows that the steps taken by TD(λ) tend to move r_t towards r^*.

Lemma 8 *Under Assumptions 1 and 2, we have*

$$(r - r^*)' E_0[s(r, X_t)] < 0, \qquad \forall r \neq r^*.$$

Proof: We have

$$\begin{aligned}
(r - r^*)' \Phi D \Big(T^{(\lambda)}(\Phi' r) - \Phi' r \Big) &= (r - r^*)' \Phi D \Big((I - \Pi) T^{(\lambda)}(\Phi' r) + \Pi T^{(\lambda)}(\Phi' r) - \Phi' r \Big) \\
&= (\Phi' r - \Phi' r^*)' D \Big(\Pi T^{(\lambda)}(\Phi' r) - \Phi' r \Big),
\end{aligned}$$

where the last equality follows because $\Phi D \Pi = \Phi D$ (see Eq. 1). As shown in the beginning of the proof of Lemma 5, $\Pi T^{(\lambda)}$ is a contraction with fixed point $\Phi' r^*$, and the contraction factor is no larger than α. Hence,

$$\|\Pi T^{(\lambda)}(\Phi' r) - \Phi' r^*\|_D \leq \alpha \|\Phi' r - \Phi' r^*\|_D,$$

and using the Cauchy–Schwartz inequality, we obtain

$$\begin{aligned}
(r - r^*)' \Phi D \Big(T^{(\lambda)}(\Phi' r) - \Phi' r \Big) &= (\Phi' r - \Phi' r^*)' D \Big(\Pi T^{(\lambda)}(\Phi' r) - \Phi' r^* + (\Phi' r^* - \Phi' r) \Big) \\
&\leq \|\Phi' r - \Phi' r^*\|_D \cdot \|\Pi T^{(\lambda)}(\Phi' r) - \Phi' r^*\|_D - \|\Phi' r - \Phi' r^*\|_D^2 \\
&\leq (\alpha - 1) \|\Phi' r - \Phi' r^*\|_D^2.
\end{aligned}$$

Since $\alpha < 1$, the result follows. **q.e.d.**

We now state without proof a general result concerning stochastic approximation, which will be used in the proof of Theorem 1. This result is a special case of a very general result on stochastic approximation algorithms (Theorem 17, in page 239 of (Benveniste et al., 1987)). It is straightforward to check that all of the assumptions in the result of (Benveniste et al., 1987) follow from the assumptions imposed in the result below. We do not show here the assumptions of (Benveniste et al., 1987) because the list is long and would require a lot in terms of new notation.

Theorem 2 *Consider an iterative algorithm of the form*

$$r_{t+1} = r_t + \gamma_t(A(X_t)r_t + b(X_t)),$$

where:
(a) The (predetermined) stepsize sequence γ_t is nonnegative, nonincreasing, and satisfies $\sum_{t=0}^{\infty} \gamma_t = \infty$ and $\sum_{t=0}^{\infty} \gamma_t^2 < \infty$.
(b) X_t is a Markov process with a unique invariant distribution and f is a bijective mapping from the states of the Markov process to a subset of $\Re^L$. Let $E_0[\cdot]$ stand for expectation with respect to this invariant distribution.
(c) $A(\cdot)$ and $b(\cdot)$ are matrix and vector valued functions, respectively. Furthermore, $A = E_0[A(X_t)]$ and $b = E_0[b(X_t)]$ are well defined and finite.
(d) The matrix A is negative definite.
(e) For any $q > 1$ there exists a constant μ_q such that for all X, t,

$$E[\|f(X_t)\|^q | X_0 = X] \leq \mu_q(1 + \|f(X)\|^q).$$

(f) There exist positive constants C_1, q_1 such that for all X,

$$\|A(X)\| \leq C_1(1 + \|f(X)\|^{q_1}),$$

and

$$\|b(X)\| \leq C_1(1 + \|f(X)\|^{q_1}).$$

(g) There exist positive constants C_2, q_2 such that for all X,

$$\sum_{t=0}^{\infty} \Big\| E[A(X_t) \mid X_0 = X] - A \Big\| \leq C_2(1 + \|f(X)\|^{q_2}),$$

and

$$\sum_{t=0}^{\infty} \Big\| E[b(X_t) \mid X_0 = X] - b \Big\| \leq C_2(1 + \|f(X)\|^{q_2}).$$

Then, r_t converges to r^*, with probability 1, where r^* is the unique vector that satisfies $Ar^* + b = 0$.

6 Proof of Theorem 1

The step $s(r_t, X_t)$ involved in the update of r_t is

$$\begin{aligned} s(r_t, X_t) &= d_t z_t \\ &= z_t g(i_t, i_{t+1}) + z_t(\alpha\phi'(i_{t+1}) - \phi'(i_t))r_t. \end{aligned}$$

Hence $s(r_t, X_t)$ takes on the form

$$s(r_t, X_t) = A(X_t)r_t + b(X_t),$$

where

$$A(X_t) = z_t(\alpha\phi'(i_{t+1}) - \phi'(i_t)),$$

and

$$b(X_t) = z_t g(i_t, i_{t+1}).$$

By Lemma 6, $A \overset{\triangle}{=} E_0[A(X_t)]$ and $b \overset{\triangle}{=} E_0[b(X_t)]$ are both well defined and finite.

By Lemma 5, we have $\Pi T^{(\lambda)}(\Phi' r^*) = \Phi' r^*$. From Equation (1), we also have $\Phi D \Pi = \Phi D$. Hence, $\Phi D T^{(\lambda)}(\Phi' r^*) = \Phi D \Phi' r^*$. We now compare with the formula for $E_0[s(r^*, X_t)]$, as given by Lemma 7, and conclude that $E_0[s(r^*, X_t)] = 0$. Hence,

$$\begin{aligned} A(r - r^*) &= E_0[s(r, X_t)] - E_0[s(r^*, X_t)] \\ &= E_0[s(r, X_t)]. \end{aligned}$$

It follows from Lemma 8 that

$$(r - r^*)' A (r - r^*) < 0,$$

for any $r \neq r^*$, and thus A is negative definite.

We will use Theorem 2 to show that r_t converges. Our analysis thus far ensures satisfaction of all required conditions, except for three technical ones – conditions (e), (f), and (g), in particular. We now show that Assumption 3 is sufficient to ensure satisfaction of these three conditions.

Without loss of generality, we assume the exponent q_1 satisfying Assumption 3(b) to be greater than 1. We define a function f mapping states $X_t = (i_t, i_{t+1}, z_t)$ to vectors $f(X_t) \in \Re^{2N+K}$ by

$$f(X_t) = \begin{bmatrix} \|\sigma(i_t)\|^{q_1 - 1} \cdot \sigma(i_t) \\ \|\sigma(i_{t+1})\|^{q_1 - 1} \cdot \sigma(i_{t+1}) \\ z_t \end{bmatrix}.$$

We start by addressing condition (e). Note that,

$$\begin{aligned} \|f(X_t)\| &\leq \|z_t\| + \|\sigma(i_t)\|^{q_1} + \|\sigma(i_{t+1})\|^{q_1} \\ &\leq \|z_0\| + \sum_{\tau=0}^{t} (\alpha\lambda)^{t-\tau} \|\phi(i_\tau)\| + \|\sigma(i_t)\|^{q_1} + \|\sigma(i_{t+1})\|^{q_1} \\ &\leq C\Big(\sum_{\tau=0}^{t} (\alpha\lambda)^{t-\tau} (1 + \|\sigma(i_\tau)\|^{q_1})\Big) + \|\sigma(i_t)\|^{q_1} + \|\sigma(i_{t+1})\|^{q_1} + \|z_0\|, \end{aligned}$$

for some positive constant C, where the final inequality follows from Assumption 3(b). It follows that, for any $q > 0$, there exists a constant C such that

$$\|f(X_t)\|^q \leq C \left(1 + \Big(\sum_{\tau=0}^{t} (\alpha\lambda)^{t-\tau} \|\sigma(i_\tau)\|^{q_1}\Big)^q + \|\sigma(i_t)\|^{q_1 q} + \|\sigma(i_{t+1})\|^{q_1 q} + \|z_0\|^q \right).$$

For $q > 1$, Jensen's inequality and the convexity of $(\cdot)^q$ give us

$$\Big(\sum_{\tau=0}^{t} (\alpha\lambda)^{\tau} |h(\tau)|\Big)^q \leq \frac{1}{(1 - \lambda\alpha)^{q-1}} \sum_{\tau=0}^{t} (\alpha\lambda)^{\tau} |h(\tau)|^q,$$

for any function h over the nonnegative integers. By specializing to the case of

$$h(\tau) = \|\sigma(i_{t-\tau})\|^{q_1},$$

we obtain

$$\Big(\sum_{\tau=0}^{t}(\alpha\lambda)^{\tau}\|\sigma(i_{t-\tau})\|^{q_1}\Big)^{q} \leq \frac{1}{(1-\lambda\alpha)^{q-1}}\sum_{\tau=0}^{t}(\alpha\lambda)^{\tau}\|\sigma(i_{t-\tau})\|^{q_1 q}.$$

It follows that, for any $q > 1$, there exists a constant C such that

$$\|f(X_t)\|^q \leq C\left(1+\sum_{\tau=0}^{t}(\alpha\lambda)^{t-\tau}\|\sigma(i_\tau)\|^{q_1 q}+\|\sigma(i_t)\|^{q_1 q}+\|\sigma(i_{t+1})\|^{q_1 q}+\|z_0\|^q\right), \qquad \forall t.$$

Taking expectations and applying Assumption 3(a), we have that for any q, there exists a constant C such that

$$E[\|f(X_t)\|^q \mid X_0] \leq C\left(1+\|\sigma(i_0)\|^{q_1 q}+\|\sigma(i_1)\|^{q_1 q}+\|z_0\|^q\right), \qquad \forall t.$$

Satisfaction of condition (e) now follows from the fact that

$$\|f(X_0)\| \geq \frac{1}{2N+K}\Big(\|\sigma(i_0)\|^{q_1}+\|\sigma(i_1)\|^{q_1}+\|z_0\|\Big).$$

As for condition (f), with $X = (i, j, z)$, we have

$$\begin{aligned}\|A(X)\| &= \|z(\alpha\phi'(j)-\phi'(i))\| \\ &\leq \|z\|(\alpha\|\phi'(j)\|+\|\phi'(i)\|),\end{aligned}$$

and the condition easily follows from Assumption 3(b). Similarly for $b(X)$ we have

$$\begin{aligned}\|b(X)\| &= zg(i,j) \\ &\leq \|z\|\cdot|g(i,j)|,\end{aligned}$$

and once again the condition easily follows from Assumption 3(b).

Finally, we show that condition (g) is satisfied. Recall that

$$A(X_t) = z_t(\alpha\phi'(i_{t+1})+\phi'(i_t)).$$

Let us concentrate on the term $z_t\phi'(i_t)$. Using the formula for z_t, we have

$$E[z_t\phi'(i_t)|i_0]-E_0[z_t\phi'(i_t)] = \sum_{m=0}^{t}(\alpha\lambda)^m E[\phi(i_{t-m})\phi'(i_t)|i_0]-\sum_{m=0}^{\infty}(\alpha\lambda)^m E_0[\phi(i_{t-m})\phi'(i_t)].$$

To condense notation, let us define $\Delta_{t-m,t}$ as follows:

$$\Delta_{t-m,t} = \Big\|E[\phi(i_{t-m})\phi'(i_t)|i_0]-E_0[\phi(i_{t-m})\phi'(i_t)]\Big\|.$$

Using the triangle inequality, we have

$$\sum_{t=0}^{\infty}\Big\|E[z_t\phi'(i_t)|i_0]-E_0[z_t\phi'(i_t)]\Big\| \leq \sum_{t=0}^{\infty}\sum_{m=0}^{t}(\alpha\lambda)^m\Delta_{t-m,t}+\sum_{t=0}^{\infty}\sum_{m=t+1}^{\infty}(\alpha\lambda)^m\|E_0[\phi(i_{t-m})\phi'(i_t)]\|.$$

We will individually bound the magnitude of each of the two summations in the right-hand-side expression above.

We can place a bound on the first summation as follows:

$$\begin{aligned}\sum_{t=0}^{\infty}\sum_{m=0}^{t}(\alpha\lambda)^m\Delta_{t-m,t} &\leq \sum_{m=0}^{\infty}(\alpha\lambda)^m\sum_{t=m}^{\infty}\Delta_{t-m,t}\\ &\leq \frac{C_2(1+\|\sigma(i_0)\|^{q_2})}{1-\alpha\lambda},\end{aligned}$$

where the final inequality follows from Assumption 3(c).

As for the second summation, we note that $E_0[\phi(i_{t-m})\phi'(i_t)] \leq G$, for some absolute constant G (Lemma 6) and

$$\begin{aligned}\sum_{t=0}^{\infty}\sum_{m=t+1}^{\infty}(\alpha\lambda)^m\|E_0[\phi(i_{t-m})\phi'(i_t)]\| &\leq G\sum_{t=0}^{\infty}\sum_{m=t+1}^{\infty}(\alpha\lambda)^m\\ &= G\sum_{t=0}^{\infty}\frac{(\alpha\lambda)^{t+1}}{1-\alpha\lambda}\\ &< \infty.\end{aligned}$$

It follows that there exists a positive constant C such that

$$\sum_{t=0}^{\infty}\left\|E[z_t\phi'(i_t)|i_0]-E_0[z_t\phi'(i_t)]\right\| \leq C(1+\|\sigma(i_0)\|^{q_2}) \leq C(1+\|f(X_0)\|^{q_2}).$$

An identical argument can be carried out for the terms $\alpha z_t\phi'(i_{t+1})$, and $z_t g(i_t,i_{t+1})$, which we omit to avoid repetition. Satisfaction of condition (g) follows.

We now have all the conditions needed to apply Theorem 2. It follows that r_t converges to r^*, which solves

$$\Phi D(T^{(\lambda)}(\Phi' r^*) - \Phi' r^*) = 0.$$

By Lemma 5 along with the fact that ΦD has full row rank (by virtue of Assumption 2(a)), r^* uniquely satisfies this equation and is the unique fixed point of $\Pi T^{(\lambda)}$. Lemma 5 also provides the desired error bound. This completes the proof to Theorem 1.

7 The Case of a Finite State Space

In this section, we show that Assumptions 1(b)-(c), 2(b), and 3 are automatically true whenever we are dealing with an ergodic Markov chain with a finite state space. This tremendously simplifies the conditions required to apply Theorem 1, reducing them to two: that the Markov chain be ergodic (Assumption 1(a)) and the basis functions be linearly independent (Assumption 2(a)). Actually, even these two assumptions can be relaxed if Theorem 1 is stated in a more general way. These two assumptions were adopted for the sake of simplicity in the proof.

Let us now assume that i_t is an aperiodic finite-state Markov chain, with a single ergodic class, and no transient states. Assumptions 1(b)-(c), 2(b), 3(a), and 3(b), are trivially satisfied when the state space is finite. We therefore only need to prove that Assumption 3(c) is satisfied.

It is well known that for any finite state ergodic Markov chain, there exist scalars $\rho < 1$ and C such that

$$|\Pr(i_t = i|i_0) - \pi(i)| \leq C\rho^t, \qquad \forall i_0 \in S.$$

Let us fix i_0. We define a sequence of $K \times K$ diagonal matrices D_t with the ith diagonal element equal to $\Pr(i_t = i | i_0)$. Note that

$$\|D_t - D\| \leq C\rho^t.$$

It is then easy to show that

$$E[\phi(i_t)\phi'(i_{t+m})|i_0] = \Phi D_t P^m \Phi',$$

the proof being essentially the same as in Lemma 6(a). We then have

$$E[\phi(i_t)\phi'(i_{t+m})|i_0] - E_0[\phi(i_t)\phi'(i_{t+m})] = \Phi(D_t - D)P^m\Phi'.$$

Note that all entries of P^m are bounded by 1 and therefore there exists a constant G such that $\|P^m\| \leq G$ for all m. We then have

$$\begin{aligned}
\sum_{t=0}^{\infty} \|\Phi(D_t - D)P^m\Phi'\| &\leq \sum_{t=0}^{\infty} K^2 \max_{k,j} |\phi_k(D_t - D)P^m\phi_j'| \\
&\leq K^2 \max_k \|\phi_k'\| G \max_j \|\phi_j'\| \sum_{t=0}^{\infty} \|D_t - D\| \\
&\leq GK^2 \max_k \|\phi_k'\|^2 \frac{C}{1-\rho}.
\end{aligned}$$

The first part of Assumption 3(c) follows. The second part is obtained using an analogous argument, which we omit.

8 Infinite State Spaces

The purpose of this section is to shed some light on the nature of our assumptions and to suggest that our results apply to infinite-state Markov chains of practical interest. For concreteness, let us assume that the state space is a subset of $\Re^N$. In terms of our notation, we have $\sigma(i) \in \Re^N$. (Strictly speaking, our results only apply if we have a countable subset of $\Re^N$, but the extension is immediate and we will be commenting on it.)

Let us first assume that the state space is a bounded subset of $\Re^N$ and that the mappings defined by $(\sigma(i), \sigma(j)) \mapsto g(i,j)$ and $\sigma(i) \mapsto \phi_k(i)$ are continuous functions on $\Re^N$. Then, Assumptions 1(b)-(c) and 2(b) are automatically valid, because continuous functions are bounded on bounded sets. The same is true regarding Assumptions 3(a)-(b). Assumption 3(c) basically refers to the speed with which the Markov chain reaches steady-state. Let $D_t(i_0)$ be a diagonal matrix whose ith entry is $\Pr(i_t = i|i_0)$. Then Assumption 3(c) is satisfied if we impose a condition of the form

$$\sum_{t=0}^{\infty} \|D_t(i_0) - D\| \leq C, \qquad \forall i_0,$$

for some finite constant C. In other words, we want the t-step transition probabilities to converge fast enough to the steady-state probabilities (for example, $\|D_t - D\|$ could drop at the rate of $1/t^2$). In addition, we need this convergence to be uniform in the initial state.

As a special case, suppose that the Markov chain has a distinguished state, say state 0, and that for some δ,

$$\Pr(i_{t+1} = 0 | i_t = i) \geq \delta \qquad \forall i.$$

Then, $D_t(i_0)$ converges to D exponentially fast, and uniformly in i_0, and Assumption 3 is satisfied.

Let us now consider the case where the state space is an unbounded subset of $\Re^N$. For many stochastic processes of practical interest (e.g., those that satisfy a large deviations principle), the tails of the probability distribution $\pi(\cdot)$ exhibit exponential decay; let us assume that this is the case. Then, as long as $g(\cdot,\cdot)$ and $\phi_k(\cdot)$ grow only polynomially (this is the content of Assumption 3(b)), Assumptions 1(b) and 2(b) are satisfied.

Assumption 3(a) is essentially a stability condition; it states that $\|\sigma(i_t)\|^q$ is not expected to grow too rapidly, and is satisfied by most stable Markov chains of practical interest. Note that by taking the steady-state limit we obtain $E_0[\|\sigma(i_t)\|^q] < \infty$ for all q, which in essence says that the tails of the steady-state distribution $\pi(\cdot)$ decay faster than any polynomial (e.g., exponentially).

Assumption 3(c) is again the most complex one. Recall that it deals with the speed of convergence of certain functions of the Markov chain to steady-state. Whether it is satisfied has to do with the interplay between the speed of convergence of $D_t(i_0)$ to D and the growth rate of the functions $\phi_k(\cdot)$ and $g(\cdot,\cdot)$. Note that the assumption allows the rate of convergence to get worse as $\sigma(i_0)$ increases; this is captured by the term $\|\sigma(i_0)\|^{q_2}$ in the right-hand side.

We close with a concrete illustration, related to queueing theory. Let i_t be a Markov chain that takes values in the nonnegative integers, and let its dynamics be

$$i_{t+1} = \max\{0,\ i_t + w_t - 1\},$$

where the w_t are independent identically distributed nonnegative integer random variables with a "nice" distribution; e.g., assume that the tail of the distribution of w_t asymptotically decays at an exponential rate. (This Markov chain corresponds to an M/G/1 queue which is observed at service completion times, with w_t being the number of new arrivals while serving a customer.) Assuming that $E[w_t] < 1$, this chain has a "downward drift," is "stable," and has a unique invariant distribution (Walrand, 1988). Furthermore, there exists some $\delta > 0$ such that $\pi(i) \leq e^{-i\delta}$, for i sufficiently large. Let $g(i,j) = i$, so that the cost function basically counts the number of customers in queue. Let us introduce the basis functions $\phi_k(i) = i^k$, $k = 0,1,2,3$. Then, Assumptions 1-2 are satisfied. Assumption 3(a) can be shown to be true by exploiting the downward drift property.

Let us now discuss Assumption 3(c). The key is again the speed of convergence of $D_t(i_0)$ to D. Starting from state i_0, with i_0 large, the Markov chain has a negative drift, and requires $O(i_0)$ steps to enter (with high probability) the vicinity of state 0 (Stamoulis and Tsitsiklis, 1990; Konstantopoulos and Baccelli, 1990). Once the vicinity of state 0 is reached, it quickly reaches steady-state. Thus, if we concentrate on $\phi_3(i) = i^3$, the difference $E[\phi(i_\tau)\phi'(i_{\tau+m})|i_0] - E_0[\phi(i_\tau)\phi'(i_{\tau+m})]$ is of the order of i_0^6 for $O(i_0)$ time steps and afterwards decays at a fast rate. This suggests that Assumption 3(c) is also satisfied, with $q_2 = 7$.

Our discussion in the preceding example was far from rigorous. Our objective was not so much to prove that our assumptions are satisfied by specific examples, but rather to demonstrate that their content is plausible. Furthermore, while the M/G/1 queue is too simple an example, we expect that stable queueing networks that have a downward drifting Lyapunov function, should also generically satisfy our assumptions.

9 The Importance of On-Line Sampling

In the introduction, we claimed that on-line sampling plays an instrumental role in ensuring convergence of TD(λ). In particular, when working with a simulation model, it is possible to define variants of TD(λ) that do not sample states with the frequencies natural to the Markov chain, and as a result, do not generally converge. Many papers, including (Boyan and Moore, 1995), (Tsitsiklis and Van Roy, 1994), and (Gordon, 1995), present such examples as counter-examples to TD(λ). In this section, we provide some insight into this issue through exploring the behavior of a variant of TD(0). More, generally, variants of TD(λ) can be defined in a similar manner, and the same issues arise in that context. We limit our discussion to TD(0) for ease of exposition.

We consider a variant of TD(0) where a states i_t are sampled independently from a distribution $q(\cdot)$ over S, and successor states j_t are generated by sampling according to $\Pr[j_t = j | i_t] = p_{i_t j}$. Each iteration of the algorithm takes on the form

$$r_{t+1} = r_t + \gamma_t \phi(i_t)(g(i_t, j_t) + \alpha\phi'(j_t)r_t - \phi'(i_t)r_t).$$

Let us refer to this algorithm as q–sampled TD(0). Note that this algorithm is closely related to the original TD(λ) algorithm as defined in Section 2. In particular, if i_t is generated by the Markov chain and $j_t = i_{t+1}$, we are back to the original algorithm. It is easy to show, using a subset of the arguments required to prove Theorem 1, that this algorithm converges when $q(i) = \pi(i)$ for all i, and Assumptions 1, 2, and 4, are satisfied. However, results can be very different when $q(\cdot)$ is arbitrary. This is captured by the following Theorem.

Theorem 3 *Let $q(\cdot)$ be a probability distribution over S, where $|S| = n$ is at least 2 (and at most countably infinite). Let the discount factor α be constrained to the open interval $(\frac{5}{6}, 1)$. Let the sequence γ_t satisfy Assumption 4. Then, there exists a stochastic matrix P, a transition cost function $g(\cdot,\cdot)$, and a matrix Φ, such that Assumptions 1 and 2 are satisfied and execution of the q-sampled TD(0) algorithm leads to*

$$\lim_{t\to\infty} \|E[r_t | r_0]\| = \infty, \qquad \forall r_0 \neq r^*,$$

for some unique vector r^*.

Proof: Without loss of generality, we will assume throughout this proof that $q(1) > 0$ and $q(1) \geq q(2)$.

We define a probability distribution $p(\cdot)$ satisfying $1 > p(2) > \frac{5}{6\alpha}$ and $p(i) > 0$ for all i. The fact that $\alpha > \frac{5}{6}$ ensures that such a probability distribution exists. We define the transition probability matrix P with each row equal to $p(\cdot)$. In other words, we have

$$P = \begin{bmatrix} p(1) & \cdots & p(n) \\ \vdots & \ddots & \vdots \\ p(1) & \cdots & p(n) \end{bmatrix}.$$

Finally, we define the transition cost function to be $g(i, j) = 0$, for all i and j. Assumption 1 is trivially satisfied by our choice of P and $g(\cdot,\cdot)$, and the invariant distribution of the Markov chain is $p(\cdot)$. Note that $J^* = 0$, since no transition incurs any cost.

Let Φ be a $1 \times n$ matrix, defined by a single scalar function $\phi(\cdot)$ with

$$\phi(i) = \begin{cases} 1, & \text{if } i = 1, \\ 2, & \text{if } i = 2, \\ 0, & \text{otherwise.} \end{cases}$$

Note that, implicit from our definition of Φ, r_t is scalar, and Assumption 2 is trivially satisfied. We let $r^* = 0$, so that $J^* = \Phi' r^*$.

In general, we can express $E[r_t|r_0]$ in terms of a recurrence of the form

$$\begin{aligned} E[r_{t+1}|r_0] &= E[r_t|r_0] + \gamma_t E\Big[\phi(i_t)(g(i_t, j_t) + \alpha\phi'(j_{t+1})r_t - \phi'(i_t)r_t)|r_0\Big] \\ &= E[r_t|r_0] + \gamma_t \Phi Q(\bar{g} + \alpha P\Phi' - \Phi')E[r_t|r_0], \end{aligned}$$

where Q is the diagonal matrix with diagonal elements $q(1), \ldots q(n)$.

Specializing to our choice of parameters, the recurrence becomes

$$\begin{aligned} E[r_{t+1}|r_0] &= E[r_t|r_0] + \gamma_t\Big[q(1) \;\; 2q(2)\Big]\left(\alpha\begin{bmatrix} p(1) + 2p(2) \\ p(1) + 2p(2) \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \end{bmatrix}\right)E[r_t|r_0] \\ &= E[r_t|r_0] + \gamma_t\Big((\alpha(p(1) + 2p(2)) - 1)q(1) + 2(\alpha(p(1) + 2p(2)) - 2)q(2)\Big)E[r_t|r_0]. \end{aligned}$$

For shorthand notation, let Δ_t be defined by

$$\Delta_t = (\alpha p(1) + 2\alpha p(2) - 1)q(1) + 2(\alpha p(1) + 2\alpha p(2) - 2)q(2).$$

Since $\alpha p(1) + 2\alpha p(2) < 2$ and $q(1) \geq q(2)$, we have

$$\begin{aligned} \Delta_t &\geq (\alpha p(1) + 2\alpha p(2) - 1)q(1) + 2(\alpha p(1) + 2\alpha p(2) - 2)q(1) \\ &= (3\alpha p(1) + 6\alpha p(2) - 5)q(1) \\ &\geq (6\alpha p(2) - 5)q(1), \end{aligned}$$

and since $p(2) > \frac{5}{6\alpha}$, there exists some $\epsilon > 0$ such that

$$\begin{aligned} \Delta_t &\geq (5 + \epsilon - 5)q(1) \\ &= \epsilon q(1). \end{aligned}$$

It follows that

$$||E[r_{t+1}|r_0]|| \geq (1 + \gamma_t \epsilon q(1))||E[r_t|r_0]||,$$

and since $\sum_{t=0}^{\infty} \gamma_t = \infty$, we have

$$\lim_{t \to \infty} ||E[r_{t+1}|r_0]|| = \infty,$$

if $r_0 \neq r^*$. **q.e.d.**

10 Conclusions

We have established the convergence of on-line temporal difference learning with linearly parameterized function approximators, when applied to infinite-horizon discounted Markov chains. We note that this result is new even for the case of lookup table representations (i.e., when there is no function approximation), but its scope is much greater. Furthermore, in addition to covering the case where the underlying Markov chain is finite, the result also applies to Markov chains over a general (infinite) state space, as long as certain technical conditions are satisfied.

The key to our development was the introduction of the norm $\| \cdot \|_D$ and the property $\|P\|_D \leq 1$. Furthermore, our development indicates that the progress of the algorithm can be monitored in two different ways: (a) we can keep track of the magnitude of the approximation error $\Phi' r^* - J^*$; the natural norm for doing so is precisely the norm $\| \cdot \|_D$; or, (b) we can keep track of the parameter error $r - r^*$; the natural norm here is the Euclidean norm, as made clear by our convergence proof.

To reinforce the central ideas in the proof, let us revisit the TD(0) method, for the case where the costs per stage are identically zero. In this case, $T^{(0)}J$ is simply $\alpha PJ - J$. The deterministic counterpart of the algorithm, as introduced in Section 3 takes the form

$$\begin{aligned} \bar{r}_{t+1} &= \bar{r}_t + \gamma_t \Phi D(\alpha P \Phi' r - \Phi' r) \\ &= \bar{r}_t + \gamma_t \Phi D(\alpha P - I)\Phi' r \end{aligned}$$

For any vector J, we have

$$J'DPJ \leq \|J\|_D \cdot \|PJ\|_D \leq \|J\|_D^2 = J'DJ.$$

This shows that the matrix $D(\alpha P - I)$ is negative definite, hence $\Phi D(\alpha P - I)\Phi'$ is also negative definite, and convergence of this deterministic iteration follows.

Besides convergence, we have also provided bounds on the distance of the limiting function $\Phi' r^*$ from the true cost-to-go function J^*. These bounds involve the expression $\|\Pi J^* - J^*\|_D$, which is natural because no approximation could have error smaller than this expression (when the error is measured in terms of $\| \cdot \|_D$). What is interesting is the term

$$1 - \frac{\alpha(1-\lambda)}{1-\alpha\lambda}$$

in the denominator. This term is 1 when $\lambda = 1$. For every $\lambda < 1$, the denominator term is smaller than 1, and the bound actually deteriorates as λ decreases. The worst bound, namely $\|\Pi J^* - J^*\|_D/(1-\alpha)$ is obtained when $\lambda = 0$. Although this is only a bound, it strongly suggests that higher values of λ are likely to produce more accurate approximations of J^*. This is consistent with the examples that have been constructed by Bertsekas (1994).

The sensitivity of the error bound to λ raises the question of whether or not it ever makes sense to set λ to values less than 1. Experimental results (Sutton, 1988; Singh and Sutton, 1994, Sutton, 1995) suggest that setting λ to values less than one can often lead to significant gains in the rate of convergence. Such acceleration may be critical when computation time and/or data (in the event that the trajectories are generated by a physical system) are limited. A full understanding of how λ influences the rate of convergence is yet to be found. Furthermore, it might be desirable to tune λ as the algorithm progresses, possibly initially starting with $\lambda = 0$ and approaching $\lambda = 1$ (although the opposite has also been advocated). These are interesting directions for future research.

In many applications of temporal difference methods, one deals with a controlled Markov chain and at each stage a decision is "greedily" chosen, by minimizing the right-hand side of Bellman's equation, and using the available approximation $\tilde{J}$ in place of J^*. Our analysis does not apply to such cases involving changing policies. Of course, if the policy eventually settles into a limiting policy, we are back to the case studied in this paper and convergence is obtained. However, there exist examples for which the policy does not converge (Bertsekas and Tsitsiklis, 1996). It remains an open problem to analyze the limiting behavior of the parameters r and the resulting approximations $\Phi' r$ for the case where the policy does not converge.

On the technical side, we mention a few straightforward extensions of our results. First, the linear independence of the basis functions ϕ_k is not essential. In the linearly dependent case, some components of z_t and r_t become linear combinations of the other components and can be simply eliminated, which takes us back to the linearly independent case. A second extension is to allow the cost per stage $g(i_t, i_{t+1})$ to be noisy, as opposed to being a deterministic function of i_t and i_{t+1}. As long as the distribution of the noise only depends on the current state, and its moments are bounded by a constant independent of the state, there is no difficulty. (It is also possible to let the moments of the noise depend on the current state, as long as they do not grow too fast.)

Finally, our results in Section 9 have elucidated the importance of sampling states according to the steady-state distribution of the Markov chain under consideration. In particular, variants of TD(λ) that samples states otherwise can lead to divergence when function approximators are employed. As a parting note, we point out that a related issue arises when one "plays" with the evolution equation for the eligibility vector z_t. (For example Singh and Sutton (1994) have suggested an alternative evolution equation for z_t known as the "replace trace.") A very general class of such mechanisms can be shown to lead to convergent algorithms for the case of lookup table representations (Bertsekas and Tsitsiklis, 1996). However, different mechanisms for adjusting the coefficients z_t lead to a change in the steady-state average value of $z_t \phi'(i_t)$ and affect the matrix A, and the negative definiteness property can be easily lost.

Acknowledgements

We would like to thank Richard Sutton for starting us on the path that lead to this work by pointing out that the counter-example in (Tsitsiklis and Van Roy, 1994) would no longer be a counter-example if on-line state sampling was used.

References

Bertsekas, D. P. (1994) "A Counter-Example to Temporal-Difference Learning," Neural Computation, vol. 7, pp. 270-279.

Bertsekas, D. P. (1995) *Dynamic Programming and Optimal Control*, Athena Scientific, Belmont, MA.

Bertsekas, D. P. & Tsitsiklis, J. N. (1996) *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, (forthcoming).

Boyan, J. A. & Moore, A. W. (1995) "Generalization in Reinforcement Learning: Safely

Approximating the Value Function," in J. D. Cowan, G. Tesauro, and D. Touretzky, editors, *Advances in Neural Information Processing Systems 7*, Morgan Kaufmann.

Dayan, P. D. (1992) "The Convergence of TD(λ) for General λ," Machine Learning, vol. 8, pp. 341-362.

Dayan, P. D. & Sejnowski, T. J. (1994) "TD(λ) Converges with Probability 1," Machine Learning, vol. 14, pp. 295-301.

Gordon, G. J. (1995) "Stable Function Approximation in Dynamic Programming," Technical Report: CMU-CS-95-103, Carnegie Mellon University.

Gurvits, L., Lin, L. J., & Hanson, S. J. (1995) "Incremental Learning of Evaluation Functions for Absorbing Markov Chains: New Methods and Theorems," preprint.

Jaakola, T., Jordan M. I., & Singh, S. P. (1994) "On the Convergence of Stochastic Iterative Dynamic Programming Algorithms," Neural Computation, vol. 6, No. 6.

Konstantopoulos, P. & Baccelli, F. (1990) "On the Cut-Off Phenomena in Some Queueing Systems," preprint.

Singh, S. P. & Sutton, R. S. (1994) "Reinforcement Learning with Replacing Eligibility Traces," to appear in Machine Learning.

Sutton, R. S., (1988) "Learning to Predict by the Method of Temporal Differences," Machine Learning, vol. 3, pp. 9-44.

Sutton, R. S. (1995) "Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding," to Appear in *Advances in Neural Information Processing Systems 8.*

Stamoulis, G. D. & Tsitsiklis, J. N. (1990) "On the Settling Time of the Congested $GI/G/1$ Queue," Advances in Applied Probability, vol. 22, PP. 929-956.

Tesauro, G., (1992) "Practical Issues in Temporal Difference Learning," Machine Learning, vol. 8, pp. 257-277.

Tsitsiklis, J. N. (1994) "Asynchronous Stochastic Approximation and Q-Learning," Machine Learning, vol. 16, pp. 185-202.

Tsitsiklis, J. N. & Van Roy, B. (1994) "Feature–Based Methods for Large Scale Dynamic Programming," Technical Report: LIDS-P-2277, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology. Also to appear in *Machine Learning.*

Warland, J. (1988) *An Introduction to Queueing Networks*, Prentice Hall, Englewood Cliffs, NJ.

Watkins, C. J. C. H. & Dayan, P. (1992) "Q–learning," Machine Learning, vol. 8, pp. 279-292.