Machine Learning, 49, 179–191, 2002

On Average Versus Discounted Reward Temporal-Difference Learning*

JOHN N. TSITSIKLIS jnt@mit.edu
Laboratory for Information and Decision Systems, M.I.T., Cambridge, MA 01239, USA

BENJAMIN VAN ROY bvr@stanford.edu
Department of Management Science and Engineering and Electrical Engineering, Stanford University, Stanford, CA 94305, USA

Editor: Satinder Singh

Abstract. We provide an analytical comparison between discounted and average reward temporal-difference (TD) learning with linearly parameterized approximations. We first consider the asymptotic behavior of the two algorithms. We show that as the discount factor approaches 1, the value function produced by discounted TD approaches the differential value function generated by average reward TD. We further argue that if the constant function—which is typically used as one of the basis functions in discounted TD—is appropriately scaled, the transient behaviors of the two algorithms are also similar. Our analysis suggests that the computational advantages of average reward TD that have been observed in some prior empirical work may have been caused by inappropriate basis function scaling rather than fundamental differences in problem formulations or algorithms.

Keywords: average reward, dynamic programming, function approximation, temporal-difference learning

1. Introduction

Sequential decision making problems under uncertainty are often formulated as infinite horizon dynamic programming problems and can be tackled by a variety of tools. For relatively small and well-modeled problems, classical methods apply (Bertsekas, 1995). When a problem is large or when a model is unavailable, one is led to consider simulation-based approximation methods, including temporal difference (TD) methods (Sutton, 1988).

The original version of TD, as well as most of the subsequent literature, has focused on problems with a discounted reward criterion. This was a natural first step for two reasons.

(a) Average reward problems can be accurately approximated (in a well-defined sense) by discounted problems, as the discount factor approaches 1.

*This research was partially supported by the National Science Foundation under grants DMI-9625489 and ACI-9873339, and by a Stanford University Terman Faculty Fellowship.

(b) The mathematical structure and analysis of average reward formulations can be complicated. This is because the favorable contraction properties of the dynamic programming operator are only present in the discounted case.

For these reasons, a discounted formulation with a discount factor very close to 1 is often used a proxy for an average reward problem. While there has been an emerging literature on the development and deployment of average reward reinforcement learning algorithms (see, e.g., Abounadi, 1998; Mahadevan, 1996; Marbach, Mihatsch, & Tsitsiklis, 1998; Schwartz, 1993; Singh, 1994; Tsitsiklis & Van Roy, 1999; Tadepalli & Ok, 1998; Van Roy, 1998), the picture is far from complete as only some of these methods can be adequately justified, mostly for the case of lookup table representations.

It is natural to inquire whether the practice of avoiding average reward formulations is necessary and/or sound. Our earlier work (Tsitsiklis & Van Roy, 1999; Van Roy, 1998) has answered the first question: it has provided a variant of TD that can be applied to average reward problems. Furthermore, it has established that average reward TD has the same theoretical guarantees as were established in Tsitsiklis and Van Roy (1997) and Van Roy (1998) for discounted TD. Assuming a linearly parameterized approximation architecture, we have convergence with probability 1, a characterization of the limit, and bounds on approximation error. Results of this type, for both discounted and average reward versions, refer to Markov chains that are operated under a fixed policy. When the policy is allowed to change during the learning process, one usually hopes that such methods will deliver good performance. This has been the case for discounted TD in many applications. The experience with average TD is limited, but has also been positive (Marbach, Mihatsch, & Tsitsiklis, 1998).

This paper focuses on the second issue, namely, soundness. Is it safe to replace an average reward problem by a discounted one? Does discounted TD become more and more ill-conditioned as the discount factor approaches 1? Could there be an inherent advantage to average reward TD? These questions have been prompted to a large extent by available experience, which suggests that average reward methods may offer computational advantages (Marbach, Mihatsch, & Tsitsiklis, 1998; Tadepalli & Ok, 1998).

The analysis provided in the remainder of the paper indicates that discounted TD (with a discount factor approaching 1) is essentially identical to average reward TD. The two methods converge to the same limit. Furthermore, *as long as* the constant function—which is typically used as one of the basis functions in discounted TD—is appropriately scaled, their transient behaviors are also similar. This analysis suggests that the observed computational advantages of average reward TD may have been caused by inappropriate basis function scaling rather than fundamental differences in problem formulations or algorithms.

We will limit the discussion in this paper to the special case of TD that is known as TD(0). This will allow us to keep the exposition simple. However, similar arguments apply to TD(λ) for general λ, and lead to the same conclusions.

The remainder of this paper is organized as follows. The next section reviews discounted and average reward temporal-difference learning and discusses how to relate the basis functions used by the two algorithms in any meaningful comparison. Section 3 establishes the similarity of the limits of convergence of the two algorithms. The transient behaviors are compared in Section 4. Concluding remarks are made in Section 5.

2. Temporal-difference learning algorithms

In this section, we review the discounted and average reward TD(0) algorithms and set the stage for their comparison. Our discussion will be somewhat brief. We refer the reader to Tsitsiklis and Van Roy (1997, 1999) for more details.

We consider a Markov chain $\{x_t \mid t = 0, 1, \ldots\}$ with state space $S = \{1, \ldots, n\}$. The Markov chain is defined in terms of a transition probability matrix P whose (i, j)th entry is the probability that $x_{t+1} = j$ given that $x_t = i$. We assume that the chain is irreducible and aperiodic, and therefore has a unique invariant distribution vector $\pi \in \Re^n$, which satisfies $\pi' P = \pi'$ and $\pi(x) > 0$ for all $x \in S$. (In our notation, π' stands for the transpose of π, and $\pi(x)$ is the xth component. A similar notational convention will be used for all other vectors to be encountered later on). For each state $x \in S$, there is a scalar $g(x)$ that represents the one-stage reward for being at that state. We let $g \in \Re^n$ be a vector whose xth component is $g(x)$.

We will make use of an inner product $\langle \cdot, \cdot \rangle_\pi$ defined by

$$\langle J, \bar{J} \rangle_\pi = \sum_{x \in S} J(x)\bar{J}(x)\pi(x), \quad \forall\, J, \bar{J} \in \Re^n.$$

The norm $\| \cdot \|_\pi$ corresponding to the above defined inner product is given by

$$\|J\|_\pi^2 = \langle J, J \rangle_\pi.$$

Finally, we will use the notation

$$e = (1, 1, \ldots, 1)'.$$

Note that $\|e\|_\pi = 1$ because $\sum_x e^2(x)\pi(x) = \sum_x \pi(x) = 1$.

2.1. *Discounted TD(0)*

Given a discount factor $\alpha \in (0, 1)$, the *discounted value function* $J^{(\alpha)}(x)$ is defined as the expectation of the discounted sum of accumulated rewards, starting from state x:

$$J^{(\alpha)}(x) = E\left[\sum_{t=0}^{\infty} \alpha^t g(x_t) \,\middle|\, x_0 = x\right].$$

In vector notation, we have

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

We consider approximations to the discounted value function using an approximation architecture of the form

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

Here, $r = (r(1), \ldots, r(K))'$ is a parameter vector and each ϕ_k is a prespecified scalar function 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 an associated weight. We will require that the basis functions be linearly independent, so that each vector $r \in \Re^K$ yields a different function $\tilde{J}(\cdot, r)$. The aim of TD methods is to settle on a value for the parameter vector r which makes the function $\tilde{J}(\cdot, r)$ a good approximation of $J^{(\alpha)}$.

Given a selection of basis functions and an arbitrary initial parameter vector $r_0 \in \Re^K$, discounted reward TD(0) observes a sample path $x_0, x_1, x_2, \ldots$ of the underlying Markov process, and generates a sequence $r_1, r_2, \ldots$, by letting

$$r_{t+1} = r_t + \gamma_t d_t \phi(x_t). \tag{1}$$

Here, γ_t is the stepsize, d_t is the temporal difference

$$d_t = g(x_t) + \alpha \tilde{J}(x_{t+1}, r_t) - \tilde{J}(x_t, r_t), \tag{2}$$

corresponding to the transition from x_t to x_{t+1}, and $\phi(x_t) = (\phi_1(x_t), \ldots, \phi_K(x_t))'$.

2.2. *Average reward TD(0)*

The *average reward* is defined by $\mu^* = \pi' g$, which is the expectation of the reward per stage $g(x_t)$, when the Markov chain is in steady state. We define the basic differential reward function $J^* : S \mapsto \Re$ by

$$J^*(x) = \lim_{T\to\infty} E\left[\sum_{t=0}^{\infty}(g(x_t) - \mu^*) \,\middle|\, x_0 = x\right].$$

In vector notation, we have

$$J^* = \sum_{t=0}^{\infty} P^t(g - \mu^* e). \tag{3}$$

Intuitively, for any two states x and y, the difference $J^*(x) - J^*(y)$ measures the relative advantage (transient gain) of starting in state x rather than state y.

Once again, we consider approximating J^* using functions of the form

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

Similar to the discounted case, we require that the basis functions $\phi_1, \ldots, \phi_K$ be linearly independent. In addition, we assume here that the vector e is not in the span of the basis functions.

In addition to a sequence $r_1, r_2, \ldots$, average reward TD(0) will generate a sequence $\mu_1, \mu_2, \ldots$ of approximations to the average reward μ^*. We define the temporal difference d_t corresponding to the transition from x_t to x_{t+1} by

$$d_t = g(x_t) - \mu_t + \tilde{J}(x_{t+1}, r_t) - \tilde{J}(x_t, r_t). \tag{4}$$

The parameter vector is then updated according to:

$$r_{t+1} = r_t + \gamma_t d_t \phi(x_t), \tag{5}$$

where the components of r_0 are initialized to arbitrary values and γ_t is a sequence of scalar step sizes. The average reward estimate is updated according to

$$\mu_{t+1} = (1 - \gamma_t)\mu_t + \gamma_t g(x_t),$$

where μ_0 is an initial estimate.

2.3. Relation between basis functions

It is interesting to note that the average reward μ^* can be computed from the discounted value function $J^{(\alpha)}$. In particular, since $\pi' P = \pi$, we have

$$\pi' J^{(\alpha)} = \pi' \sum_{t=0}^{\infty} \alpha^t P^t g = \sum_{t=0}^{\infty} \alpha^t \pi' g = \frac{\mu^*}{1 - \alpha}.$$

Note that for any J, we have $\pi' J = \langle e, J \rangle_\pi$. Therefore, we can also write

$$\mu^* = (1 - \alpha)\left\langle e, J^{(\alpha)} \right\rangle_\pi.$$

Given the last equality, we can interpret the part of $J^{(\alpha)}$ that is aligned with e, i.e., $e\langle e, J^{(\alpha)}\rangle_\pi$ as an encoding of the average reward μ^*. It turns out that the rest of $J^{(\alpha)}$, i.e., the difference $J^{(\alpha)} - e\langle e, J^{(\alpha)}\rangle_\pi$, can be interpreted as an encoding of the differential value function. We now formalize this point.

We define an $n \times n$ matrix Γ by

$$\Gamma = I - e\pi'.$$

It is easily checked that Γ is a projection (with respect to the norm $\|\cdot\|_\pi$) on the set of vectors that are orthogonal to e. It is known that as α approaches 1 the projection $\Gamma J^{(\alpha)}$ converges to the differential value function J^*.

To verify this property, we start with the equalities $Pe = e$ and $\pi' P = \pi'$, which lead to $e\pi' P = Pe\pi'$. Since $\Gamma = I - e\pi'$, we obtain that

$$\Gamma P = P\Gamma.$$

We also note that $\Gamma g = g - e\pi' g = g - \mu^* e$. Therefore,

$$\Gamma J^{(\alpha)} = \Gamma \left(\sum_{t=0}^{\infty} \alpha^t P^t g \right) = \sum_{t=0}^{\infty} \alpha^t P^t (g - \mu^* e). \tag{6}$$

It is well known that for irreducible and aperiodic Markov chains, P^t converges to its limit $e\pi'$ (the matrix with all rows equal to π') at the rate of a geometric progression. It follows that $P^t(g - \mu^* e)$ also approaches its limit, which is zero, geometrically fast. Thus, comparing Eqs. (3) and (6), we see that, for some constants $B > 0$ and $\beta < 1$,

$$\left\| \Gamma J^{(\alpha)} - J^* \right\|_\pi \leq \sum_{t=0}^{\infty} (1 - \alpha^t) \| P^t (g - \mu^* e) \|_\pi \leq \sum_{t=0}^{\infty} (1 - \alpha^t) B \beta^t,$$

which approaches 0 as α approaches 1.

To summarize the preceding discussion, we have seen that the discounted value function $J^{(\alpha)}$ provides the average reward and an approximation to the differential value function, in the following sense:

$$J^* \approx \Gamma J^{(\alpha)} \quad \text{and} \quad \mu^* = (1 - \alpha) \left\langle e, J^{(\alpha)} \right\rangle_\pi .$$

The limits of convergence of discounted and average reward TD(0), exhibit similar relationships, which we will explore in subsequent sections.

Recall our assumption that in average reward TD(0), the constant vector e is not within the span of the basis functions. On the other hand, the estimate μ in average reward TD(0) plays the same role as the one that would be played in discounted TD(0) by a basis function aligned with e. Hence, when we compare discounted and average reward TD(0), we should include an extra basis function in the former. In particular, we will study and compare the following two situations:

1. Average reward TD(0), with linearly independent basis functions $\phi_1, \ldots, \phi_K$ that do not include e in their span.
2. Discounted TD(0) with basis functions $\phi_1, \ldots, \phi_K, Ce$ (a total of $K + 1$ basis functions). Here, C is a fixed real number, which is used to scale the constant basis function e.

The next section compares the ultimate approximations generated by these two approaches, while Section 4 discusses the incremental updates that occur during execution of the algorithms. It turns out that the algorithms become very similar when α is close to 1.

3. Asymptotic behavior

If X is a subspace of $\Re^n$ and J is some vector, its (orthogonal) projection (with respect to the metric $\|\cdot\|_\pi$) is defined to be a vector $\bar{J} \in X$ for which the distance $\|J - \bar{J}\|_\pi$ is minimized. The mapping from J to $\bar{J}$ is linear and can be represented by an $n \times n$

matrix. In this spirit, let Π be the matrix that projects on the subspace spanned by the basis functions $\{\phi_1, \ldots, \phi_K, e\}$ of discounted TD(0). Similarly, let Π_0 be the matrix that projects on the subspace spanned by the basis functions $\{\phi_1, \ldots, \phi_K\}$ of average reward TD(0).

Under the assumptions introduced so far, together with some standard technical conditions on the step size sequence γ_t, the sequence of value functions $\tilde{J}(\cdot, r_t)$ generated by TD(0) methods converge, with probability 1, to limiting value functions, which we denote by $\tilde{J}^{(\alpha)}$ and $\tilde{J}$ for the discounted and average reward cases, respectively. Furthermore, these limits have been characterized as the unique solutions to the following linear equations:

$$\tilde{J}^{(\alpha)} = \Pi\big(\alpha P\tilde{J}^{(\alpha)} + g\big), \quad \tilde{J} = \Pi_0(P\tilde{J} + g). \tag{7}$$

Finally, μ_t in average reward TD(0) converges to μ^* (Tsitsiklis & Van Roy, 1997, 1999).

Our main result below indicates that for α close to 1, discounted TD is essentially identical to average reward TD, in the following sense. There is a part of $\tilde{J}^{(\alpha)}$ which is aligned with e, namely, $(1-\alpha)e\langle e, \tilde{J}^{(\alpha)}\rangle_\pi$, which provides an encoding of the average reward μ^*. In addition, the part of $\tilde{J}^{(\alpha)}$ which is orthogonal to the constant function, namely $\Gamma\tilde{J}^{(\alpha)}$, approximates the corresponding part $\Gamma\tilde{J}$. Note that $\Gamma\tilde{J}$ only differs from $\tilde{J}$ by a multiple of e. Such constant differences are irrelevant when value functions are used in decision making. Thus, $\tilde{J}^{(\alpha)}$ (through $\Gamma\tilde{J}^{(\alpha)}$) encompasses any relevant information that is contained in $\tilde{J}$.

Proposition 1. *We assume that the underlying Markov chain is irreducible and aperiodic. Then the solutions of the equations shown in* (7) *satisfy*:

$$\lim_{\alpha\uparrow 1} \Gamma\tilde{J}^{(\alpha)} = \Gamma\tilde{J} \quad \text{and} \quad \mu^* = (1-\alpha)\big\langle e, \tilde{J}^{(\alpha)}\big\rangle_\pi.$$

Proof: Recall that Π projects on the span X of $\{\phi_1, \ldots, \phi_K, e\}$, whereas $I-\Gamma$ projects on the span of $\{e\}$, which is a subspace of X. In this context, it is easily seen that one can project in any order, and we have $(I-\Gamma)\Pi = \Pi(I-\Gamma)$, which also implies that $\Gamma\Pi = \Pi\Gamma$. We also recall from Section 2 that $\Gamma P = P\Gamma$.

We start with the equation on the left-hand side of (7), right-multiply both sides by Γ and use the commutation properties that we have just derived. We obtain

$$\Gamma\tilde{J}^{(\alpha)} = \Pi\big(\alpha P\Gamma\tilde{J}^{(\alpha)} + \Gamma g\big).$$

As shown in Tsitsiklis and Van Roy (1997), the matrix $\alpha\Pi P$ is a contraction with respect to the norm $\|\cdot\|_\pi$, for any $\alpha \in [0, 1)$. Therefore, $\Gamma\tilde{J}^{(\alpha)}$ is the unique solution of the equations

$$U = \Pi(\alpha PU + \Gamma g), \quad \Gamma U = U, \tag{8}$$

for $\alpha \in [0, 1)$.

Let us now consider the equation on the right-hand side of 7. For any vector J, there exists a scalar c such that

$$\Pi J = \Pi_0 J + ce.$$

This is simply because Π_0 projects onto the subspace of X that is orthogonal to e. We apply this property to the equation on the right-hand side of (7), which yields

$$\tilde{J} = \Pi(P\tilde{J} + g) + ce.$$

We then multiply by Γ, use commutativity and the property $\Gamma e = 0$, to obtain $\Gamma\tilde{J} = \Pi(P\Gamma\tilde{J} + \Gamma g)$. In particular, $\Gamma\tilde{J}$ is a solution to the equations

$$V = \Pi(PV + \Gamma g), \quad \Gamma V = V. \tag{9}$$

We now argue that these equations have a unique solution. Consider any two solutions, and let J be their difference. We then have $J = \Pi PJ$ and $\Gamma J = J$. Since $\|P\|_\pi \leq 1$ and $\|\Pi\|_\pi \leq 1$ (the latter being a generic property of projections), we must have $\|\Pi PJ\|_\pi = \|PJ\|_\pi = \|J\|_\pi$. The fact that $\|\Pi PJ\|_\pi = \|PJ\|_\pi$ implies that $\Pi PJ = PJ$. We conclude that $PJ = J$, and therefore J is a multiple of e. Together with the property $J = \Gamma J$, we obtain $J = 0$. Since the difference of any two solutions is zero, there must be a unique solution.

The proof can now be completed. The system (9) has a unique solution. As $\alpha \uparrow 1$, the coefficient matrices in the system (8) converge to the coefficient matrices for the system (9). It is then immediate that the unique solution $\Gamma\tilde{J}^{(\alpha)}$ of (8) converges to the unique solution $\Gamma\tilde{J}$ of (9).

To complete the argument, we have

$$\langle e, \tilde{J}^{(\alpha)}\rangle_\pi = \pi' g + \pi'\alpha P\tilde{J}^{(\alpha)} = \mu^* + \alpha\pi'\tilde{J}^{(\alpha)} = \mu^* + \alpha\langle e, \tilde{J}^{(\alpha)}\rangle_\pi,$$

and it follows that

$$\langle e, \tilde{J}^{(\alpha)}\rangle_\pi = \frac{\mu^*}{1-\alpha}. \qquad \square$$

4. Transient behavior

The result in the previous section holds no matter how we choose the scaling factor C for the basis function Ce in discounted TD(0). In particular, the point to which the algorithm converges does not depend on C. However, a poorly chosen scaling factor can impact significantly the transient behavior. We will now show that the transient behaviors of discounted and average TD(0) are essentially the same. However, this argument has to assume a very particular choice for the scaling factor, namely,

$$C = \frac{1}{1-\alpha} - \sum_{k=1}^{K}\langle e, \phi_k\rangle_\pi^2.$$

This formula is not practical, because the sum on the right is an unknown quantity (although it can be easily estimated with a Monte Carlo simulation run). What is important, however, is that the scaling factor increases with $1/(1-\alpha)$ as α approaches 1. In the absence of that, we would expect the transient behavior of discounted TD(0) to be very different from its average reward counterpart, and to deteriorate as α increases to 1.

Let r_t and μ_t be the parameters generated by average reward TD(0), and let $r_t^{(\alpha)}$ be the parameters in the discounted case. In the previous section, we focused on the correspondence between $\Gamma \tilde{J}_t$ and $\Gamma \tilde{J}_t^{(\alpha)}$, as well as on the correspondence between μ_t and $(1-\alpha)\langle e, \tilde{J}_t^{(\alpha)}\rangle_\pi$. We showed that the correspondence becomes exact in the limit. We will now consider a situation where the two algorithms start at some time t with an exact such correspondence, and investigate whether the correspondence will tend to be preserved.

We introduce the notation

$$\bar{\phi}_k = \Gamma \phi_k, \quad k = 1, \ldots, K.$$

The average reward TD(0) update is given by

$$\begin{aligned} r_{t+1}(k) &= r_t(k) + \gamma_t \phi(x_t) \left(g(x_t) - \mu_t + \sum_{j=1}^{K} \phi_j(x_{t+1}) r_t(j) - \sum_{j=1}^{K} \phi_j(x_t) r_t(j) \right) \\ &= r_t(k) + \gamma_t \phi(x_t) \left(g(x_t) - \mu_t + \sum_{j=1}^{K} \bar{\phi}_j(x_{t+1}) r_t(j) - \sum_{j=1}^{K} \bar{\phi}_j(x_t) r_t(j) \right), \end{aligned}$$

because the component of each ϕ_k that is aligned with e evaluates to the same value for both x_t and x_{t+1}.

Suppose that $\mu_t = (1-\alpha)\langle e, \tilde{J}_t^{(\alpha)}\rangle_\pi$. Note that

$$\tilde{J}_t^{(\alpha)} = \Gamma \tilde{J}_t^{(\alpha)} + (I - \Gamma)\tilde{J}_t^{(\alpha)} = \sum_{j=1}^{K} \bar{\phi}_j r_t^{(\alpha)}(j) + \left\langle e, \tilde{J}_t^{(\alpha)} \right\rangle_\pi.$$

The update in the discounted case is then given by

$$\begin{aligned} & r_{t+1}^{(\alpha)}(k) \\ &\quad = r_t^{(\alpha)}(k) + \gamma_t \phi_k(x_t) \left(g(x_t) + \alpha \tilde{J}_t^{(\alpha)}(x_{t+1}) - \tilde{J}_t^{(\alpha)}(x_t) \right) \\ &\quad = r_t^{(\alpha)}(k) + \gamma_t \phi_k(x_t) \left(g(x_t) - \mu_t + \alpha \sum_{j=1}^{K} \bar{\phi}_j(x_{t+1}) r_t^{(\alpha)}(j) - \sum_{j=1}^{K} \bar{\phi}_j(x_t) r_t^{(\alpha)}(j) \right). \end{aligned}$$

We observe that as α approaches 1, the update equation becomes identical to the one used in the average reward case.

We now study the equations relating to the average reward estimates. In average reward TD(0), we have

$$\mu_{t+1} = \mu_t + \gamma_t (g(x_t) - \mu_t).$$

In the discounted case, the change in the average reward estimate $\langle e, \tilde{J}_t^{(\alpha)}\rangle_\pi$ takes a significantly different form. Always assuming that $\mu_t = (1-\alpha)\langle e, \tilde{J}_t^{(\alpha)}\rangle_\pi$, we have

$$
\begin{aligned}
&(1-\alpha)\left\langle e, \tilde{J}_{t+1}^{(\alpha)}\right\rangle_\pi \\
&\quad = (1-\alpha)\left\langle e, \sum_{k=1}^{K} r_{t+1}^{(\alpha)}(k)\phi_k + r_{t+1}^{(\alpha)}(K+1)Ce\right\rangle_\pi \\
&\quad = (1-\alpha)\sum_{k=1}^{K} r_{t+1}^{(\alpha)}(k)\langle e, \phi_k\rangle_\pi + (1-\alpha)r_{t+1}^{(\alpha)}(K+1)C \\
&\quad = (1-\alpha)\sum_{k=1}^{K}\left(r_t^{(\alpha)}(k) + \gamma_t\phi_k(x_t)\left(g(x_t) - \mu_t + \alpha\sum_{j=1}^{K}\bar{\phi}_j(x_{t+1})r_t^{(\alpha)}(j)\right.\right. \\
&\qquad \left.\left. - \sum_{j=1}^{K}\bar{\phi}_j(x_t)r_t^{(\alpha)}(j)\right)\right)\langle e, \phi_k\rangle_\pi + (1-\alpha)\left(Cr_t^{(\alpha)}(K+1) + \gamma_t C\right. \\
&\qquad \left.\times\left(g(x_t) - \mu_t + \alpha\sum_{j=1}^{K}\bar{\phi}_j(x_{t+1})r_t^{(\alpha)}(j) - \sum_{j=1}^{K}\bar{\phi}_j(x_t)r_t^{(\alpha)}(j)\right)\right) \\
&\quad = (1-\alpha)\left\langle e, \sum_{k=1}^{K} r_t^{(\alpha)}(k)\phi_k\right\rangle_\pi + (1-\alpha)Cr_t^{(\alpha)}(K+1) \\
&\qquad + (1-\alpha)\gamma_t\left(\sum_{k=1}^{K}\langle e, \phi_k\rangle_\pi\phi_k(x_t) + C\right) \\
&\qquad \times\left(g(x_t) - \mu_t + \alpha\sum_{j=1}^{K}\bar{\phi}_j(x_{t+1})r_t^{(\alpha)}(j) - \sum_{j=1}^{K}\bar{\phi}_j(x_t)r_t^{(\alpha)}(j)\right) \\
&\quad = \mu_t + (1-\alpha)\gamma_t\left(\sum_{k=1}^{K}\langle e, \phi_k\rangle_\pi\phi_k(x_t) + C\right) \\
&\qquad \times\left(g(x_t) - \mu_t + \alpha\sum_{j=1}^{K}\bar{\phi}_j(x_{t+1})r_t^{(\alpha)}(j) - \sum_{j=1}^{K}\bar{\phi}_j(x_t)r_t^{(\alpha)}(j)\right).
\end{aligned}
$$

Denoting the difference $(1-\alpha)\langle e, \tilde{J}_{t+1}^{(\alpha)}\rangle_\pi - (1-\alpha)\langle e, \tilde{J}_t^{(\alpha)}\rangle_\pi$ by Δ_t, we have

$$
\begin{aligned}
\Delta_t \;=\; & (1-\alpha)\gamma_t\left(\sum_{k=1}^{K}\langle e, \phi_k\rangle_\pi^2 + C\right)\left(g(x_t) - \mu_t + \alpha\sum_{j=1}^{K}\bar{\phi}_j(x_{t+1})r_t^{(\alpha)}(j)\right. \\
& \left. - \sum_{j=1}^{K}\bar{\phi}_j(x_t)r_t^{(\alpha)}(j)\right) + (1-\alpha)\gamma_t\sum_{k=1}^{K}\langle e, \phi_k\rangle_\pi\bar{\phi}_k(x_t) \\
& \times\left(g(x_t) - \mu_t + \alpha\sum_{j=1}^{K}\bar{\phi}_j(x_{t+1})r_t^{(\alpha)}(j) - \sum_{j=1}^{K}\bar{\phi}_j(x_t)r_t^{(\alpha)}(j)\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \gamma_t(g(x_t) - \mu_t) + \gamma_t \left(\alpha \sum_{j=1}^{K} \bar{\phi}_j(x_{t+1}) r_t^{(\alpha)}(j) - \sum_{j=1}^{K} \bar{\phi}_j(x_t) r_t^{(\alpha)}(j) \right) \\
&\quad + (1-\alpha)\gamma_t \sum_{k=1}^{K} \langle e, \phi_k \rangle_\pi \bar{\phi}_k(x_t) \\
&\quad \times \left(g(x_t) - \mu_t + \alpha \sum_{j=1}^{K} \bar{\phi}_j(x_{t+1}) r_t^{(\alpha)}(j) - \sum_{j=1}^{K} \bar{\phi}_j(x_t) r_t^{(\alpha)}(j) \right),
\end{aligned}
$$

where we have used the fact that

$$
C = \frac{1}{1-\alpha} - \sum_{k=1}^{K} \langle e, \phi_k \rangle_\pi^2.
$$

Let us discuss the three terms involved in Δ_t. The first term

$$
\gamma_t(g(x_t) - \mu_t),
$$

is identical to the one in the update equation for μ_t in average reward TD(0).

The third term

$$
\begin{aligned}
&(1-\alpha)\gamma_t \sum_{k=1}^{K} \langle e, \phi_k \rangle_\pi \bar{\phi}_k(x_t) \\
&\quad \times \left(g(x_t) - \mu_t + \alpha \sum_{j=1}^{K} \bar{\phi}_j(x_{t+1}) r_t^{(\alpha)}(j) - \sum_{j=1}^{K} \bar{\phi}_j(x_t) r_t^{(\alpha)}(j) \right),
\end{aligned}
$$

is absent in average reward TD(0). However, it is easy to see that as α approaches one, this term vanishes.

The second term

$$
\gamma_t \left(\alpha \sum_{j=1}^{K} \bar{\phi}_j(x_{t+1}) r_t^{(\alpha)}(j) - \sum_{j=1}^{K} \bar{\phi}_j(x_t) r_t^{(\alpha)}(j) \right),
$$

does not vanish. However, we will now argue that when we sum this term over time we are left with a quantity which is of second order in the step size magnitude, and is therefore negligible when compared to the first order term $\gamma_t(g(x_t) - \mu_t)$.

As shorthand, we let

$$
s_{t,\tau} = \sum_{j=1}^{K} \bar{\phi}_j(x_t) r_\tau^{(\alpha)}(j).
$$

Hence, the second term making up Δ_t can be rewritten as $\gamma_t(\alpha s_{t+1,t} - s_{t,t})$, and as α approaches 1, this difference becomes $\gamma_t(s_{t+1,t} - s_{t,t})$. Consider the accumulation of the

term of interest, which is

$$\sum_{\tau=0}^{t} \gamma_\tau (s_{\tau+1,\tau} - s_{\tau,\tau}) = \gamma_t s_{t+1,t} - \gamma_0 s_{0,0} + \sum_{\tau=0}^{t-1} \gamma_\tau (s_{\tau+1,\tau} - s_{\tau+1,\tau+1}).$$

Roughly speaking, we need only be concerned with the summation because the other terms do not accumulate over time. The difference between $s_{\tau+1,\tau}$ and $s_{\tau+1,\tau+1}$ is of order γ_τ since the only source of difference is the replacement of r_τ with $r_{\tau+1}$. Therefore, the sum

$$\sum_{\tau=0}^{t-1} \gamma_\tau (s_{\tau+1,t} - s_{\tau+1,\tau+1}),$$

involves terms of order γ_τ^2, and the accumulation of such terms becomes insignificant relative to the sum of terms of order γ_τ involved in updating the average cost estimate.

It is well known in stochastic approximation theory that any effects that are of second order in the stepsize can be safely neglected, especially under the standard assumption $\sum_t \gamma_t^2 < \infty$. In a more rigorous analysis, we could develop an asymptotic result (in the limit of small stepsizes) of the central limit type, that provides information on the variance of the transient fluctuations. Second order, $O(\gamma_t^2)$ terms, usually disappear in such an analysis and one would then obtain a formal characterization of the similarities of the two transient behaviors. However, we feel that the argument given here captures the essence of the situation, while also being accessible.

5. Conclusion

The general conclusion suggested by our analysis seems to be that the choice between discounted and average reward TD(0) is of little consequence, as long as one is careful with the scaling of the constant function. Esthetic considerations (why introduce an extraneous discount factor when none is called for), together with the possibility of erroneous judgments in the choice of scaling could argue in favor of average reward TD(0).

Our theoretical analysis still needs to be corroborated by numerical experimentation. Finally, we have only dealt with the case of autonomous Markov chains (or a Markov decision process controlled under a fixed stationary policy). One may conjecture that the similarities extend to the controlled case.

References

Abounadi, J. (1998). Stochastic approximation for non-expansive maps: Application to Q-learning algorithms. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA.

Bertsekas, D. P. (1995). *Dynamic programming and optimal control*, Belmont, MA: Athena Scientific.

Mahadevan, S. (1996). Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning, 22*, 1–38.

Marbach, P., Mihatsch, O., & Tsitsiklis, J. N. (1998). Call admission control and routing in integrated service networks using reinforcement learning. In *Proceedings of the 1998 IEEE CDC*, Tampa, FL.

Schwartz, A. (1993). A reinforcement learning algorithm for maximizing undiscounted rewards. In *Proceedings of the Tenth Machine Learning Conference*.

Singh, S. P. (1994). Reinforcement learning algorithms for average payoff markovian decision processes. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*.

Sutton, R. S. (1988). Learning to predict by the method of temporal differences. *Machine Learning, 3*, 9–44.

Tadepalli, P., & Ok, D. (1998). Model-based average reward reinforcement learning. *Artificial Intelligence, 100*, 177–224.

Tsitsiklis, J. N., & Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control, 42:5*, 674–690.

Tsitsiklis, J. N., & Van Roy, B. (1999). Average cost temporal-difference learning. *Automatica, 35:11*, 1799–1808.

Van Roy, B. (1998). Learning and value function approximation in complex decision processes. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

Received March 3, 1999
Revised March 16, 2001
Accepted March 19, 2001
Final manuscript March 19, 2001