An Analysis of Temporal-Difference Learning with Function Approximation

John N. Tsitsiklis, *Member, IEEE*, and Benjamin Van Roy

***Abstract*— We discuss the temporal-difference learning algorithm, as applied to approximating the cost-to-go function of an infinite-horizon discounted Markov chain. The algorithm we analyze updates parameters of a linear function approximator on-line during a single endless trajectory of an irreducible aperiodic Markov chain with a finite or infinite state space. We present a proof of convergence (with probability one), a characterization of the limit of convergence, and a bound on the resulting approximation error. Furthermore, our analysis is based on a new line of reasoning that provides new intuition about the dynamics of temporal-difference learning.**

In addition to proving new and stronger positive results than those previously available, we identify the significance of on-line updating and potential hazards associated with the use of nonlinear function approximators. First, we prove that divergence may occur when updates are not based on trajectories of the Markov chain. This fact reconciles positive and negative results that have been discussed in the literature, regarding the soundness of temporal-difference learning. Second, we present an example illustrating the possibility of divergence when temporal-difference learning is used in the presence of a nonlinear function approximator.

***Index Terms*— Dynamic programming, function approximation, Markov chains, neuro-dynamic programming, reinforcement learning, temporal-difference learning.**

I. 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 [1].

Temporal-difference learning, originally proposed by Sutton [2], is a method for approximating long-term future cost as a function of current state. The algorithm is recursive, efficient, and simple to implement. A function approximator is used to approximate the mapping from state to future cost. Parameters of the function approximator 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 [2]–[7], 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 [8]–[12]. The latter three establish convergence with probability one. 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 [8] establishes convergence in the mean for the general class of linear function approximators, i.e., function approximators involving linear combinations of fixed basis functions, where the weights of the basis functions are tunable parameters. 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. Schapire and Warmuth [9] carry out a (nonprobabilistic) worst case analysis of an algorithm similar to temporal-difference learning. Fewer assumptions are required by their analysis, but the end results do not imply convergence and establish error bounds that are weak relative to those that can be deduced in the standard probabilistic framework.

In addition to the positive results, counterexamples to variants of the algorithm have been provided in several papers; these include [10], [11], [13], and [14]. As suggested by Sutton [15], 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 counterexamples by showing that for linear function approximators, convergence is guaranteed if states are sampled according

Manuscript received March 20, 1996; revised November 11, 1996. Recommended by Associate Editor, E. K. P. Chong. This work was supported by the NSF under Grant DMI-9625489 and the ARO under Grant DAAL-03-92-G-0115.

The authors are with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: jnt@mit.edu).

Publisher Item Identifier S 0018-9286(97)03437-5.

to the steady-state probabilities, while divergence is possible when states are sampled from distributions independent of the dynamics of the Markov chain of 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 addition to the analysis of temporal-difference learning in conjunction with linear function approximators, we provide an example demonstrating that the algorithm may diverge when a nonlinear function approximator is employed. This example should be viewed as a warning rather than a ban on all nonlinear function approximators. In particular, the function approximator used in the example is somewhat contrived, and it is not clear whether or not divergence can occur with specific classes of nonlinear function approximators such as neural networks.

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 absorbing (and typically finite state) 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 be applied to the simpler context of absorbing Markov chains. Though this extension is omitted from this paper, it can be found in [16], which also contains a more accessible version of the results in this paper for the case of finite state spaces.

The contributions in this paper are as follows.

1) Convergence (with probability one) 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 TD(λ) 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) 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.
4) We reconcile positive and negative results concerning temporal-difference learning by proving a theorem that identifies the importance of on-line sampling.
5) We provide an example demonstrating the possibility of divergence when temporal-difference learning is used in conjunction with a nonlinear function approximator.

At about the same time that this paper was initially submitted, Gurvits [17] independently established convergence with probability one in the context of absorbing Markov chains. Also, Pineda [18] derived a stable differential equation for the "mean field" of temporal-difference learning, in the case of finite-state absorbing Markov chains. He also suggested a convergence proof based on a weighted maximum norm contraction property, which, however, is not satisfied in the presence of function approximation. (The proof was corrected after the paper became available.)

This paper is organized as follows. In Section II, we provide a precise definition of the algorithm that we will be studying. Sections III–IX deal only with the use of linear function approximators. In Section III, we recast temporal-difference learning in a way that sheds light into its mathematical structure. Section IV contains our main convergence result together with our assumptions. We develop some mathematical machinery in Section V, which captures the fundamental ideas involved in the analysis. Section VI presents a proof of the convergence result, which consists primarily of the technicalities required to integrate the machinery supplied by Section V. Our analysis is valid for general state spaces, subject to certain technical assumptions. In Section VII, we show that these technical assumptions are automatically valid for the case of irreducible aperiodic finite-state Markov chains. In Section VIII, we argue that the class of infinite-state Markov chains that satisfy our assumptions is broad enough to be of practical interest. Section IX contains our converse convergence result, which establishes the importance of on-line sampling. Section X departs from the setting of linear function approximators, presenting a divergent example involving a nonlinear function approximator. Finally, Section XI contains some concluding remarks.

II. 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 irreducible aperiodic Markov chain whose states lie in a finite or countably infinite space S. By indexing the states with positive integers, we can view the state space as the set $S = \{1, \cdots, n\}$, where n is possibly infinite. Note that the positive integers only serve as indexes here. In particular, each state might actually correspond to some other entity such as a vector of real numbers describing the state of a physical system. In such a case, the actual state space would comprise of a countable subset of a Euclidean space.

The sequence of states visited by the Markov chain is denoted by $\{i_t \mid t = 0, 1, \cdots\}$. The dynamics of the Markov chain are 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

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

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

We consider approximations of J^*: $S \mapsto \Re$ using a function $\tilde{J}$: $S \times \Re^K \mapsto \Re$, which we refer to as a function approximator. To approximate the cost-to-go function one usually tries to choose a parameter vector $r \in \Re^K$ so as to minimize some error metric between the functions $\tilde{J}(\cdot, r)$ and $J^*(\cdot)$.

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

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

where r_0 is initialized to some arbitrary vector, γ_t is a sequence of scalar step sizes, λ is a parameter in $[0, 1]$, and the gradient $\nabla \tilde{J}(i, r)$ is the vector of partial derivatives with respect to the components of r. Since temporal-difference learning is actually a continuum of algorithms, parameterized by λ, it is often referred to as TD(λ).

In the special case of linear function approximators, the function $\tilde{J}$ takes the form

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

Here, $r = (r(1), \cdots, r(K))$ is the 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.

It is convenient to define a vector-valued function ϕ: $S \mapsto \Re^K$ by letting $\phi'(i) = (\phi_1(i), \cdots, \phi_K(i))$. 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 an $|S| \times K$ matrix whose kth column is equal to ϕ_k; that is

$$\Phi = \begin{bmatrix} | & & | \\ \phi_1 & \cdots & \phi_K \\ | & & | \end{bmatrix} = \begin{bmatrix} - & \phi'(1) & - \\ & \vdots & \\ - & \phi'(n) & - \end{bmatrix}.$$

Note that the gradient vector here is given by

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

and we have

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

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

In the case of linear function approximators, a more convenient representation of TD(λ) is obtained by defining a sequence of *eligibility vectors* z_t (of dimension K) by

$$z_t = \sum_{k=0}^{t} (\alpha\lambda)^{t-k} \nabla \tilde{J}(i_k, r_t) = \sum_{k=0}^{t} (\alpha\lambda)^{t-k} \phi(i_k).$$

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 $z_{-1} = 0$.

In the next few sections, we focus on temporal-difference learning as used with linear function approximators. Only in Section X do we return to the more general context of nonlinear function approximators.

III. 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, in the context of linear function approximators. Our goal 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.

A. Inner Product Space Concepts and Notation

We begin by introducing some notation that will make our discussion here, as well as the analysis later in the paper, more concise. Let $\pi(1), \cdots, \pi(n)$ 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 $\pi(1), \cdots, \pi(n)$. It is easy to see that $\langle x, y \rangle_D \triangleq x'Dy$ satisfies the requirements for an inner product. We denote the norm on the associated 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 will assume that each basis function ϕ_k is an element of $L_2(S, D)$ so that $\{\Phi r | r \in \Re^K\} \subset L_2(S, D)$. For any pair of functions $J, \overline{J} \in L_2(S, D)$, we say that J is D-orthogonal to $\overline{J}$ (denoted by $J \perp_D \overline{J}$) if and only if $\langle J, \overline{J} \rangle_D = 0$. For any $J \in L_2(S, D)$, there exists a unique element $\overline{J} \in \{\Phi r^* | r \in \Re^K\}$ minimizing $\|J - \overline{J}\|_D$ over $\{\Phi r^* | r \in \Re^K\}$. This $\overline{J}$ is referred to as the *projection* of J on $\{\Phi r | r \in \Re^K\}$ with respect to $\langle \cdot, \cdot \rangle_D$. We define a "projection matrix" (more precisely, projection operator) Π that generates such a $\overline{J}$ when applied to J. Assuming that the basis functions $\phi_1, \cdots, \phi_K$ 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.) For any $J \in L_2(S, D)$, we then have

$$\Pi J = \arg \min_{\overline{J} \in \{\Phi r | r \in \Re^K\}} \|J - \overline{J}\|_D.$$

Furthermore, $\overline{J} = \Pi J$ is the unique element of $\{\Phi r^* | r \in \Re^K\}$ such that $(J - \overline{J}) \perp_D \phi_k$ for all $k \in \{1, \cdots, K\}$. In other words, the difference between J and $\overline{J}$ is D-orthogonal to the space spanned by the basis functions.

The projection ΠJ^* is a natural approximation to J^*, given the fixed set of basis functions. In particular, 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 were continuous instead of countable, this sum would be replaced by an integral.)

B. The TD(λ) Operator

To streamline our analysis of TD(λ) we introduce an operator that is useful in characterizing the algorithm's dynamics. This operator, which we will refer to as the TD(λ) operator, is indexed by a parameter $\lambda \in [0, 1]$ and is denoted by $T^{(\lambda)}$: $L_2(S, D) \mapsto L_2(S, D)$. It is defined by

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

for $\lambda < 1$, and

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

for $\lambda = 1$, so that $\lim_{\lambda \uparrow 1} (T^{(\lambda)} J)(i) = (T^{(1)} J)(i)$ (under some technical conditions). The fact that $T^{(\lambda)}$ maps $L_2(S, D)$ into $L_2(S, D)$ will be established in a later section. To interpret the TD(λ) 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 $\| \cdot \|_D$.

C. Dynamics of the Algorithm

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:

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

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

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

and thus the deterministic algorithm takes the form

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

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 feasible. 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 $\overline{r}_t$, we rewrite the deterministic algorithm in the form

$$\overline{r}_{t+1} = \overline{r}_t + \gamma_t \nabla \tilde{J}(\overline{r}_t) D(T^{(\lambda)}(\Phi \overline{r}_t) - \Phi \overline{r}_t). \tag{2}$$

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

$$\overline{r}_{t+1} = \overline{r}_t - \frac{\gamma_t}{2} \nabla \| J^* - \Phi \overline{r}_t \|_D^2$$

which is a steepest descent iteration for the problem of minimizing

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

with respect to r. It is easy to show that if the step sizes are appropriately chosen, $\Phi \overline{r}_t$ 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

$$\sum_{i \in S} \pi(i)((T^{(\lambda)}(\Phi \overline{r}_t))(i) - \tilde{J}(r, i))^2$$

with respect to r, given a fixed $\overline{r}_t$. Note, however, that the error function changes from one time step to the next, and therefore it is not a true steepest descent method. Nevertheless, if we think of $T^{(\lambda)}(\Phi \overline{r}_t)$ 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 $q(1), \cdots, q(n)$. If we replace the matrix D in the deterministic variant of TD(1) with the matrix Q, we obtain

$$\overline{r}_{t+1} = \overline{r}_t - \frac{\gamma_t}{2} \nabla \| J^* - \Phi \overline{r}_t \|_Q^2$$

which is a steepest descent method that minimizes

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

with respect to r. If step sizes are appropriately chosen, $\Phi \overline{r}_t$ 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 IX.

To get a better grasp on the issues involved here, let us consider the following variant of the algorithm:

$$\overline{r}_{t+1} = \overline{r}_t + \gamma_t \nabla \tilde{J}(\overline{r}_t) Q(T^{(\lambda)}(\Phi \overline{r}_t) - \Phi \overline{r}_t). \tag{3}$$

Note that by letting $Q = D$, we recover the deterministic variant of TD(λ). Each iteration given by (3) can be thought of as a steepest descent iteration on an error function given by

$$\sum_{i \in S} q(i)((T^{(\lambda)}(\Phi \overline{r}_t))(i) - \tilde{J}(r, i))^2.$$

(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 $\Pi_Q T^{(\lambda)}(\Phi \overline{r}_t)$. Hence, letting $J_t = \Phi \overline{r}_t$, we might think of $\Pi_Q T^{(\lambda)}(J_t)$ as a "target vector," given a current vector J_t. We can define an algorithm of the form

$$J_{t+1} = \Pi_Q T^{(\lambda)}(J_t) \tag{4}$$

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

Intuitively, the iteration of (3) can be thought of as an incremental form of (4). Hence, one might expect the two algorithms to have similar convergence properties. In fact, they do. Concerning convergence of the algorithm given by (4), note that if $T^{(\lambda)}$ is a contraction of the norm $\| \cdot \|_Q$, then the composition $\Pi_Q T^{(\lambda)}(\cdot)$ 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, $\Pi_Q T^{(\lambda)}(\cdot)$ may not be a contraction and might even be an expansion. Hence, convergence guarantees for the algorithms of (3) and (4) rely on a relationship between $T^{(\lambda)}$ and Π. This idea captures 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 (2), can be rewritten in the form

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

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.

IV. 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, when linear function approximators are employed. 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 irreducible and aperiodic and that the steady-state variance of transition costs is finite. The formal statement follows.

Assumption 1:

1) The Markov chain i_t is irreducible and aperiodic. Furthermore, there is a unique 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.
2) Transition costs $g(i_t, i_{t+1})$ satisfy
$$E_0[g^2(i_t, i_{t+1})] < \infty.$$

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

Assumption 2:

1) The matrix Φ has full column rank; that is, the basis functions $\{\phi_k \mid k = 1, \cdots, K\}$ are linearly independent.
2) For every k, the basis function ϕ_k satisfies
$$E_0[\phi_k^2(i_t)] < \infty.$$

The next assumption essentially requires that the Markov chain has a certain "degree of stability." As will be shown in Section VI, 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 can be found in Section VII.

Assumption 3: There exists a function $f\colon S \mapsto \Re^+$ (the range is the set of nonnegative reals) satisfying the following requirements.

1) For all i_0 and $m \geq 0$
$$\sum_{\tau=0}^{\infty} ||E[\phi(i_\tau)\phi'(i_{\tau+m})|i_0] - E_0[\phi(i_t)\phi'(i_{t+m})]|| \leq f(i_0)$$
and
$$\sum_{\tau=0}^{\infty} ||E[\phi(i_\tau)g(i_{\tau+m}, i_{\tau+m+1})|i_0] - E_0[\phi(i_t)g(i_{t+m}, i_{t+m+1})]|| \leq f(i_0).$$
2) Any $q > 1$, there exists a constant μ_q such that for all i_0, t
$$E[f^q(i_t)|i_0] \leq \mu_q f^q(i_0).$$

Implicit in the statement of this assumption is that certain expectations are finite. It will be seen later that their finiteness is a consequence of earlier assumptions.

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

Assumption 4: The step sizes γ_t are positive, 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 1–4, the following hold.

1) The cost-to-go function J^* is in $L_2(S, D)$.
2) For any $\lambda \in [0, 1]$, the TD(λ) algorithm with linear function approximators, as defined in Section II, converges with probability one.
3) The limit of convergence r^* is the unique solution of the equation
$$\Pi T^{(\lambda)}(\Phi r^*) = \Phi r^*.$$
4) Furthermore, r^* satisfies[1]
$$||\Phi r^* - J^*||_D \leq \frac{1 - \lambda\alpha}{1 - \alpha} ||\Pi J^* - J^*||_D.$$

In order to place Theorem 1 in perspective, let us discuss its relation 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 [5] (actually, that paper dealt with the on-line TD(λ) algorithm only for Markov chains involving a termination state). With a lookup table representation, the operator $T^{(\lambda)}$ is easily shown to be a maximum norm contraction, the projection operator Π is simply the identity matrix, and general results on stochastic approximation methods based on maximum norm contractions [4], [5] become applicable. However, once function approximation is introduced, the composition $\Pi T^{(\lambda)}$ need not be a maximum norm contraction, and this approach does not extend.

Closer to our results is the work of Dayan [8] who considered TD(λ) for the case of linear function approximators and established a weak form of convergence (convergence

[1] It has been brought to our attention by V. Papavassiliou that this bound can be improved to
$$\|\Phi r^* - J^*\|_D \leq \frac{1 - \lambda\alpha}{\sqrt{(1 - \alpha)(1 + \alpha - 2\lambda\alpha)}} \|\Pi J^* - J^*\|_D.$$

in the mean). Finally, the work of Dayan and Sejnowski [6] contains a sketch of a proof of convergence with probability one. 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 [7].) Some of the ideas in our method of proof originate in the work of Sutton [2] and Dayan [8]. 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.

V. Preliminaries

In this section we present a series of lemmas that provide the essential ideas behind Theorem 1. Lemma 1 states a general property of Markov chains that is central to the analysis of TD(λ). Lemma 2 ensures that our assumptions are sufficient to have a well-defined cost-to-go function J^* in $L_2(S, D)$. Lemmas 3–6 deal with properties of the TD(λ) operator $T^{(\lambda)}$ and the composition $\Pi T^{(\lambda)}$, as well as their fixed points. Lemma 7 characterizes the steady-state expectations of various variables involved in the dynamics of TD(λ), and these results are used in the proof of Lemma 8, which deals with the steady-state dynamics. Lemma 9 establishes that these dynamics lead to convergence of the deterministic version of the algorithm. Finally, we state a theorem concerning stochastic approximation that will be used in Section VI, along with the lemmas, to establish convergence of the stochastic algorithm.

We begin with the fundamental lemma on Markov chains.

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

Proof: The proof involves Jensen's inequality, the Tonelli–Fubini theorem, and the property $\pi' P = \pi'$

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

■

Our first use of Lemma 1 will be in showing that J^* is in $L_2(S, D)$. In particular, we have the following result, where we use the notation $\overline{g}$ 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 1-1) and 2), $J^*(i)$ is well defined and finite for every $i \in S$. Furthermore, J^* is in $L_2(S, D)$, and

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

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

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

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

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

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

Using the Tonelli–Fubini theorem to switch the order of expectation and summation in the definition of J^*, we obtain

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

and it follows that

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

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

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

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

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

by Assumption 1-2). It follows that J^* is in $L_2(S, D)$. ■

The next lemma states 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 $J \in L_2(S, D)$ and $\lambda \in [0, 1]$, $T^{(\lambda)}(J)$ is in $L_2(S, D)$, and for $\lambda \in [0, 1)$, we have

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

Proof: We have

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

and the formula in the statement of the lemma follows.

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

$$\begin{aligned}&\left\|(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\sum_{t=0}^{m}(\alpha P)^t\overline{g}\right\|_D\\&\quad\le(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\sum_{t=0}^{m}\alpha^t\|\overline{g}\|_D\\&\quad<\infty.\end{aligned}$$

Similarly

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

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

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 1-1), for any $J, \overline{J} \in L_2(S, D)$ and $\lambda \in [0, 1]$, we have

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

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

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

■

The next lemma states that J^* is the unique fixed point of $T^{(\lambda)}$.

Lemma 5: Under Assumption 1, for any $\lambda \in [0, 1]$, the cost-to-go function J^* uniquely solves the system of equations given by

$$J^* = T^{(\lambda)}J^*.$$

Proof: For the case of $\lambda = 1$, the result follows directly from the definition of $T^{(\lambda)}$. For $\lambda \in [0, 1)$, the fact that J^* is a fixed point follows from Lemmas 2 and 3, the Tonelli–Fubini theorem, and some simple algebra:

$$\begin{aligned}T^{(\lambda)}J^*&=(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\left(\sum_{t=0}^{m}(\alpha P)^t\overline{g}+(\alpha P)^{m+1}J^*\right)\\&=(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\left(\sum_{t=0}^{m}(\alpha P)^t\overline{g}\right.\\&\qquad\left.+(\alpha P)^{m+1}\sum_{t=0}^{\infty}(\alpha P)^t\overline{g}\right)\\&=(1-\lambda)\sum_{m=0}^{\infty}\lambda^m\left(\sum_{t=0}^{\infty}(\alpha P)^t\overline{g}\right)\\&=(1-\lambda)\sum_{m=0}^{\infty}\lambda^m J^*.\end{aligned}$$

The contraction property (Lemma 4) implies that the fixed point is unique. ■

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 2-2)]. The lemma establishes existence and uniqueness of this fixed point, which we will denote by Φr^*. Note that in the special case of $\lambda = 1$ the lemma implies that $\Phi r^* = \Pi J^*$, in agreement with the definition of $T^{(1)}$.

Lemma 6: Under Assumptions 1 and 2, $\Pi T^{(\lambda)}(\cdot)$ is a contraction and has a unique fixed point which is of the form Φr^* for a unique choice of r^*. Furthermore, r^* satisfies the following bound:

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

Proof: Lemma 4 ensures that $T^{(\lambda)}$ is a contraction from $L_2(S, D)$ into itself, and from J^* is its fixed point by Lemma 5. Note that for $J \in L_2(S, D)$, by the Babylonian–Pythagorean theorem we have

$$\|J\|_D^2 = \|\Pi J\|_D^2 + \|J - \Pi J\|_D^2$$

since $\Pi J \perp_D (J - \Pi J)$. It follows that Π is nonexpansive, and thus the composition $\Pi T^{(\lambda)}(\cdot)$ is a contraction. Hence, $\Pi T^{(\lambda)}(\cdot)$ has a unique fixed point of the form Φr^*, for some r^*. Because the functions $\phi_k(\cdot)$ are assumed to be linearly independent, it follows that the choice of r^* is unique.

Using the fact that J^* is in $L_2(S, D)$ (Lemma 2) and is the fixed point of $T^{(\lambda)}$ (Lemma 5), we establish the desired

bound. In particular, we have

$$\begin{aligned}\|\Phi r^* - J^*\|_D &\le \|\Phi r^* - \Pi J^*\|_D + \|\Pi J^* - J^*\|_D \\ &= \|\Pi T^{(\lambda)}(\Phi r^*) - \Pi J^*\|_D + \|\Pi J^* - J^*\|_D \\ &\le \|T^{(\lambda)}(\Phi r^*) - J^*\|_D + \|\Pi J^* - J^*\|_D \\ &\le \frac{\alpha(1-\lambda)}{1-\alpha\lambda}\|\Phi r^* - J^*\|_D + \|\Pi J^* - J^*\|_D\end{aligned}$$

and it follows that

$$\begin{aligned}\|\Phi r^* - J^*\|_D &\le \frac{\|\Pi J^* - J^*\|_D}{1-\alpha(1-\lambda)/(1-\alpha\lambda)} \\ &= \frac{1-\lambda\alpha}{1-\alpha}\|\Pi J^* - J^*\|_D.\end{aligned}$$

■

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 is 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 one 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 III (the evolution equation is the same). The only difference is that the process X_t of Section III was initialized with $z_{-1} = 0$, whereas here we have a stationary process X_t. 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 7: Under Assumptions 1 and 2, the following relations hold.

1) $E_0[\phi(i_t)\phi'(i_{t+m})] = \Phi' D P^m \Phi$, for $m \ge 0$.
2) There exists a finite constant G such that $\|E_0[\phi(i_t)\phi'(i_{t+m})]\| \le G$, for all $m \ge 0$.
3) $E_0[z_t\phi'(i_t)] = \sum_{m=0}^{\infty}(\alpha\lambda)^m \Phi' D P^m \Phi$.
4) $E_0[z_t\phi'(i_{t+1})] = \sum_{m=0}^{\infty}(\alpha\lambda)^m \Phi' D P^{m+1} \Phi$.
5) $E_0[z_t g(i_t, i_{t+1})] = \sum_{m=0}^{\infty}(\alpha\lambda)^m \Phi' D P^m \overline{g}$.

Furthermore, each of the above expressions is well defined and finite.

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

$$\begin{aligned}&E_0[J(i_t)\overline{J}(i_{t+m})] \\ &\quad= \sum_{i\in S}\pi(i)\sum_{j\in S}\Pr(i_{t+m} = j \mid i_t = i)J(i)\overline{J}(j) \\ &\quad= \sum_{i\in S}\pi(i)J(i)[P^m\overline{J}](i) \\ &\quad= J' D P^m \overline{J}.\end{aligned}$$

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

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

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

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

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

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

which is a finite constant G, by Assumption 2-2). We have used here the notation ϕ_k to indicate the kth column of the matrix Φ, with entries $\phi_k(1), \cdots, \phi_k(n)$. Note that the second inequality above follows from the Cauchy–Schwartz inequality.

We have so far verified parts 1) and 2) of the lemma. We now begin with the analysis for part 3). 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 1).

The results of parts 4) and 5) are proved by entirely similar arguments, which we omit. ■

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

Lemma 8: 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 7, we have

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

For $\lambda = 1$, it follows that

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

Note that for $\lambda \in [0, 1)$ and any $J \in L_2(S, D)$, we have

$$\sum_{m=0}^{\infty} (\alpha\lambda P)^m J = (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \sum_{t=0}^{m} (\alpha P)^t J.$$

Hence, for $\lambda \in [0, 1)$, we have

$$E_0[s(r, X_t)] = \Phi' D \Bigg((1-\lambda) \sum_{m=0}^{\infty} \lambda^m \sum_{t=0}^{m} (\alpha P)^t \overline{g} + \Bigg((1-\lambda) \sum_{m=0}^{\infty} \lambda^m (\alpha P)^{m+1} - I \Bigg) \Phi r \Bigg)$$
$$= \Phi' D(T^{(\lambda)}(\Phi r) - \Phi r)$$

by Lemma 3. Each expression is finite and well defined by Lemma 7. ∎

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

Lemma 9: Under Assumptions 1 and 2, we have

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

Proof: We have

$$(r - r^*)' \Phi' D(T^{(\lambda)}(\Phi r) - \Phi r)$$
$$= (r - r^*)' \Phi' D((I - \Pi) T^{(\lambda)}(\Phi r) + \Pi T^{(\lambda)}(\Phi r) - \Phi r)$$
$$= (\Phi r - \Phi r^*)' D(\Pi T^{(\lambda)}(\Phi r) - \Phi r)$$

where the last equality follows because $\Phi' D\Pi = \Phi' D$ [see (1)]. As shown in the beginning of the proof of Lemma 5, $\Pi T^{(\lambda)}$ is a contraction with fixed point Φr^*, 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

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

Since $\alpha < 1$, the result follows. ∎

We now state without proof a result concerning stochastic approximation which will be used in the proof of Theorem 1. This is a special case of a very general result on stochastic approximation algorithms [19, Th. 17, p. 239]. It is straightforward to check that all of the assumptions in the result of [19] follow from the assumptions imposed in the result below. We do not show here the assumptions of [19] because the list is long and would require a lot in terms of new notation. However, we note that in our setting here, the potential function $U(\cdot)$ that would be required to satisfy the assumptions of the theorem from [19] is given by $U(r) = \|r - r^*\|^2$.

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

1) the (predetermined) step-size sequence γ_t is positive, nonincreasing, and satisfies $\sum_{t=0}^{\infty} \gamma_t = \infty$ and $\sum_{t=0}^{\infty} \gamma_t^2 < \infty$;
2) X_t is a Markov process with a unique invariant distribution, and there exists a mapping h from the states of the Markov process to the positive reals, satisfying the remaining conditions. Let $E_0[\cdot]$ stand for expectation with respect to this invariant distribution;
3) $A(\cdot)$ and $b(\cdot)$ are matrix and vector valued functions, respectively, for which $A = E_0[A(X_t)]$ and $b = E_0[b(X_t)]$ are well defined and finite;
4) the matrix A is negative definite;
5) there exist constants C and q such that for all X

$$\sum_{t=0}^{\infty} \|E[A(X_t) \mid X_0 = X] - A\| \leq C(1 + h^q(X))$$

and

$$\sum_{t=0}^{\infty} \|E[b(X_t) \mid X_0 = X] - b\| \leq C(1 + h^q(X));$$

6) for any $q > 1$ there exists a constant μ_q such that for all X, t

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

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

VI. Proof of Theorem 1

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

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

Hence, $s(r_t, X_t)$ takes 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 7, $A \triangleq E_0[A(X_t)]$ and $b \triangleq E_0[b(X_t)]$ are both well defined and finite.

By Lemma 6, we have $\Pi T^{(\lambda)}(\Phi r^*) = \Phi r^*$. From (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 8, 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 9 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 validity of all conditions except for 5) and 6). We now show that Assumption 3 is sufficient to ensure validity of these two conditions.

We begin by bounding the summations involved in 5). Letting $X_t = (i_t, i_{t+1}, z_t)$, 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

$$\begin{aligned} &E[z_t\phi'(i_t)|X_0] - E_0[z_t\phi'(i_t)] \\ &\quad = (\alpha\lambda)^{t+1}E[z_{-1}\phi'(i_t)] \\ &\qquad + \sum_{m=0}^{t}(\alpha\lambda)^m E[\phi(i_{t-m})\phi'(i_t)|X_0] \\ &\qquad - \sum_{m=0}^{\infty}(\alpha\lambda)^m E_0[\phi(i_{t-m})\phi'(i_t)]. \end{aligned}$$

Using the triangle inequality, we obtain

$$\begin{aligned} &\sum_{t=0}^{\infty}||E[z_t\phi'(i_t)|X_0] - E_0[z_t\phi'(i_t)]|| \\ &\quad \le \sum_{t=0}^{\infty}(\alpha\lambda)^{t+1}||E[z_{-1}\phi'(i_t)|X_0]|| + \sum_{t=0}^{\infty}\sum_{m=0}^{t} \\ &\qquad \cdot (\alpha\lambda)^m||E[\phi(i_{t-m})\phi'(i_t)|X_0] - E_0[\phi(i_{t-m})\phi'(i_t)]|| \\ &\qquad + \sum_{t=0}^{\infty}\sum_{m=t+1}^{\infty}(\alpha\lambda)^m||E_0[\phi(i_{t-m})\phi'(i_t)]||. \end{aligned}$$

We will individually bound the magnitude of each summation in the right-hand side.

First we have

$$\begin{aligned} &\sum_{t=0}^{\infty}(\alpha\lambda)^{t+1}||E[z_{-1}\phi'(i_t)|X_0]|| \\ &\quad = ||z_{-1}||\sum_{t=0}^{\infty}(\alpha\lambda)^{t+1}||E[\phi(i_t)|X_0]|| \\ &\quad = \frac{1}{\alpha\lambda}||z_0 - \phi(i_0)||\sum_{t=0}^{\infty}(\alpha\lambda)^{t+1}||E[\phi(i_t)|X_0]|| \end{aligned}$$

where the second inequality follows from the fact that $z_0 = (\alpha\lambda)z_{-1} + \phi(i_0)$. Assumption 3-1) implies that

$$||E[\phi(i_t)|X_0]|| \le C(1 + f(i_0) + f(i_1))^q$$

for some constants C and q and any $t \ge 0$. It follows that

$$\begin{aligned} &\sum_{t=0}^{\infty}(\alpha\lambda)^{t+1}||E[z_{-1}\phi'(i_t)|X_0]|| \\ &\quad \le C(1 + ||z_0|| + f(i_0) + f(i_1))^q \end{aligned}$$

for some constants C and q.

Next, we deal with the second summation. Letting $\Delta_{t-m,\,t}$ be defined by

$$\Delta_{t-m,\,t} = ||E[\phi(i_{t-m})\phi'(i_t)|X_0] - E_0[\phi(i_{t-m})\phi'(i_t)]||$$

we have

$$\begin{aligned} &\sum_{t=0}^{\infty}\sum_{m=0}^{t}(\alpha\lambda)^m\Delta_{t-m,\,t} \\ &\quad = \sum_{m=0}^{\infty}(\alpha\lambda)^m\left(\Delta_{0,\,m} + \sum_{t=m+1}^{\infty}\Delta_{t-m,\,t}\right) \\ &\quad \le C(f(i_0) + f(i_1)) \end{aligned}$$

for some constant C, where the inequality follows from Assumption 3-1).

Finally, recalling that $E_0[\phi(i_{t-m})\phi'(i_t)] \le G$, for some absolute constant G (Lemma 7), we have

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

Given these bounds, it follows that there exist positive constants C and q such that

$$\begin{aligned} &\sum_{t=0}^{\infty}||E[z_t\phi'(i_t)|X_0] - E_0[z_t\phi'(i_t)]|| \\ &\quad \le C(1 + ||z_0|| + f(i_0) + f(i_1))^q. \end{aligned}$$

In other words, the summation above is bounded by a polynomial function of $||z_0||$, $f(i_0)$, and $f(i_1)$. 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. Using these arguments, we can place bounds that are polynomial in $||z_0||$, $f(i_0)$, and $f(i_1)$, on the summations in Condition 5) of Theorem 2. We can thus satisfy the condition with a function $h(X)$ ($X = (i, j, z)$) that is polynomial in $||z||$, $f(i)$, and $f(j)$. The fact that such a function h would satisfy Condition 6) then follows from Assumption 3-2).

We now have all the conditions needed to apply Theorem 2. It follows that r_t converges to r^*, which solves $Ar^* + b = 0$. Since $Ar^* + b = E_0[s(r^*, X_t)]$, Lemma 8 implies that

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

By Lemma 6 along with the fact that $\Phi' D$ has full row rank [by virtue of Assumption 2-1)], r^* uniquely satisfies this

equation and is the unique fixed point of $\Pi T^{(\lambda)}$. Lemma 6 also provides the desired error bound. This completes the proof to Theorem 1.

VII. The Case of a Finite State Space

In this section, we show that Assumptions 1-2), 2-2), and 3 are automatically true whenever we are dealing with an irreducible aperiodic Markov chain with a finite state space. This tremendously simplifies the conditions required to apply Theorem 1, reducing them to a requirement that the basis functions be linearly independent [Assumption 2a)]. Actually, even this assumption can be relaxed if Theorem 1 is stated in a more general way. This assumption was adopted for the sake of simplicity in the proof.

Let us now assume that i_t is an irreducible aperiodic finite-state Markov chain [Assumption 1-1)]. Assumptions 1-2) and 2-2) are trivially satisfied when the state space is finite. We therefore only need to prove that Assumption 3 is satisfied.

It is well known that for any irreducible aperiodic finite-state 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 7-1). 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 one, and therefore there exists a constant G such that $||P^m|| \leq G$ for all m. We then have

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

The first part of Assumption 3-1) is thus satisfied by a function $f(i)$ that is equal to a constant for all i. An analogous argument, which we omit, can be used to establish that the same is true for the second part of Assumption 3-1). Assumption 3-2) follows from the fact that $f(i)$ is constant.

VIII. 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 countable subset of $\Re^N$. Each state $x \in \Re^n$ is associated with an integer index $i \in \{1, 2, 3, \cdots\}$ and denoted by $x(i)$.

Let us first assume that the state space is a bounded subset of $\Re^N$ and that the mappings defined by $(x(i), x(j)) \mapsto g(i, j)$ and $x(i) \mapsto \phi_k(i)$ are continuous functions on $\Re^N \times \Re^N$ and $\Re^N$. Then, Assumptions 1-2) and 2-2) are automatically valid because continuous functions are bounded on bounded sets.

Assumption 3-1) 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 3-3) is satisfied by a function $f(i) = C$ 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 zero, and that for some $\delta > 0$

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

Then, $D_t(i_0)$ converges to D exponentially fast, and uniformly in i_0, and Assumption 3-1) is satisfied with $f(i) = C$. Validity of Assumption 3-2) easily follows.

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 $x(i) \mapsto \pi(i)$ exhibit exponential decay; let us assume that this is the case.

For the purposes of Assumption 3, it is natural in this context to employ a function $f(i) = C(1 + ||x(i)||^q)$, for some C and q. Assumption 3-2) is essentially a stability condition; given our definition of f, it states that $||x(i_t)||^q$ is not expected to grow too rapidly, and this is satisfied by most stable Markov chains of practical interest. Note that by taking the steady-state limit we obtain $E_0[||x(i_t)||^q] < \infty$ 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 3-1) is 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 $||x(i_0)||$ increases; this is captured by the term $f(i_0)$ 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 [20]. 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 and 2 are satisfied. Assumption 3-2) can be shown to be true for functions of the form $f(i) = C(1+\|x(i)\|^q)$ by exploiting the downward drift property (in this example, it is natural to simply let $x(i) = i$).

Let us now discuss Assumption 3-1). 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 zero [21], [22]. Once the vicinity of state zero 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 3-1) is satisfied by a function f that grows polynomially with $\|x(i)\|$.

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.

IX. 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 [10], [11], [13], and [14], present such examples as counterexamples to TD(λ). In this section, we provide some insight into this issue by 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 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 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(0) algorithm as defined in Section II. 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 a countable set S with at least two elements. 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 $1 > p(2) > 5/6\alpha$ 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 an $n \times 1$ 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

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

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

Specializing to our choice of parameters, the recurrence becomes

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

For shorthand notation, let Δ be defined by

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

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

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

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

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

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 $r_0 \neq r^*$. ∎

X. Divergence with a Nonlinear Approximator

Our analysis of temporal-difference learning up until now has focused on linear function approximators. In many situations, it may be natural to employ classes of nonlinear function approximators. Neural networks present one popular example. One might hope that the analysis we have provided for the linear case generalizes to nonlinear parameterizations, perhaps under some simple regularity conditions. Unfortunately, this does not seem to be the case. To illustrate potential difficulties, we present an example for which TD(0) diverges due to the structure of a nonlinear function approximator. (By divergence here, we mean divergence of both the approximate cost-to-go function and the parameters.) For the sake of brevity, we limit our study to a characterization of steady-state dynamics, rather than presenting a rigorous proof, which would require arguments formally relating the steady-state dynamics to the actual stochastic algorithm.

We consider a Markov chain with three states ($S = \{1, 2, 3\}$), all transition costs equal to zero, and a discount factor $\alpha \in (0, 1)$. The cost-to-go function $J^* \in \Re^3$ is therefore given by $J^* = (0, 0, 0)'$. Let the function approximator

$$\tilde{J}(r) = (\tilde{J}(1, r), \tilde{J}(2, r), \tilde{J}(3, r))'$$

be parameterized by a single scalar r. Let the form of $\tilde{J}$ be defined by letting $\tilde{J}(0)$ be some nonzero vector satisfying $e'\tilde{J}(0) = 0$, where $e = (1, 1, 1)'$, requiring that $\tilde{J}(r)$ be the unique solution to the linear differential equation

$$\frac{d\tilde{J}}{dr}(r) = (Q + \epsilon I)\tilde{J}(r) \tag{6}$$

where I is the 3×3 identity matrix, ϵ is a small positive constant, and Q is given by

$$Q = \begin{bmatrix} 1 & 1/2 & 3/2 \\ 3/2 & 1 & 1/2 \\ 1/2 & 3/2 & 1 \end{bmatrix}.$$

Given our definition of $\tilde{J}$, it is easy to show that all functions representable by $\tilde{J}$ lie on the plane $\{J \in \Re^2 | e'J = 0\}$. Furthermore, the set of functions $\{\tilde{J}(r) | r \in \Re\}$ forms a spiral that diverges as r grows to infinity (see Fig. 1).

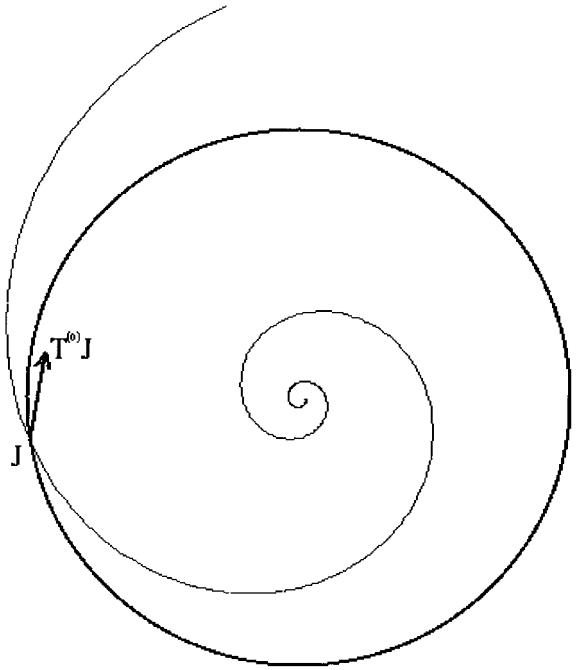

Fig. 1. Example of divergence with a nonlinear function approximator. The plot is of points in the plane $\{J \in \Re^3 | e'J = 0\}$.

We let the transition probability matrix of the Markov chain be

$$P = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \end{bmatrix}.$$

Since all transition costs are zero, the TD(0) operator is given by $T^{(0)}J = \alpha PJ$, for all $J \in \Re^3$. It turns out that there is an acute angle θ and a scalar $\beta \in (0, 1)$ such that for any r, $T^{(0)}\tilde{J}(r)$ is equal to the vector $\tilde{J}(r)$ scaled by β and rotated by θ degrees in the plane $\{J \in \Re^3 | e'J = 0\}$. The points labeled J and $T^{(0)}J$ in Fig. 1 illustrate the nature of this transformation.

Before discussing divergence of TD(0), let us motivate the underlying intuition by observing the qualitative behavior of a simpler algorithm. In particular, suppose we generated a sequence of approximations $\tilde{J}(r_t)$, where each r_{t+1} satisfies

$$r_{t+1} = \arg\min_r \|\tilde{J}(r) - T^{(0)}\tilde{J}(r_t)\|.$$

(Note that the steady-state distribution is uniform so that the Euclidean norm is the appropriate one for this context.) In Fig. 1, the point on the spiral closest to $T^{(0)}J$ is further from the origin than J, even though $T^{(0)}J$ is closer to the origin than J (the origin is located at the center of the circle in the diagram). Therefore, if $\tilde{J}(r_0) = J$, then $\|\tilde{J}(r_1)\| > \|\tilde{J}(r_0)\|$. Furthermore, since each application of $T^{(0)}$ induces the same degree of rotation and scaling, we might expect that each subsequent iteration takes the approximation further from the origin in a completely analogous way. Hence, the underlying dynamics suggest that divergence is conceivable.

Let us now more concretely identify divergent behavior in the steady-state dynamics of TD(0). The TD(0) algorithm applies the update equation

$$r_{t+1} = r_t + \gamma_t \frac{d\tilde{J}}{dr}(r)(\alpha \tilde{J}(i_{t+1}, r_t) - \tilde{J}(i_t, r))$$

where i_t is the state visited by the trajectory at time t. Since the steady-state distribution resulting from P is uniform, the steady-state expectation of the update direction, within a factor of three, is given by

$$\sum_{i=1}^{3} \frac{d\tilde{J}}{dr}(i, r)\left(\alpha \sum_{i=1}^{3} p_{ij}\tilde{J}(j, r) - \tilde{J}(i, r)\right).$$

This is the inner product of the vector $d\tilde{J}/dr$, which is $(Q + \epsilon I)\tilde{J}(r)$, with the vector with components $\alpha \sum_{i=1}^{3} p_{ij}\tilde{J}(j, r) - \tilde{J}(r)$, which is the vector $\alpha P\tilde{J}(r) - \tilde{J}(r)$.

As the step size becomes extremely small, we can think of the deterministic version of the algorithm as an approximation to a differential equation. Given the average direction of motion of the parameter r, the appropriate differential equation for our example is

$$\begin{aligned}\frac{dr}{dt} &= (Q + \epsilon I)\tilde{J}(r))'(\alpha P - I)\tilde{J}(r) \\ &= \tilde{J}'(r)(Q' + \epsilon I)(\alpha P - I)\tilde{J}(r).\end{aligned}$$

For $\epsilon = 0$, we have

$$\begin{aligned}\frac{dr}{dt} &= \tilde{J}'(r)Q'(\alpha P - I)\tilde{J}(r) \\ &= \alpha \tilde{J}'(r)Q'P\tilde{J}(r) \\ &= \frac{\alpha}{2}\,\tilde{J}'(r)(Q'P + P'Q)\tilde{J}(r)\end{aligned}$$

where the first equality follows from the fact that $\tilde{J}'(r)Q'\tilde{J}(r) = 0$, for any r. Note that

$$(Q'P + P'Q) = \begin{bmatrix} 2.5 & 1.75 & 1.75 \\ 1.75 & 2.5 & 1.75 \\ 1.75 & 1.75 & 2.5 \end{bmatrix}$$

which is easily verified to be positive definite. Hence, there exists a positive constant c such that

$$\frac{dr}{dt} \geq c\|\tilde{J}(r)\|^2. \tag{7}$$

By a continuity argument, this inequality remains true (possibly with a smaller positive constant c) if ϵ is positive but sufficiently small. The combination of this inequality and the fact that

$$\begin{aligned}\frac{d}{dr}\|\tilde{J}(r)\|^2 &= \tilde{J}'(r)(Q + Q')\tilde{J}(r) + 2\epsilon\|\tilde{J}(r)\|^2 \\ &\geq 2\epsilon\|\tilde{J}(r)\|^2\end{aligned}$$

implies that both r and $\|\tilde{J}(r)\|$ diverge to infinity.

XI. Conclusions

We have established the convergence of on-line temporal-difference learning with linear function approximators when applied to irreducible aperiodic 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: 1) we can keep track of the magnitude of the approximation error $\Phi r^* - J^*$; the natural norm for doing so is $\|\cdot\|_D$, or 2) 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 αPJ. The deterministic counterpart of the algorithm, as introduced in Section III, takes the form

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

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 Φr^* 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 factor of

$$\frac{1 - \alpha\lambda}{1 - \alpha}.$$

This expression is one when $\lambda = 1$. For every $\lambda < 1$, it is larger than one, 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 [23].

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 one. Experimental results [2], [24], and [25] 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 learning, 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 [16]. It remains an open problem to analyze the limiting behavior of the parameters r and the resulting approximations Φr 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}. In particular, we can replace the Markov process $X_t = (i_t, i_{t+1}, z_t)$ that was constructed for the purposes of our analysis with a process $X_t = (i_t, i_{t+1}, z_t, g_t)$, where g_t is the cost associated with the transition from i_t to i_{t+1}. Then, as long as the distribution of the noise only depends on the current state and its moments are such that the assumptions of Theorem 2 are still satisfied, our proof can easily be modified to accommodate this situation. Finally, the assumption that the Markov chain was aperiodic can be alleviated. No part of our convergence proof truly required this assumption—it was introduced merely to simplify the exposition.

Our results in Section IX 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 sample 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 [24] 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 [16]. However, different mechanisms for adjusting the coefficients z_t lead to a change in the steady-state average value of $z_t\phi'(i_t)$, affect the matrix A, and the negative definiteness property can be easily lost.

Finally, the example of Section X identifies the possibility of divergence when TD(λ) is used in conjunction with nonlinear function approximators. However, the example is somewhat contrived, and it is unclear whether divergence can occur with special classes of function approximators, such as neural networks. This presents an interesting question for future research.

ACKNOWLEDGMENT

The authors would like to thank R. S. Sutton for starting them on the path that led to this work by pointing out that the counterexample in [10] would no longer be a counterexample if on-line state sampling was used. They also thank him for suggesting an algebraic simplification to the original expression for the error bound in Theorem 1, which resulted in its current form. The authors would like to thank the reviewers for their feedback, especially the one who provided them with four pages of detailed corrections and useful comments.

REFERENCES

[1] D. P. Bertsekas, *Dynamic Programming and Optimal Control.* Belmont, MA: Athena Scientific, 1995.
[2] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Mach. Learning,* vol. 3, pp. 9–44, 1988.
[3] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learning,* vol. 8, pp. 279–292, 1992.
[4] J. N. Tsitsiklis, "Asynchronous stochastic approximation and Q-learning," *Mach. Learning,* vol. 16, pp. 185–202, 1994.
[5] T. Jaakkola, M. I. Jordan, and S. P. Singh, "On the convergence of stochastic iterative dynamic programming algorithms," *Neural Comp.,* vol. 6, no. 6, pp. 1185–1201, 1994.
[6] P. D. Dayan and T. J. Sejnowski, "TD(λ) converges with probability 1," *Mach. Learning,* vol. 14, pp. 295–301, 1994.
[7] L. Gurvits, L. J. Lin, and S. J. Hanson, "Incremental learning of evaluation functions for absorbing Markov chains: New methods and theorems," 1994, preprint.
[8] P. D. Dayan, "The convergence of TD(λ) for general λ," *Mach. Learning,* vol. 8, pp. 341–362, 1992.
[9] R. E. Schapire and M. K. Warmuth, "On the worst-case analysis of temporal-difference learning algorithms," *Mach. Learning,* vol. 22, pp. 95–122, 1996.
[10] J. N. Tsitsiklis and B. Van Roy, "Feature-based methods for large scale dynamic programming," *Mach. Learning,* vol. 22, pp. 59–94, 1996.
[11] G. J. Gordon, "Stable function approximation in dynamic programming," Carnegie Mellon Univ., Tech. Rep. CMU-CS-95-103, 1995.
[12] S. P. Singh, T. Jaakkola, and M. I. Jordan, "Reinforcement learning with soft state aggregation," in *Advances in Neural Information Processing Systems,* vol. 7, G. Tesauro, D. S. Touretzky, and T. K. Leen, Eds. Cambridge, MA: MIT Press, 1995.
[13] L. C. Baird, "Residual algorithms: Reinforcement learning with function approximation," in *Machine Learning: Proceedings 12th Int. Conf.,* July 9–12, Prieditis and Russell, Eds. San Francisco, CA: Morgan Kaufman, 1995.
[14] J. A. Boyan and A. W. Moore, "Generalization in reinforcement learning: Safely approximating the value function," in *Advances in Neural Information Processing Systems,* vol. 7. MIT Press, 1995.
[15] R. S. Sutton, "On the virtues of linear learning and trajectory distributions," in *Proc. Wkshp. Value Function Approximation, Mach. Learning Conf.,* Carnegie Mellon Univ., Tech. Rep. CMU-CS-95-206, 1995.
[16] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming.* Belmont, MA: Athena Scientific, 1996.
[17] L. Gurvits, 1996, private communication.
[18] F. Pineda, "Mean-field analysis for batched TD(λ)," 1996, preprint.
[19] A. Benveniste, M. Métivier, and P. Prioret, *Adaptive Algorithms and Stochastic Approximations.* Berlin: Springer-Verlag, 1990.
[20] J. Walrand, *An Introduction to Queueing Networks.* Englewood Cliffs, NJ: Prentice Hall, 1988.

[21] G. D. Stamoulis and J. N. Tsitsiklis, "On the settling time of the congested GI/G/1 queue," *Adv. Appl. Probability*, vol. 22, pp. 929–956, 1990.
[22] P. Konstantopoulos and F. Baccelli, "On the cut-off phenomenon in some queueing systems," *J. Appl. Probability*, vol. 28, pp. 683–694, 1991.
[23] D. P. Bertsekas, "A counterexample to temporal-difference learning," *Neural Comp.*, vol. 7, pp. 270–279, 1994.
[24] S. P. Singh and R. S. Sutton, "Reinforcement learning with replacing eligibility traces," *Mach. Learning*, vol. 22, pp. 123–158, 1996.
[25] R. S. Sutton, "Generalization in reinforcement learning: Successful examples using sparse coarse coding," in *Advances in Neural Information Processing Systems*, vol. 8, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds. Cambridge, MA: MIT Press, 1996.

John N. Tsitiklis (S'81–M'83) was born in Thessaloniki, Greece, in 1958. He received the B.S. degree in mathematics in 1980 and the B.S., M.S., and Ph.D. degrees in electrical engineering, all from the Massachusetts Institute of Technology, Cambridge, MA, in 1980, 1981, and 1984, respectively.

During the academic year 1983–1984, he was an Acting Assistant Professor of Electrical Engineering at Stanford University, Stanford, CA. Since 1984, he has been with the Massachusetts Institute of Technology, where he is currently Professor of Electrical Engineering. His research interests include systems and control theory, neural networks, and operations research. He has written more than 70 journal papers on these subjects and is a coauthor of *Parallel and Distributed Computation: Numerical Methods* (1989), *Neuro-Dynamic Programming* (1996), and *Introduction to Linear Optimization* (1997).

Dr. Tsitiklis has been a recipient of an IBM Faculty Development Award (1983), an NSF Presidential Young Investigator Award (1986), an Outstanding Paper Award by the IEEE Control Systems Society, the MIT Edgerton Faculty Achievement Award (1989), and the Bodossakis Foundation Prize (1994). He was a plenary speaker at the 1992 IEEE Conference on Decision and Control. He is an associate editor of *Applied Mathematics Letters* and has been an associate editor of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL and *Automatica*.

Benjamin Van Roy was born in Bangkok, Thailand, in 1971. He received the B.S. degree in computer science in 1993 and the M.S. degree in electrical engineering and computer science in 1995, both from the Massachusetts Institute of Technology, Cambridge, MA, where he is currently a Ph.D. candidate.

He is a coauthor of *Solving Pattern Recognition Problems* (1995), and he consults regularly with private industry.

Mr. Van Roy has been a recipient of a Digital Equipment Corporation Scholarship, the MIT George C. Newton Award for "the best undergraduate electrical engineering laboratory project," and the MIT Morris J. Levin Memorial Award for "an outstanding Master's thesis presentation."