May, 1997

LIDS-P 2390

Research Supported By:

NSF grant DMI-9625489
AFOSR grant F49620-95-1-0219

Average Cost Temporal-Difference Learning

Tsitsiklis, J.N.
Roy, B.V.

LIDS-P-2390

May 16, 1997

Average Cost Temporal–Difference Learning[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 DMI-9625489 and the AFOSR under grant F49620-95-1-0219.

ABSTRACT

We propose a variant of temporal–difference learning that approximates average and differential costs of an irreducible aperiodic Markov chain. Approximations are comprised of linear combinations of fixed basis functions whose weights are incrementally updated during a single endless trajectory of the Markov chain. We present a proof of convergence (with probability 1), and a characterization of the limit of convergence. We also provide a bound on the resulting approximation error that exhibits an interesting dependence on the "mixing time" of the Markov chain. The results parallel previous work by the authors, involving approximations of discounted cost–to–go.

1 Introduction

Temporal–difference learning, originally proposed by Sutton (1988), is an algorithm for approximating the cost–to–go function of a Markov chain (the expected future cost, as a function of the initial state). Given a set of basis functions, the algorithm tunes a vector of weights so that the weighted combination of the basis functions approximates the cost–to–go function. The weights are iteratively adapted based on information drawn from either simulation or observation of a Markov process. Updates occur upon each state transition with the objective of improving the approximation as time progresses.

The reason for our interest in cost–to–go functions is their central role in dynamic programming algorithms for solving Markov decision problems. In particular, given the cost–to–go function associated with a given control policy, one can perform an iteration of the classical policy iteration method to obtain an improved policy (Bertsekas, 1995). However, when the state space is large, the exact computation of the cost–to–go function becomes infeasible, and this is why we are interested in approximations.

A comprehensive convergence analysis for the case of discounted Markov chains has been provided by the authors (Tsitsiklis and Van Roy, 1997). A simplified version of that work, together with extensions to the case of undiscounted absorbing Markov chains, is presented in (Bertsekas and Tsitsiklis, 1995). Related analyses are given by (Sutton, 1988), (Dayan, 1992), (Gurvits et al., 1994), and (Pineda, 1996). The purpose of the present paper is to propose a variant of temporal–difference learning that is suitable for approximating differential cost functions of undiscounted Markov chains (i.e., solutions to Poisson's equation), and to extend the analysis of (Tsitsiklis and Van Roy, 1997) to this new context. The contributions of this paper include the following:

1. A temporal–difference learning algorithm that approximates differential cost functions is proposed.

2. Convergence (with probability 1) is established for the case where approximations are generated by linear combinations of basis functions over a finite state space.

3. 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. Furthermore, a relationship between the error bound and the "mixing time" of the Markov chain is identified.

This paper is not the first to consider simulation–based methods that iteratively evaluate differential cost functions. However, the algorithms that have been explored in this context generally make use of look–up table representations, which involve storing and updating one value per state in the state space. We refer the reader to (Mahadevan, 1996) for a survey of relevant experimental work and to (Abounadi, Bertsekas, and Borkar, 1997) for a theoretical treatment.

There is no prior work explicitly dealing with approximations of differential cost functions. It is known that the differential cost function of an infinite horizon Markov chain is the same as the cost–to–go function of an auxiliary absorbing Markov chain (Bertsekas, 1995; Bertsekas and Tsitsiklis, 1996). This relationship motivates one way of using temporal–difference learning to approximate a differential cost function, namely, deriving the auxiliary absorbing Markov chain and then employing an existing version of temporal–difference

learning. However, this reduction can affect approximations in undesirable ways, as we discuss next.

In temporal–difference learning, each weight update is dependent on a history of visited states. When temporal–difference learning is applied to an absorbing Markov chain, multiple finite trajectories (each terminating at an absorbing state) are simulated. Weight updates occur during these simulations, and the history of visited states is erased upon the termination of each trajectory. Even though restarting the record of visited states is appropriate for an absorbing Markov chain, it is unnatural for the original infinite horizon Markov chain. Due to this peculiarity introduced by the reduction, it is preferable to use a variant of temporal–difference learning designed specifically for approximating differential cost functions, as the one we will introduce in this paper.

The remainder of this paper is organized as follows. In Section 2, we provide a precise definition of the algorithm. Section 3 presents our convergence result, together with assumptions and a proof. In Section 4, we develop a bound for the approximation error associated with the limit of convergence. Section 5 presents and analyzes another variant of temporal–difference learning. Some new insights that stem from the analysis are also discussed. Finally, concluding remarks are made in Section 6.

2 Average Cost Temporal–Difference Learning

In this section, we define precisely the nature of average cost temporal–difference learning. While the temporal–difference learning method as well as our subsequent results can be generalized to Markov chains with infinite state spaces, we restrict our attention to the case where the state space is finite.

We consider a Markov chain with a state space $S = \{1, \ldots, n\}$. The sequence of states visited by the Markov chain is denoted by $\{i_t \mid t = 0, 1, \ldots\}$. The Markov chain is defined by a transition probability matrix P whose (i, j)th entry, denoted by p_{ij}, is the probability that $i_{t+1} = j$ given that $i_t = i$. We make the following assumption concerning the dynamics of the Markov chain:

Assumption 1 *The Markov chain corresponding to P is irreducible and aperiodic.*

It follows from this assumption that the Markov chain has a unique invariant distribution π that satisfies $\pi' P = \pi'$ with $\pi(i) > 0$ for all i. Let $E_0[\cdot]$ denote expectation with respect to this distribution.

For any state $i \in S$, a scalar $g(i)$ represents the cost of remaining in the state for one time step. We define the average cost by $\mu^* = E_0[g(i_t)]$, and a differential–cost function is any function $J : S \mapsto \Re$ satisfying Poisson's equation, which takes the form

$$J = g - \mu^* e + PJ,$$

where $e \in \Re^n$ is the vector with each component equal to 1, and J and g are viewed as vectors in $\Re^n$. Under Assumption 1, it is known that differential cost functions exist and the set of all differential cost functions takes the form $\{J^* + ce | c \in \Re\}$, for some function J^* satisfying $\pi' J^* = 0$ (Gallager, 1996). We will refer to J^* as the *basic* differential cost function, and it is known that, under Assumption 1, this function is given by

$$J^* = \sum_{t=0}^{\infty} P^t (g - \mu^* e). \tag{1}$$

Furthermore, if J is any differential cost function, then $J(i) - J(j)$ represents the difference in expected future costs if the initial state j were to be replaced by i, and can be used to guide the policy iteration algorithm.

We consider approximations to differential cost functions 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 n), while each $r(k)$ can be viewed as the associated weight. In approximating a differential cost function, one wishes to choose the parameter vector r so as to minimize some error metric between the function $\tilde{J}(\cdot,r)$ and the space of differential cost functions.

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 an $n \times K$ matrix whose kth column is equal to ϕ_k; that is,

$$\Phi = \begin{bmatrix} | & & | \\ \phi_1 & \cdots & \phi_K \\ | & & | \end{bmatrix}.$$

We make the following assumption concerning the choice of basis functions:

Assumption 2 *(a) The basis functions $\{\phi_k \mid k = 1,\ldots,K\}$ are linearly independent (i.e., Φ has full rank).*
(b) For every $r \in \Re^K$, $\Phi r \neq e$.

Suppose that we observe a sequence of states i_t generated according to the transition probability matrix P. Given that at a time t, the parameter vector r has been set to some value r_t, and we have an approximation μ_t to the average cost μ^*, we define the temporal difference d_t corresponding to the transition from i_t to i_{t+1} by

$$d_t = g(i_t) - \mu_t + \tilde{J}(i_{t+1},r_t) - \tilde{J}(i_t,r_t). \tag{2}$$

For each time $t = 0,1,\ldots$, the average cost temporal–difference learning algorithm updates the average cost estimate μ_t and the parameter vector r_t. The average cost estimate is updated according to

$$\mu_{t+1} = (1-\eta_t)\mu_t + \eta_t g(i_t),$$

where μ_0 is an initial estimate and η_t is a sequence of scalar step sizes. The parameter vector evolves according to a more complex iteration:

$$r_{t+1} = r_t + \gamma_t d_t \sum_{k=0}^{t} \lambda^{t-k}\phi(i_k), \tag{3}$$

where the components of r_0 are initialized to arbitrary values, 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

$$z_t = \sum_{k=0}^{t} \lambda^{t-k} \phi(i_k). \tag{4}$$

With this new notation, the parameter updates are given by

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

and the eligibility vectors can be updated according to

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

initialized with $z_{-1} = 0$. Note that it is important to set the parameter λ to values less than one, since the eligibility vector becomes unstable if $\lambda \geq 1$.

We make one final assumption, regarding the step size sequences.

Assumption 3 *(a) The sequence γ_t is positive, deterministic, and satisfies $\sum_{t=0}^{\infty} \gamma_t = \infty$ and $\sum_{t=0}^{\infty} \gamma_t^2 < \infty$.*
(b) There exists a positive scalar c such that the sequence η_t satisfies $\eta_t = c\gamma_t$, for all t.

The TD(λ) algorithm we have described simultaneously tunes estimates of the average cost and a differential cost function. This will be the primary algorithm studied in this paper. However, we will also consider an interesting alternative that involves first finding an estimate μ to the average cost μ^*, and then carrying out the updates given by Equation (3) with d_t defined by

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

instead of Equation (2). We analyze this algorithm in Section 5, and in the process, we will uncover an intriguing relationship between the error $|\mu^* - \mu|$ of the average cost estimate and the error of the approximated differential cost function.

3 Convergence Result

In this section, we present the main result of this paper, which establishes convergence and characterizes the limit of convergence of average cost temporal–difference learning. We begin by introducing some notation that helps to streamline the formal statement of results, as well as the analysis.

Recall that $\pi(1), \ldots, \pi(n)$ denote the steady-state probabilities for the process i_t. We define an $n \times n$ diagonal matrix D with diagonal entries $\pi(1), \ldots, \pi(n)$. It is easy to see that $\langle x, y \rangle_D = x'Dy$ defines a Hilbert space with norm $\| \cdot \|_D = \sqrt{\langle \cdot, \cdot \rangle_D}$. We say that two vectors $J, \overline{J}$ are D–orthogonal if $J'D\overline{J} = 0$. Regarding notation, we will also use $\| \cdot \|$, without a subscript, to denote the Euclidean norm on vectors or the Euclidean-induced norm on matrices. (That is, for any matrix A, we have $\|A\| = \max_{\|x\|=1} \|Ax\|$.)

We define a projection matrix Π that projects onto the subspace spanned by the basis functions. In particular, we let $\Pi = \Phi(\Phi' D \Phi)^{-1} \Phi' D$. For any $J \in \Re^n$, we then have

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

We define an operator that is useful in characterizing the dynamics of average cost temporal–difference learning. 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 $T^{(\lambda)} : \Re^n \mapsto \Re^n$. It is defined by

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

To interpret $T^{(\lambda)}$ in a meaningful manner, note that, for each m, the term

$$\sum_{t=0}^{m}P^t(g-\mu^* e)+P^{m+1}J$$

is an approximation to the basic differential cost function where the summation in Equation (1) is truncated after m terms, and the remainder of the summation is approximated by $P^{m+1}J$. In fact, the remainder of the summation is exactly equal to $P^{m+1}J^*$, so $P^{m+1}J$ is a reasonable approximation when J^* is unknown and J is its estimate. The function $T^{(\lambda)}J$ is therefore a geometrically weighted average of approximations to the differential cost function.

Our convergence result follows.

Theorem 1 *Under Assumptions 1, 2, and 3, the following hold:*
(a) For any $\lambda \in [0,1)$, the average cost TD(λ) algorithm, as defined in Section 2, converges with probability 1.
(b) The limit of the sequence μ_t is the average cost μ^*.
(c) The limit r^* of the sequence r_t is the unique solution of the equation

$$\Pi T^{(\lambda)}(\Phi r^*) = \Phi r^*.$$

3.1 Preview of Proof

The next few subsections are dedicated to the development of a proof. Before diving into technicalities, let us clarify the fundamental structure of the algorithm and discuss the approach we take for its analysis.

We construct a process $X_t = (i_t, i_{t+1}, z_t)$, where z_t is the eligibility vector defined by Equation (4). 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 values of μ_t and r_t, provides all necessary information for computing μ_{t+1} and r_{t+1}.

So that we can think of the TD(λ) algorithm as adapting only a single vector, we introduce a sequence $\theta_t \in \Re^{K+1}$ with components $\theta_t(1) = \mu_t$ and $\theta_t(i) = r_t(i-1)$ for $i \in \{2,\ldots,n+1\}$, or using more compact notation,

$$\theta_t = \begin{bmatrix}\mu_t \\ r_t\end{bmatrix}.$$

The TD(λ) updates can be rewritten as

$$\theta_{t+1} = \theta_t + \gamma_t(A(X_t)\theta_t + b(X_t)), \tag{5}$$

for certain matrix and vector–valued functions $A(\cdot)$ and $b(\cdot)$. In particular, for any $X = (i, j, z)$, $A(\cdot)$ is given by

$$A(X) = \begin{bmatrix} -c & 0 \cdots 0 \\ -z & z(\phi'(j) - \phi'(i)) \end{bmatrix},$$

and $b(\cdot)$ is given by

$$b(X) = \begin{bmatrix} cg(i) \\ zg(i) \end{bmatrix},$$

where c is the constant in Assumption 3(b). Note that $z(\phi'(j) - \phi'(i))$ is a $K \times K$ matrix and z is a K–dimensional vector. Hence, for any X, $A(X)$ is a $(K+1) \times (K+1)$ matrix, while $b(X)$ is a $(K+1)$–dimensional vector.

As we will show later, $A(X_t)$ and $b(X_t)$ have well defined "steady-state" expectations, which we denote by A and b. General results concerning stochastic approximation algorithms can be used to show that the asymptotic behavior of the sequence generated by Equation (5) mimics that of an ordinary differential equation:

$$\dot{\theta}_t = A\theta_t + b.$$

Our analysis can be broken down into two parts. The first establishes that the relevant ordinary differential equation converges (we will show that the matrix A is stable). The second involves the application of a result from stochastic approximation theory to show that the algorithm delivers similar behavior.

The following subsections are organized as follows. Subsection 3.2 proves a few lemmas pertinent to characterizing the matrix A and the vector b, and establishing stability of A. Subsection 3.3 presents the stochastic approximation result that will be employed. Finally, in Subsection 3.4, the machinery provided by Subsection 3.2 is integrated with the stochastic approximation result in order to prove the theorem.

3.2 Preliminaries

We begin with a fundamental lemma on Markov chains, which is central to the analysis of TD(λ).

Lemma 1 *Under Assumption 1, for all $J \in \Re^n$,*

$$\|PJ\|_D \leq \|J\|_D.$$

Furthermore, unless J is proportional to e,

$$PJ \neq J.$$

Proof: The fact that $\|PJ\|_D \leq \|J\|_D$ is proven as Lemma 1 in (Tsitsiklis and Van Roy, 1997). The second part is implied by Assumption 1(a), which ensures that $\lim_{m\to\infty} P^m J \in \{ce|c \in \Re\}$, ruling out the possibility that $PJ = J$ for $J \notin \{ce|c \in \Re\}$. **q.e.d.**

The next lemma shows that analogous properties are enjoyed by a geometrically weighted convex combination of powers of P.

Lemma 2 *Let* $P^{(\lambda)} = (1-\lambda)\sum_{m=0}^{\infty}\lambda^m P^{m+1}$. *Then, under Assumption 1, for any* $\lambda \in [0,1)$ *and* $J \in \Re^n$,

$$\|P^{(\lambda)}J\|_D \leq \|J\|_D.$$

Furthermore, unless J *is proportional to* e,

$$P^{(\lambda)}J \neq J.$$

Proof: The fact that $\|P^{(\lambda)}J\|_D \leq \|J\|_D$ follows from the first part of Lemma 1 and the triangle inequality.

Suppose that $P^{(\lambda)}J = J$. Then, J is a convex combination of the vectors $P^m J$, all of which belong to the set $\{\overline{J} \mid \|\overline{J}\|_D \leq \|J\|_D\}$. Since J is an extreme point of this set, we must have $J = P^m J$ for all m. By Lemma 1, this implies that J is proportional to e. **q.e.d.**

The following lemma establishes that the set of fixed points of $T^{(\lambda)}$ is the set of differential cost functions.

Lemma 3 *Under Assumption 1, for any* $\lambda \in [0,1)$, *we have*

$$T^{(\lambda)}J = J \quad \text{if and only if} \quad J \in \{J^* + ce | c \in \Re\}.$$

Proof: Suppose that $J = J^* + ce$, for some scalar c. Then,

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

On the other hand, suppose that J is not of the form $J^* + ce$. Then,

$$\begin{aligned}
T^{(\lambda)}J &= T^{(\lambda)}J^* + P^{(\lambda)}(J-J^*)\\
&= J^* + P^{(\lambda)}(J-J^*)\\
&\neq J^* + (J-J^*)\\
&= J,
\end{aligned}$$

where the inequality follows from Lemma 2. **q.e.d.**

We next set out to characterize the "steady-state" expectations of $A(X_t)$ and $b(X_t)$. While this can be done by taking limits of expectations as t goes to infinity, it is simpler to characterize expectations of a process that is already in steady-state. We therefore make a short digression to construct a stationary version of X_t.

We proceed as follows. Let $\{i_t | -\infty < t < \infty\}$ 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

$$z_t = \sum_{\tau=-\infty}^{t} \lambda^{t-\tau} \phi(i_\tau). \tag{6}$$

Note that z_t is constructed by taking the stationary process $\phi(i_t)$, whose magnitude is bounded by a constant, and passing it through an exponentially stable linear time invariant filter. The output z_t of this filter is stationary and its magnitude is bounded by a constant (the same constant applies to all sample paths). 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 process constructed in Subsection 3.2. Furthermore, the state space of this process, which we will denote by $\mathcal{S}$, is bounded. We can now identify $E_0[\,\cdot\,]$ with the expectation with respect to the invariant distribution of this process.

Let us now provide a lemma, characterizing the steady–state expectations of several expressions of interest. We omit the proof, since it would follow the same steps as that of Lemma 7 in (Tsitsiklis and Van Roy, 1997).

Lemma 4 *Under Assumption 1, the following relations hold:*
(a) $E_0[z_t \phi'(i_t)] = \sum_{m=0}^{\infty} \lambda^m \Phi' D P^m \Phi$,
(b) $E_0[z_t \phi'(i_{t+1})] = \sum_{m=0}^{\infty} \lambda^m \Phi' D P^{m+1} \Phi$,
(c) $E_0[z_t] = \frac{1}{1-\lambda} \Phi' D e$.
(d) $E_0[z_t g(i_t)] = \sum_{m=0}^{\infty} \lambda^m \Phi' D P^m g$.

Recall that the TD(λ) algorithm can be written as

$$\theta_{t+1} = \theta_t + \gamma_t (A(X_t)\theta_t + b(X_t)),$$

as explained in Subsection 3.2. The following lemma characterizes the steady–state expectations $E_0[A(X_t)]$ and $E_0[b(X_t)]$, which we will denote by A and b.

Lemma 5 *Under Assumption 1, the steady-state expectations* $A = E_0[A(X_t)]$ *and* $b = E_0[b(X_t)]$ *are given by*

$$A = \begin{bmatrix} -c & 0 \cdots 0 \\ -\frac{1}{1-\lambda} \Phi' D e & \Phi' D (P^{(\lambda)} - I) \Phi \end{bmatrix},$$

and

$$b = \begin{bmatrix} c\mu^* \\ \Phi' D (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \sum_{t=0}^{m} P^t g \end{bmatrix}.$$

Proof: Using Lemma 4, and the relation

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

we have

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

Since A is given by

$$A = \begin{bmatrix} -c & 0 \cdots 0 \\ -E_0[z_t] & E_0[z_t(\phi'(i_{t+1}) - \phi'(i_t))] \end{bmatrix},$$

this establishes the desired characterization of A. As for the case of b, using Lemma 4, we have

$$\begin{aligned} E_0[z_t g(i_t)] = &= \sum_{m=0}^{\infty} \lambda^m \Phi' D P^m g \\ &= \Phi' D(1-\lambda)\sum_{m=0}^{\infty} \lambda^m \sum_{t=0}^{m} P^t g. \end{aligned}$$

Combining this with the fact that

$$b = \begin{bmatrix} cE_0[g(i_t)] \\ E_0[z_t g(i_t)] \end{bmatrix},$$

completes the proof. **q.e.d.**

The following lemma establishes that the expectations of $A(X_t)$ and $b(X_t)$ converge to their steady–state values quickly. This fact is used in the stochastic approximation result we will employ.

Lemma 6 *Under Assumption 1, there exist scalars C and $\rho \in (0,1)$ such that for any $X_0 \in \mathcal{S}$ and $t \geq 0$, we have*

$$\Big\| E[A(X_t)|X_0] - A \Big\| \leq C\rho^t,$$

and

$$\Big\| E[b(X_t)|X_0] - b \Big\| \leq C\rho^t.$$

Proof: It is well known that for any irreducible aperiodic finite state Markov chain, there exist scalars C and $\rho \in (0,1)$ such that

$$|\Pr(i_t = i|i_0) - \pi(i)| \leq C\rho^t, \qquad \forall i_0, i \in S, t \geq 0.$$

Using this fact it is easy to show that there exist scalars C and $\rho \in (0,1)$ such that

$$\Big\| E[z(i_t)\phi'(i_{t+m})|i_0] - E_0[z(i_t)\phi'(i_{t+m})] \Big\| \leq C\rho^t, \qquad \forall i_0 \in S, m \geq 0, t \geq 0,$$

and

$$\Big\| E[z(i_t)|i_0] - E_0[z(i_t)] \Big\| \leq C\rho^t, \qquad \forall i_0 \in S, t \geq 0.$$

(The details of this argument are straightforward, and can be found in (Tsitsiklis and Van Roy, 1997).) The result pertaining to the matrix A follows. The proof for the case involving the vector b involves similar arguments. **q.e.d.**

The matrix $\Phi' D(P^{(\lambda)} - I)\Phi$ can be shown to be negative definite, and this fact formed the basis for the convergence analysis of discounted cost TD(λ) in (Tsitsiklis and Van Roy, 1997). In the context of average cost TD(λ), the matrix A is not necessarily negative definite, and a slightly different property must be used. In particular, the next lemma shows that the matrix A becomes negative definite under an appropriate coordinate scaling.

Lemma 7 *Under Assumptions 1 and 2, there exists a diagonal matrix L with positive diagonal entries, such that the matrix LA is negative definite.*

Proof: We begin by showing that the matrix $\Phi' D(P^{(\lambda)} - I)\Phi$ is negative definite. By Lemma 2, for any $J \notin \{ce | c \in \Re\}$, we have $\|P^{(\lambda)}J\|_D \leq \|J\|_D$ and $P^{(\lambda)}J \neq J$. In other words, for any $J \notin \{ce | c \in \Re\}$, J and $P^{(\lambda)}J$ are distinct elements of the set $\{\overline{J} \mid \|\overline{J}\|_D \leq \|J\|\}$. Hence, given a vector $J \notin \{ce | c \in \Re\}$, there are three possibilities: (a) $\|P^{(\lambda)}J\|_D < \|J\|_D$; (b) $\|P^{(\lambda)}J\|_D = \|J\|_D$ and the vectors $P^{(\lambda)}J$ and J are not collinear; (c) $\|P^{(\lambda)}J\|_D = \|J\|_D$ and $P^{(\lambda)}J = -J$. In case (a) we have

$$J'DP^{(\lambda)}J \leq \|J\|_D \|P^{(\lambda)}J\|_D < \|J\|_D^2 = J'DJ,$$

while in cases (b) and (c) we have

$$J'DP^{(\lambda)}J < \|J\|_D \|P^{(\lambda)}J\|_D \leq \|J\|_D^2 = J'DJ.$$

Therefore, for all $J \notin \{ce | c \in \Re\}$,

$$J'D(P^{(\lambda)} - I)J < 0.$$

Since the columns of Φ are linearly independent and do not span $\{ce | c \in \Re\}$ (Assumption 2), the negative definiteness of $\Phi' D(P^{(\lambda)} - I)\Phi$ follows.

Let L be a diagonal matrix with the first diagonal entry equal to some scalar $\ell > 0$ and every other diagonal entry equal to one. Then, given a vector

$$\theta = \begin{bmatrix} \mu \\ r \end{bmatrix},$$

we have

$$\theta' LA\theta = -\ell c\mu^2 - \frac{1}{1-\lambda}\mu e' D\Phi r + r'\Phi' D(P^{(\lambda)} - I)\Phi r.$$

Note that

$$\frac{1}{1-\lambda}|e' D\Phi r| \leq \frac{1}{1-\lambda}\|e\|_D \|\Phi r\|_D = \frac{1}{1-\lambda}\|\Phi r\|_D \leq C_1 \|r\|,$$

for some constant $C_1 > 0$, and since $\Phi' D(P^{(\lambda)} - I)\Phi$ is negative definite,

$$r'\Phi' D(P^{(\lambda)} - I)\Phi r \leq -C_2 \|r\|^2,$$

for some constant $C_2 > 0$. It follows that

$$\theta' LA\theta \leq -\ell c\mu^2 + C_1 \mu \|r\| - C_2 \|r\|^2,$$

and by setting ℓ to a value satisfying $C_1^2 < 4\ell c C_2$, we have

$$\theta' LA\theta < 0,$$

for any $\theta \neq 0$. **q.e.d.**

3.3 A Result on Stochastic Approximation

To establish convergence of TD(λ) based on the steady–state dynamics, we rely on results from stochastic approximation theory. The following Theorem (Proposition 4.8 from page 174 of (Bertsekas and Tsitsiklis, 1996)) is a special case of a very general result (Theorem 17 on page 239 of (Benveniste et al., 1987)), and it provides the basis for a corollary that will suit our needs.

Theorem 2 *Consider an iterative algorithm of the form*

$$\theta_{t+1} = \theta_t + \gamma_t(A(X_t)\theta_t + b(X_t)),$$

where:
(a) The step sizes γ_t are positive, deterministic, and satisfy $\sum_{t=0}^{\infty}\gamma_t = \infty$ and $\sum_{t=0}^{\infty}\gamma_t^2 < \infty$.
(b) The Markov process X_t, which evolves in a state space $\mathcal{S}$, has an invariant (steady–state) distribution. Let $E_0[\cdot]$ stand for the expectation with respect to this invariant distribution.
(c) The matrix A defined by $A = E_0[A(X_t)]$ is negative definite.
(d) There exists a constant C such that $\|A(X)\| \leq C$ and $\|b(X)\| \leq C$, for all $X \in \mathcal{S}$.
(e) There exist scalars C and $\rho \in (0,1)$ such that

$$\Big\|E[A(X_t)|X_0 = X] - A\Big\| \leq C\rho^t, \qquad \forall t \geq 0, X \in \mathcal{S},$$

and

$$\Big\|E[b(X_t)|X_0 = X] - b\Big\| \leq C\rho^t, \qquad \forall t \geq 0, X \in \mathcal{S},$$

where $b = E_0[b(X_t)]$.
Then, θ_t converges to θ^*, with probability 1, where θ^* is the unique vector that satisfies $A\theta^* + b = 0$.

We next state and prove the following corollary, which suits the needs of our proof of Theorem 1. Note that the only difference between this corollary and Theorem 2 is in Condition (c) pertaining to negative definiteness.

Corollary 1 *The conclusions of Theorem 2 remain valid if Condition (c) is replaced by the following condition:*
(c') Let the matrix A be defined by $A = E_0[A(X_t)]$. There exists a diagonal matrix L with positive diagonal entries such that LA is negative definite.

Proof: We first note that LA is negative definite if and only if $L^{\frac{1}{2}}AL^{-\frac{1}{2}}$ is negative definite. This follows from the fact that

$$\theta' LA\theta = \tilde{\theta}' L^{\frac{1}{2}} AL^{-\frac{1}{2}}\tilde{\theta},$$

for any θ, where $\tilde{\theta} = L^{\frac{1}{2}}\theta$, or for any $\tilde{\theta}$, where $\theta = L^{-\frac{1}{2}}\tilde{\theta}$.

Let $\tilde{\theta}_t = L^{\frac{1}{2}}\theta_t$. We then have

$$\tilde{\theta}_{t+1} = \tilde{\theta}_t + \gamma_t\Big(L^{\frac{1}{2}}A(X_t)L^{-\frac{1}{2}}\tilde{\theta}_t + L^{\frac{1}{2}}b(X_t)\Big).$$

The steady–state expectation $L^{\frac{1}{2}}AL^{-\frac{1}{2}} = E_0[L^{\frac{1}{2}}A(X_t)L^{-\frac{1}{2}}]$ is negative definite because LA is negative definite. It follows from Theorem 2 that $\tilde{\theta}_t$ converges to the vector $\tilde{\theta}^*$ that solves $L^{\frac{1}{2}}AL^{-\frac{1}{2}}\tilde{\theta}^* + L^{\frac{1}{2}}b = 0$. Therefore, θ_t converges to θ^*, and we have $A\theta^* + b = 0$. **q.e.d.**

3.4 Proof of Theorem 1

Let us verify that the conditions of Corollary 1 are satisfied by the TD(λ) algorithm. Condition (a) is satisfied by the requirements on step sizes that we made for TD(λ). We have already discussed condition (b), and validity of condition (c') is established by Lemma 7. Note that z_t, $g(i_t)$, and $\phi(i_t)$, are bounded since i_t lies in a finite state space S. Hence, it is easy to show that $A(X_t)$ and $b(X_t)$ are bounded, satisfying condition (d). Validity of condition (e) is established by Lemma 6. Since all conditions are satisfied, the limit of convergence θ^* of the TD(λ) algorithm satisfies

$$A\theta^* + b = 0,$$

with probability 1.

Invoking Lemma 5, we recall that $b(1) = c\mu^*$, and observe that $(A\theta^*)(1) = -c\theta^*(1)$. We therefore have $\theta^*(1) = \mu^*$, i.e., the sequence μ_t converges to μ^*. Let the vector $r^* \in \Re^n$ be given by $r^* = (\theta^*(2), \ldots, \theta^*(n+1))'$. Then, using Lemmas 3 and 5, the relation $1/(1-\lambda) = (1-\lambda)\sum_{m=0}^{\infty} \lambda^m (m+1)$, and the equation $A\theta^* + b = 0$, we obtain

$$\begin{aligned}
-\Phi' D(1-\lambda) \sum_{m=0}^{\infty} \lambda^m \sum_{t=0}^{m} P^t g &= \Phi' D(P^{(\lambda)} - I)\Phi r^* - \frac{\mu^*}{1-\lambda}\Phi' D e, \\
-\Phi' D(1-\lambda) \sum_{m=0}^{\infty} \lambda^m \sum_{t=0}^{m} P^t g &= \Phi' D(P^{(\lambda)} - I)\Phi r^* - \Phi' D(1-\lambda) \sum_{m=0}^{\infty} \lambda^m \sum_{t=0}^{m} \mu^* e, \\
\Phi' D \Phi r^* &= \Phi' D \left(P^{(\lambda)} \Phi r^* + (1-\lambda) \sum_{m=0}^{\infty} \lambda^m \sum_{t=0}^{m} P^t (g - \mu^* e) \right), \\
\Phi' D \Phi r^* &= \Phi' D T^{(\lambda)}(\Phi r^*), \\
\Phi (\Phi' D \Phi)^{-1} \Phi' D \Phi r^* &= \Phi (\Phi' D \Phi)^{-1} \Phi' D T^{(\lambda)}(\Phi r^*), \\
\Phi r^* &= \Pi T^{(\lambda)}(\Phi r^*).
\end{aligned}$$

This completes the proof. **q.e.d.**

4 Approximation Error

We have proven that the sequence r_t generated by TD(λ) converges to r^*, which uniquely solves $\Pi T^{(\lambda)}(\Phi r^*) = \Phi r^*$. However, we have not commented on the quality of this final approximation. In this section, we propose a definition of approximation error, study a few of its properties, and derive error bounds.

4.1 A Definition of Error

In our analysis of discounted cost TD(λ) (Tsitsiklis and Van Roy, 1997), we employed the error metric $\|\Phi r^* - J^*\|_D$, where J^* was the cost–to–go function for a discounted Markov chain. This formulation enabled the development of a graceful error bound. In the context of average cost TD(λ), one might proceed similarly and define the error to be $\|\Phi r^* - J^*\|_D$, where J^* is the basic differential cost function. However, because we would be content with a good approximation of *any* differential cost function, and not just the *basic* one,

this definition may not accurately reflect our preferences. In particular, there may exist a parameter vector $\overline{r}$ such that $\|\Phi\overline{r} - J\|_D$ is very small for some differential cost function J, while $\|\Phi r - J^*\|_D$ is large for all r. To accommodate this possibility, we will employ as our definition of approximation error the infimum of the weighted Euclidean distance from the set of all differential cost functions, which is given by the following expression:

$$\inf_{J \in \{J^* + ce | c \in \Re\}} \|\Phi r^* - J\|_D = \inf_{c \in \Re} \|\Phi r^* - (J^* + ce)\|_D.$$

In addition to catering intuitive appeal, this definition will lead to a graceful error bound.

4.2 An Alternative Characterization

The error metric provided in the previous subsection can be expressed in a form that does not involve infimization. To derive this alternative representation, we first note that any vector $J \in \Re^n$ can be decomposed into a component $\mathcal{P}J$ that is D-orthogonal to e, and a component $(I - \mathcal{P})J$ that is a multiple of e, where $\mathcal{P}$ is the projection matrix defined by

$$\mathcal{P} = I - ee'D.$$

It is easily checked that

$$\mathcal{P} = I - e\pi' = I - \lim_{t \to \infty} P^t.$$

This implies that P and $\mathcal{P}$ commute (i.e., $P\mathcal{P} = \mathcal{P}P$). By definition of J^*, we have

$$e'DJ^* = \pi' J^* = 0.$$

It follows that $\mathcal{P}J^* = J^*$. Since the minimum distance of the vector $\Phi r^* - J^*$ from the subspace $\{ce | c \in \Re\}$ is equal to the magnitude of the projection onto the orthogonal complement of the subspace, we have

$$\inf_{c \in \Re} \|\Phi r^* - (J^* + ce)\|_D = \|\mathcal{P}\Phi r^* - J^*\|_D.$$

From here on, we will use this simpler characterization rather than the original definition of approximation error.

4.3 A Decomposition of Basis Functions

The projection introduced in the previous subsection can be applied to each basis function ϕ_k to obtain the function $\mathcal{P}\phi_k$, which is D–orthogonal to e. In this subsection, we show that replacing each ϕ_k by $\mathcal{P}\phi_k$ does not change the limit to which TD(λ) converges or the resulting approximation error.

Recall that TD(λ) converges to the unique solution r^* of the equation $\Pi T^{(\lambda)}(\Phi r^*) = \Phi r^*$. Let

$$\overline{\Phi} = \mathcal{P}\Phi,$$

and note that $\overline{\Phi}$ replaces Φ, if each basis functions ϕ_k is replaced by $\mathcal{P}\phi_k$. When the basis functions $\mathcal{P}\phi_1, \ldots, \mathcal{P}\phi_K$ are employed, TD(λ) converges to a vector $\overline{r}$ that satisfies

$$\overline{\Pi} T^{(\lambda)}(\overline{\Phi}\overline{r}) = \overline{\Phi}\overline{r},$$

where the matrix $\overline{\Pi}$ is defined by

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

We will now show that $r^* = \overline{r}$.

Using the definition of $T^{(\lambda)}$ and the property $e'DP = \pi'P = \pi'$, it is easily verified that for any r,

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

By the fixed point equation $\Pi T^{(\lambda)}(\Phi r^*) = \Phi r^*$, we also have

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

for each basis function ϕ_k. It follows that for any projected basis function $\overline{\phi}_k = \mathcal{P}\phi_k$, there is a scalar c such that

$$\overline{\phi}_k' D(T^{(\lambda)}(\Phi r^*) - \Phi r^*) = (\phi_k + ce)' D(T^{(\lambda)}(\Phi r^*) - \Phi r^*) = 0.$$

The fact that

$$T^{(\lambda)}(\overline{\Phi} r^*) = T^{(\lambda)}(\Phi r^* + \hat{c}e) = T^{(\lambda)}(\Phi r^*) + \hat{c}e,$$

for some constant $\hat{c}$, then leads to the conclusion that

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

Hence, $\overline{\Pi} T^{(\lambda)}(\overline{\Phi} r^*) = \overline{\Phi} r^*$ and $r^* = \overline{r}$.

4.4 Mixing Times

In the next subsection, we will provide a bound on the error associated with the limiting weight vector r^*. Central to the development of this bound will be the notion of a "mixing time," which represents the time it takes for a Markov process to reach steady state. In this section, we motivate the relationship between mixing times and approximation error, and we will define "mixing factors" that will be used in the statement of our bound.

We begin by presenting some intuition concerning the bound that was established in (Tsitsiklis and Van Roy, 1997) in the context of discounted cost temporal–difference learning. To simplify the exposition here, we will focus on the case of $\lambda = 0$. The operator $T^{(0)}$ for discounted cost Markov chains is defined by $T^{(0)}J = g + \alpha PJ$, where $\alpha \in (0, 1)$ is the discount factor. To reduce notation, let $T = T^{(0)}$. This operator can be used to compute expected costs over a finite number of time steps. In particular, the expected (discounted) cost over k time steps, starting at a state i, is given by $(T^k 0)(i)$, and $T^k 0 \to J^*$, where J^* is the infinite–horizon cost–to–go function. Now consider approximating $T^k 0$, for each k, using $(\Pi T)^k 0$. This approximation can be thought of as the result of a sequence of approximations – each iterate generated by an application of T is approximated by projecting onto the subspace spanned by the basis functions. Because error is introduced at each iteration, one might conjecture that the error accumulates and diverges as k grows. However this is not the case, and $(\Pi T)^k 0$ actually converges to an approximation of J^*. In (Tsitsiklis and Van Roy, 1997), the fact that the future costs are discounted was instrumental in the development of an error bound. In particular, the discount factor reduces the effects of

the errors accumulated over repeated applications of ΠT. One way of understanding this phenomenon is by noting that

$$(\Pi T)^k 0 = \sum_{t=0}^{k-1} (\alpha \Pi P)^t g,$$

and that the terms in the summation that involve many repeated projections are heavily discounted. This discounting keeps the accumulated error bounded.

In average cost TD(0), there is no discount factor, so the development of an error bound must rely on other properties of the Markov chain. The fact that a Markov chain settles into steady state serves this purpose. In particular, accurate approximations of expected future costs for times beyond the mixing time of the Markov chain are unnecessary. This is because the Markov chain will be in steady state, and therefore, expected per-stage costs are virtually equal to the average cost of the Markov chain, regardless of the current state. This phenomenon plays a role analogous to that of discounting in the context of discounted Markov chains.

We now discuss some possible characterizations of mixing. Let J be some function defined on the state space. Mixing can be viewed as an assumption that $E[J(i_t)|i_0]$ converges to $E_0[J(i_t)]$ at the rate of α^t, where the "mixing factor" $\alpha \in [0,1)$ is a constant that captures the rate at which mixing occurs. In fact, $(I - \mathcal{P})J$ is aligned with e and is immaterial to our context, whereas $E[(\mathcal{P}J)(i_t) \mid i_0]$ converges to zero as t approaches infinity. Thus, one possible assumption could be that $E[(\mathcal{P}J)(i_t) \mid i_0]$ decreases like α^t, for all functions J. In terms of the transition probability matrix P, this would be captured by an assumption that $\|\mathcal{P}P\|_D \leq \alpha$.

For the purposes of our error bounds, we do not need every possible function J to converge rapidly to steady-state. Rather, it suffices to consider only those functions that are representable by our approximation architecture, i.e., linear combinations of the basis functions ϕ_k. We can capture this effect by projecting, using the projection matrix $\overline{\Pi}$, and place an assumption on the induced norm $\|\overline{\Pi}\mathcal{P}P\|_D$, which is actually the same as $\|\overline{\Pi}P\|_D$ since $\overline{\Pi}\mathcal{P} = \overline{\Pi}$ (this follows from the fact that $\overline{\Pi}$ projects onto a subspace of the range onto which $\mathcal{P}$ projects).

Finally, it turns out that an even weaker assumption will do, using the following idea. Given any $\delta \in (0,1)$, we define an auxiliary Markov chain with a transition matrix $P_\delta = I + \delta(P - I)$ and a cost function $g_\delta = \delta g$. The basic differential cost function for this Markov chain remains unchanged. This is because

$$\delta g - \delta \mu^* e + (I + \delta(P - I))J^* = \delta(g - \mu^* e + PJ^*) + (1 - \delta)J^* = J^*.$$

Similarly, it is easy to show that TD(0) generates the same limit of convergence for this auxiliary Markov chain as it did for the original one. In this spirit, we can consider $\|\overline{\Pi}P_\delta\|_D$ as the relevant mixing factor. Furthermore, since δ is arbitrary, we can obtain the tightest possible bound by minimizing over all possible choices of δ.

For the more general case of $\lambda \in [0,1)$, the pertinent mixing time is that of the stochastic matrix $P^{(\lambda)} = (1 - \lambda)\sum_{m=0}^{\infty} \lambda^m P^{m+1}$. (Note that $P^{(0)} = P$, which brings us back to our previous discussion concerning the case of $\lambda = 0$.) Similar to the context of TD(0), we define $P_\delta^{(\lambda)} = I + \delta(P^{(\lambda)} - I)$, and we define a scalar α_λ for each $\lambda \in [0,1)$ by

$$\alpha_\lambda = \inf_{\delta > 0} \|\overline{\Pi}P_\delta^{(\lambda)}\|_D.$$

This mixing factor will be used to establish our error bound.

4.5 The Error Bound

We now state a theorem that provides a bound on approximation error. A proof is provided in the next subsection.

Theorem 3 *Let Assumptions 1 and 2 hold. For each $\lambda \in [0,1)$, let $r_\lambda^* \in \Re^K$ be the vector satisfying*

$$\Phi r_\lambda^* = \Pi T^{(\lambda)}(\Phi r_\lambda^*).$$

Then:
(a) For each $\lambda \in [0,1)$, the mixing factor α_λ is in $[0,1)$ and $\lim_{\lambda \uparrow 1} \alpha_\lambda = 0$.
(b) The following bound holds:

$$\|\mathcal{P}\Phi r_\lambda^* - J^*\|_D \leq \frac{1}{\sqrt{1-\alpha_\lambda^2}} \inf_{r \in \Re^K} \|\mathcal{P}\Phi r - J^*\|_D.$$

Note that the bound is a multiple of

$$\inf_{r \in \Re^K} \|\mathcal{P}\Phi r - J^*\|_D,$$

which is the minimal error possible given the fixed set of basis functions. This term becomes zero if there exists a parameter vector r and a scalar c for which $\Phi r = J^* + ce$, that is, if our "approximation architecture" is capable of representing exactly some differential cost function.

The term $1/\sqrt{1-\alpha_\lambda^2}$ decreases as α_λ decreases. Hence, the term is guaranteed to approach its optimal value of 1 as λ approaches 1. This suggests that larger values of λ may lead to lower approximation error.

4.6 Proof of Theorem 3

We begin by establishing part (a) of the theorem. Since α_λ is the infimum of a set of nonnegative reals, $\alpha_\lambda \geq 0$. From Lemma 2, we have $\|P^{(\lambda)}\|_D \leq 1$ and $P^{(\lambda)}J \neq J$ if J is not proportional to e. It follows that for any $\delta \in (0,1)$ and any J that is not proportional to e, we have

$$\|\mathcal{P}P_\delta^{(\lambda)}J\|_D = \|\mathcal{P}(\delta P^{(\lambda)}J + (1-\delta)J)\|_D < \|\mathcal{P}J\|_D \leq \|J\|_D.$$

(This is because $\mathcal{P}J$ and $\mathcal{P}P^{(\lambda)}J$ are distinct elements of $\{\mathcal{P}\overline{J} \mid \|\mathcal{P}\overline{J}\|_D \leq \|\mathcal{P}J\|_D\}$, so their strictly convex combination cannot be an extreme point.) Note that $\|\mathcal{P}P_\delta^{(\lambda)}J\|_D$ is a continuous function of J and that the set $\{J \mid \|J\|_D \leq 1\}$ is compact. It follows from Weierstrass' theorem that for any $\delta \in (0,1)$, $\|\mathcal{P}P_\delta^{(\lambda)}\|_D < 1$. Since $\overline{\Pi} = \overline{\Pi}\mathcal{P}$, we then have

$$\alpha_\lambda = \inf_{\delta > 0} \|\overline{\Pi}P_\delta^{(\lambda)}\|_D \leq \inf_{\delta > 0} \|\mathcal{P}P_\delta^{(\lambda)}\|_D \leq \inf_{\delta \in (0,1)} \|\mathcal{P}P_\delta^{(\lambda)}\|_D < 1.$$

As for the limit as λ approaches 1, we have

$$\lim_{\lambda \uparrow 1} \alpha_\lambda = \lim_{\lambda \uparrow 1} \inf_{\delta > 0} \|\overline{\Pi}P_\delta^{(\lambda)}\|_D \leq \lim_{\lambda \uparrow 1} \|\overline{\Pi}P^{(\lambda)}\|_D \leq \lim_{\lambda \uparrow 1} \|\mathcal{P}P^{(\lambda)}\|_D.$$

Assumption 1 implies that

$$\lim_{t\to\infty} \|\mathcal{P}P^t\|_D = 0.$$

It follows that

$$\lim_{\lambda\uparrow 1} \|\mathcal{P}P^{(\lambda)}\|_D = \lim_{\lambda\uparrow 1} \left\| (1-\lambda)\sum_{t=0}^{\infty} \lambda^t \mathcal{P}P^{t+1} \right\|_D = 0.$$

This completes the proof for part (a).

Let $T_\delta^{(\lambda)} = (1-\delta)I + \delta T^{(\lambda)}$. It is easy to see that $T_\delta^{(\lambda)} J^* = J^*$ and $\overline{\Pi} T_\delta^{(\lambda)}(\overline{\Phi} r_\lambda^*) = \overline{\Phi} r_\lambda^*$. For any nonnegative scalar δ, we have

$$\begin{aligned}
\|\mathcal{P}\Phi r_\lambda^* - J^*\|_D^2 &= \|\overline{\Phi} r_\lambda^* - J^*\|_D^2 \\
&= \|\overline{\Pi} T_\delta^{(\lambda)}(\overline{\Phi} r_\lambda^*) - T_\delta^{(\lambda)} J^*\|_D^2 \\
&\leq \|\overline{\Pi} T_\delta^{(\lambda)}(\overline{\Phi} r_\lambda^*) - \overline{\Pi} T_\delta^{(\lambda)} J^*\|_D^2 + \|T_\delta^{(\lambda)} J^* - \overline{\Pi} T_\delta^{(\lambda)} J^*\|_D^2 \\
&= \|\overline{\Pi} P_\delta^{(\lambda)}(\overline{\Phi} r_\lambda^*) - \overline{\Pi} P_\delta^{(\lambda)} J^*\|_D^2 + \|J^* - \overline{\Pi} J^*\|_D^2 \\
&\leq \|\overline{\Pi} P_\delta^{(\lambda)}\|_D^2 \|\overline{\Phi} r_\lambda^* - J^*\|_D^2 + \|J^* - \overline{\Pi} J^*\|_D^2.
\end{aligned}$$

Since δ is an arbitrary nonnegative scalar, we have

$$\|\mathcal{P}\Phi r_\lambda^* - J^*\|_D^2 \leq \alpha_\lambda^2 \|\overline{\Phi} r_\lambda^* - J^*\|_D^2 + \|J^* - \overline{\Pi} J^*\|_D^2,$$

and it follows that

$$\|\mathcal{P}\Phi r_\lambda^* - J^*\|_D \leq \frac{1}{\sqrt{1-\alpha_\lambda^2}} \|J^* - \overline{\Pi} J^*\|_D.$$

Since

$$\|J^* - \overline{\Pi} J^*\|_D = \inf_r \|\mathcal{P}\Phi r - J^*\|_D,$$

this completes the proof for part (b).

5 Using a Fixed Average Cost Estimate

Recall that the basic differential cost function is defined by

$$J^* = \sum_{t=0}^{\infty} P^t (g - \mu^* e).$$

When $\mu_t = \mu^*$, the TD(λ) algorithm essentially tunes r_t to approximate this function. On the other hand, when $\mu_t \neq \mu^*$, the algorithm tunes r_t as if

$$\sum_{t=0}^{\infty} P^t (g - \mu_t e)$$

were to be approximated. This series diverges since

$$\lim_{t\to\infty} P^t (g - \mu e) \neq 0,$$

for $\mu \neq \mu^*$. The fact that TD(λ) converges at all may therefore seem strange. It turns out that Assumption 2(b), which prevents the basis functions from spanning the space

of constant functions, is instrumental in preventing divergence here. In some sense, the absence of this subspace "filters out" effects that would lead to divergence when $\mu_t \neq \mu^*$. In this section, we present and analyze a variant of TD(λ) that sheds some additional light on this phenomenon.

We consider the use of a fixed average cost estimate μ instead of the adapted sequence μ_t. Only the weight vector r_t is updated. In particular, we have

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

where the temporal–difference d_t is given by

$$d_t = (g(i_t) - \mu) + \phi'(i_{t+1})r_t - \phi'(i_t)r_t,$$

and the eligibility vectors are updated according to

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

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

Recall that the error bound associated with the original algorithm (Theorem 3) involved a mixing factor

$$\alpha_\lambda = \inf_{\delta \geq 0} \|\overline{\Pi} P_\delta^{(\lambda)}\|.$$

In addition to this mixing factor, the bound for our new algorithm will depend upon a second mixing factor

$$\beta_\lambda = \inf_{\delta \in [0,1]} \|\Pi P_\delta^{(\lambda)}\|.$$

This mixing factor is similar in spirit to α_λ, but involves a projection onto the range of Φ instead of $\overline{\Phi}$. The restriction of δ to values less than or equal to one simplifies the upcoming bound and proof.

We have the following theorem establishing convergence and error bounds for TD(λ) with a fixed average cost estimate:

Theorem 4 *Under Assumptions 1, 2, and 3, for any $\lambda \in [0,1)$, the following hold:*
(a) The TD(λ) algorithm with a fixed average cost estimate, as defined above, converges with probability 1.
(b) The limit of convergence $\overline{r}_\lambda$ is the unique solution of the equation

$$\Pi T^{(\lambda)}(\Phi \overline{r}_\lambda) + \frac{\mu^* - \mu}{1 - \lambda} \Pi e = \Phi \overline{r}_\lambda.$$

(c) For any $\lambda \in [0,1)$, the mixing factor β_λ is in $[0,1)$, and $\lim_{\lambda \uparrow 1} \beta_\lambda = 0$.
(d) The limit of convergence $\overline{r}_\lambda$ satisfies

$$\|\mathcal{P}\Phi \overline{r}_\lambda - J^*\|_D \leq \frac{1}{\sqrt{1 - \alpha_\lambda^2}} \inf_{r \in \Re^K} \|\mathcal{P}\Phi r - J^*\|_D + \frac{|\mu^* - \mu|}{(1 - \beta_\lambda)(1 - \lambda)} \|\Pi e\|_D,$$

where α_λ and $\mathcal{P}$ are defined as in Section 5.

There are two somewhat unrelated terms involved in the bound of Theorem 4. The first term is equal to the error bound of Theorem 3, and can be viewed as error brought about by the choice of basis functions. The second term is proportional to the error in the average cost estimate. The term is also proportional to $\|\Pi e\|_D$, which is zero if the space spanned by the basis functions is D–orthogonal to e. The dependence on λ and β_λ is a little more complicated. If either λ or β_λ approaches one, the coefficient approaches infinity. In contrast to the discussion in the preceding section, we now have a situation where values of λ close to 1 cease to be preferable.

Our proof of Theorem 4 relies on the same ideas that were used to prove Theorems 1 and 3. However, this proof will actually be simpler, since we no longer have to deal with the sequence μ_t. Our presentation of the proof will be brief, relying heavily on the recollection of arguments used to prove Theorems 1 and 3.

Proof of Theorem 4

The new variant of TD(λ) can be written in the form

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

where $X_t = (i_t, i_{t+1}, z_t)$, and the matrix and vector valued functions $A(\cdot)$ and $b(\cdot)$ are defined by

$$A(X) = z(\phi'(j) - \phi'(i))$$

and

$$b(X) = z(g(i) - \mu),$$

for any $X = (i, j, z)$. Letting A and b represent steady–state expectations of these functions, it is easy to show that

$$A = \Phi' D(P^{(\lambda)} - I)\Phi$$

and

$$b = \Phi' D(1-\lambda) \sum_{m=0}^{\infty} \lambda^m \sum_{t=0}^{m} P^t(g - \mu e),$$

the proof being essentially the same as that of Lemma 5. The results of Lemma 6, concerning the rate of convergence of the expectations to their steady–state values, remain valid. Part of the proof of Lemma 7 shows that the matrix A (as defined in this section) is negative definite. Given these facts, it follows from Theorem 2 that r_t converges to a vector $\overline{r}_\lambda$ that is the unique solution to

$$A\overline{r}_\lambda + b = 0.$$

Based on this equation, some simple algebra establishes that

$$\Phi\overline{r}_\lambda = \Pi T^{(\lambda)}(\Phi\overline{r}_\lambda) + \frac{\mu^* - \mu}{1-\lambda}\Pi e.$$

This completes the proof of parts (a) and (b).

Using arguments similar to the proof of Theorem 3, it can be shown that for any $\lambda \in [0, 1)$, β_λ is in $[0, 1)$, and that $\lim_{\lambda \uparrow 1} \beta_\lambda = 0$, establishing part (c) of the theorem.

We will now derive the error bound. As in previous sections, we let r_λ^* denote the unique vector satisfying $\Phi r_\lambda^* = \Pi T^{(\lambda)}(\Phi r_\lambda^*)$. For any $\delta \in [0,1]$, we have

$$\begin{aligned}
\|\Phi\bar{r}_\lambda - \Phi r_\lambda^*\|_D &= \|(1-\delta)\Phi\bar{r}_\lambda + \delta\Phi\bar{r}_\lambda - (1-\delta)\Phi r_\lambda^* - \delta\Phi r_\lambda^*\|_D \\
&= \left\|(1-\delta)\Pi\Phi\bar{r}_\lambda + \delta\left(\Pi T^{(\lambda)}(\Phi\bar{r}_\lambda) + \frac{\mu^*-\mu}{1-\lambda}\Pi e\right) - (1-\delta)\Pi\Phi r_\lambda^* - \delta\Pi T^{(\lambda)}(\Phi r_\lambda^*)\right\|_D \\
&= \left\|\Pi T_\delta^{(\lambda)}(\Phi\bar{r}_\lambda) + \frac{\delta(\mu^*-\mu)}{1-\lambda}\Pi e - \Pi T_\delta^{(\lambda)}(\Phi r_\lambda^*)\right\|_D \\
&\leq \left\|\Pi T_\delta^{(\lambda)}(\Phi\bar{r}_\lambda) - \Pi T_\delta^{(\lambda)}(\Phi r_\lambda^*)\right\|_D + \frac{|\mu^*-\mu|}{1-\lambda}\|\Pi e\|_D \\
&= \left\|\Pi P_\delta^{(\lambda)}(\Phi\bar{r}_\lambda) - \Pi P_\delta^{(\lambda)}(\Phi r_\lambda^*)\right\|_D + \frac{|\mu^*-\mu|}{1-\lambda}\|\Pi e\|_D \\
&\leq \|\Pi P_\delta^{(\lambda)}\|_D\|\Phi\bar{r}_\lambda - \Phi r_\lambda^*\|_D + \frac{|\mu^*-\mu|}{1-\lambda}\|\Pi e\|_D.
\end{aligned}$$

Since δ is an arbitrary scalar in $[0,1]$, we have

$$\|\Phi\bar{r}_\lambda - \Phi r_\lambda^*\|_D \leq \beta_\lambda\|\Phi\bar{r}_\lambda - \Phi r_\lambda^*\|_D + \frac{|\mu^*-\mu|}{1-\lambda}\|\Pi e\|_D,$$

and it follows that

$$\|\Phi\bar{r}_\lambda - \Phi r_\lambda^*\|_D \leq \frac{|\mu^*-\mu|}{(1-\beta_\lambda)(1-\lambda)}\|\Pi e\|_D.$$

The desired bound then follows from Theorem 3 and the triangle inequality.

q.e.d.

6 Conclusion

We have proposed a variant of temporal–difference learning that is suitable for approximating differential cost functions, and we have established the convergence of this algorithm when applied to finite state irreducible aperiodic Markov chains. In addition, we have provided bounds on the distance of the limiting function Φr_λ^* from the space of differential cost functions. These bounds involve the expression $\inf_r \|\mathcal{P}\Phi r - J^*\|_D$, which is natural because no approximation could have error smaller than this expression (when the error is measured in terms of $\|\mathcal{P}(\cdot)\|_D$). What is interesting is the factor of $1/\sqrt{1-\alpha_\lambda^2}$. The value of α_λ is in $[0,1)$ and generally decreases as λ increases, approaching zero as λ approaches one. Although this is only a bound, it strongly suggests that higher values of λ may lead to more accurate approximations. However, as has often been observed in the context of discounted cost temporal–difference learning, lower values of λ may lead to substantial gains in computational efficiency.

It is interesting to note that even if a given Markov chain takes a long time to reach steady state, the mixing factor α_λ may be small due to the choice of basis functions. In particular, the expected future value $E[\phi_k(i_t)|i_0]$ of a basis function may converge rapidly even though $E[J(i_t)|i_0]$ converges slowly for some other function J. This may partially explain why small values of λ seem to lead to good approximations even with Markov chains that converge to steady state rather slowly. The impressive Tetris player that was

constructed using methods similar to temporal–difference learning with small values of λ may exemplify this possibility (Bertsekas and Ioffe, 1996).

The main algorithm we analyzed adapts approximations of average cost and differential costs simultaneously. To better understand how error in the average cost estimate affects the approximation of the differential cost function, we analyzed a second algorithm, in which the average cost estimate μ is fixed, while the approximation of the differential cost function is adapted. This algorithm converges, but the error bound includes an extra term that is proportional to the error $|\mu - \mu^*|$ in the average cost estimate and a term $\|\Pi e\|_D$ that is influenced by the orientation of the basis functions. It is interesting that this term is equal to zero if the basis functions are all D–orthogonal to e. However, it is difficult in practice to ensure that such a property will be satisfied when basis functions are selected, especially since the steady–state probabilities (the elements of D) are unknown.

On the technical side, we mention a few straightforward extensions to our results.

1. With some additional technical assumptions, the proof of Theorem 1 can be extended to the case of infinite state Markov chains where approximations are generated using unbounded basis functions. This extension has been omitted for the sake of brevity, but largely involves arguments of the same type as in (Tsitsiklis and Van Roy, 1997).

2. 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.

3. Another extension is to allow the cost per stage $g(i_t)$ to be dependent on the next state (i.e., employ a function $g(i_t, i_{t+1})$) or even 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, our proof can easily be modified to accommodate this situation.

4. The assumption that the Markov chain was aperiodic can also be alleviated. No part of our convergence proof truly required this assumption – it was introduced merely to simplify the exposition.

5. Assumption 3(b) on the step size sequences was adopted for convenience, and weaker assumptions will certainly suffice, although this might require a substantially more sophisticated proof.

6. Finally, if Assumption 2(b) is removed, then our line of analysis can be used to show that $\mathcal{P}\Phi r_t$ still converges, but $(I - \mathcal{P}\Phi r_t)$ is aligned to e and need not converge.

Acknowledgments

We thank Peter Dayan for originally suggesting that ideas from (Tsitsiklis and Van Roy, 1997) might be applicable to the average cost setting.

References

Abounadi, J., Bertsekas, D.P., and Borkar, V.S. (1997) "ODE Analysis for Q–Learning Algorithms," Laboratory for Information and Decision Systems Draft Report, Massachusetts Institute of Technology, Cambridge, MA.

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

Bertsekas, D. P & Ioffe, S. (1996) "Temporal Differences–Based Policy Iteration and Applications in Neuro–Dynamic Programming," Laboratory for Information and Decision Systems Report LIDS–P–2349, Massachusetts Institute of Technology, Cambridge, MA.

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

Benveniste, A., Metivier, M., & Priouret, P., (1990) *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin.

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

Gallager, R. G. (1996) *Discrete Stochastic Processes*, Kluwer Academic Publishers, Boston, MA.

Gurvits, L., Lin, L. J., & Hanson, S. J. (1994) "Incremental Learning of Evaluation Functions for Absorbing Markov Chains: New Methods and Theorems," preprint.

Mahadevan, S. (1996) "Average Reward Reinforcement Learning: Foundations, Algorithms, and Empirical Results," *Machine Learning*, Vol. 22, pp. 1–38.

Pineda, F. (1996) "Mean–Field Analysis for Batched TD(λ)," unpublished.

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

Tsitsiklis, J. N. & Van Roy, B., (1997) "An Analysis of Temporal–Difference Learning with Function Approximation," the *IEEE Transactions on Automatic Control*, Vol. 42, No. 5, pp. 674-690.