Machine Learning, 8, 341-362 (1992)

The Convergence of TD(λ) for General λ

PETER DAYAN dayan@helmholtz.sdsc.edu
Centre for Cognitive Science & Department of Physics, University of Edinburgh, EH8 9LW, Scotland

Abstract. The method of temporal differences (TD) is one way of making consistent predictions about the future. This paper uses some analysis of Watkins (1989) to extend a convergence theorem due to Sutton (1988) from the case which only uses information from adjacent time steps to that involving information from arbitrary ones.

It also considers how this version of TD behaves in the face of linearly dependent representations for states—demonstrating that it still converges, but to a different answer from the least mean squares algorithm. Finally it adapts Watkins' theorem that $\mathcal{Q}$-learning, his closely related prediction and action learning method, converges with probability one, to demonstrate this strong form of convergence for a slightly modified version of TD.

Keywords. Reinforcement learning, temporal differences, asynchronous dynamic programming

1. Introduction

Many systems operate in temporally extended circumstances, for which whole sequences of states rather than just individual ones are important. Such systems may frequently have to predict some future outcome, based on a potentially stochastic relationship between it and their current states. Furthermore, it is often important for them to be able to learn these predictions based on experience.

Consider a simple version of this problem in which the task is to predict the expected ultimate terminal values starting from each state of an absorbing Markov process, and for which some further random processes generate these terminating values at the absorbing states. One way to make these predictions is to learn the transition matrix of the chain and the expected values from each absorbing state, and then solve a simultaneous equation in one fell swoop. A simpler alternative is to learn the predictions directly, without first learning the transitions.

The methods of temporal differences (TD), first defined as such by Sutton (1984; 1988), fall into this simpler category. Given some parametric way of predicting the expected values of states, they alter the parameters to reduce the inconsistency between the estimate from one state and the estimates from the next state or states. This learning can happen incrementally, as the system observes its states and terminal values. Sutton (1988) proved some results about the convergence of a particular case of a TD method.

Many control problems can be formalized in terms of controlled, absorbing, Markov processes, for which each *policy*, i.e., mapping from states to actions, defines an absorbing Markov chain. The engineering method of dynamic programming (DP) (Bellman & Dreyfus, 1962) uses the predictions of the expected terminal values as a way of judging and hence improving policies, and TD methods can also be extended to accomplish this. As discussed extensively by Watkins (1989) and Barto, Sutton and Watkins (1990), TD is actually very closely related to DP in ways that significantly illuminate its workings.

This paper uses Watkins' insights to extend Sutton's theorem from a special case of TD, which considers inconsistencies only between adjacent states, to the general case in which arbitrary states are important, weighted exponentially less according to their temporal distances. It also considers what TD converges to if the representation adopted for states is linearly dependent, and proves that one version of TD prediction converges with probability one, by casting it in the form of Q-learning.

Some of the earliest work in temporal difference methods was due to Samuel (1959; 1967). His checkers (draughts) playing program tried to learn a consistent function for evaluating board positions, using the discrepancies between the predicted values at each state based on limited depth games-tree searches, and the subsequently predicted values after those numbers of moves had elapsed. Many other proposals along similar lines have been made: Sutton acknowledged the influence of Klopf (1972; 1982) and in Sutton (1988) discussed Holland's bucket brigade method for classifier systems (Holland, 1986), and a procedure by Witten (1977). Hampson (1983; 1990) presented empirical results for a quite similar navigation task to the one described by Barto, Sutton and Watkins (1990). Barto, Sutton and Anderson (1983) described an early TD system which learns how to balance an upended pole, a problem introduced in a further related paper by Michie and Chambers (1968). Watkins (1989) also gave further references.

The next section defines TD(λ), shows how to use Watkins' analysis of its relationship with DP to extend Sutton's theorem, and makes some comments about unhelpful state representations. Section 3 looks at Q-learning, and uses a version of Watkins' convergence theorem to demonstrate in a particular case the strongest guarantee known for the behavior of TD(0).

2. TD(λ)

Sutton (1988) developed the rationale behind TD methods for prediction, and proved that TD(0), a special case with a time horizon of only one step, converges in the mean for observations of an absorbing Markov chain. Although his theorem applies generally, he illustrated the case in point with an example of the simple random walk shown in figure 1. Here, the chain always starts in state D, and moves left or right with equal probabilities from each state until it reaches the left absorbing barrier A or the right absorbing barrier G. The problem facing TD is estimating the probability it absorbs at the right hand barrier rather than the left hand one, given any of the states as a current location.

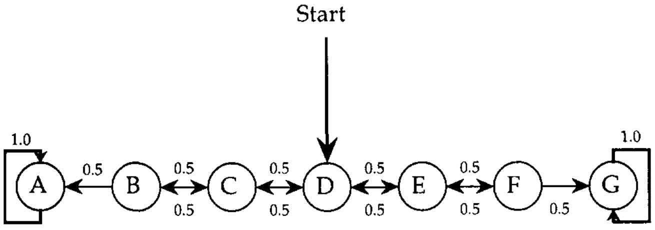

Figure 1. Sutton's Markov example. Transition probabilities given above (right to left) and below (left to right) the arrows.

The raw information available to the system is a collection of sequences of states and terminal locations generated from the random walk—it initially has no knowledge of the transition probabilities. Sutton described the supervised least mean squares (LMS) (Widrow & Stearns, 1985) technique, which works by making the estimates of the probabilities for each place visited on a sequence closer to 1 if that sequence ended up at the right hand barrier, and closer to 0 if it ended up at the left hand one. He showed that this technique is exactly TD(1), one special case of TD, and constrasted it with TD(λ) and particularly TD(0), which tries to make the estimate of probability from one state closer to the estimate from the next, without waiting to see where the sequence might terminate. The discounting parameter λ in TD(λ) determines exponentially the weights of future states based on their temporal distance—smoothly interpolating between $\lambda = 0$, for which only the next state is relevant, and $\lambda = 1$, the LMS case, for which all states are equally weighted. As described in the introduction, it is its obeisance to the temporal order of the sequence that marks out TD.

The following subsections describe Sutton's results for TD(0) and separate out the algorithm from the vector representation of states. They then show how Watkins' analysis provides the wherewithal to extend it to TD(λ) and finally re-incorporate the original representation.

2.1. The convergence theorem

Following Sutton (1988), consider the case of an absorbing Markov chain, defined by sets and values:

$\mathcal{T}$		terminal states
$\mathcal{N}$		non-terminal states
$q_{ij} \in [0, 1]$	$i \in \mathcal{N}, j \in \mathcal{N} \cup \mathcal{T}$	transition probabilities
$\mathbf{x}_i \in \mathcal{R}^c$	$i \in \mathcal{N}$	vectors representing non-terminal states
$\bar{z}_j$	$j \in \mathcal{T}$	expected terminal values from state j
μ_i	$i \in \mathcal{N}$	probabilities of starting at state i, where $\sum_{i\in\mathcal{N}} \mu_i = 1.$

The payoff structure of the chain shown in figure 1 is degenerate, in the sense that the values of the terminal states A and G are deterministically 0 and 1 respectively. This makes the expected value from any state just the probability of absorbing at G.

The estimation system is fed complete sequences $\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots \mathbf{x}_{i_m}$ of observation vectors, together with their scalar terminal value z. It has to generate for every non-terminal state $i \in \mathcal{N}$ a prediction of the expected value $\mathbb{E}[z\,|\,i]$ for starting from that state. If the transition matrix of the Markov chain were completely known, these predictions could be computed as:

$$\mathbb{E}[z\,|\,i] = \sum_{j\in\mathcal{T}} q_{ij}\bar{z}_j + \sum_{j\in\mathcal{N}} q_{ij} \sum_{k\in\mathcal{T}} q_{jk}\bar{z}_k + \sum_{j\in\mathcal{N}} q_{ij} \sum_{k\in\mathcal{N}} q_{jk} \sum_{l\in\mathcal{T}} q_{kl}\bar{z}_l + \ldots . \tag{1}$$

Again, following Sutton, let $[M]_{ab}$ denote the ab^{th} entry of any matrix M, $[\mathbf{u}]_a$ denote the a^{th} component of any vector $\mathbf{u}$. Q denote the square matrix with components $[Q]_{ab} = q_{ab}$, $a, b \in \mathcal{N}$, and $\mathbf{h}$ denote the vector whose components are $[\mathbf{h}]_a = \Sigma_{b\in\mathcal{T}} q_{ab}\bar{z}_b$, for $a \in \mathcal{N}$. Then from equation (1):

$$\mathbb{E}[z\,|\,i] = \left[\sum_{k=0}^{\infty} Q^k \mathbf{h}\right]_i = [(I - Q)^{-1}\mathbf{h}]_i \tag{2}$$

As Sutton showed, the existence of the limit in this equation follows from the fact that Q is the transition matrix for the nonterminal states of an absorbing Markov chain, which, with probability one will ultimately terminate.

During the learning phase, linear TD(λ) generates successive vectors $\mathbf{w}_1^\lambda$, $\mathbf{w}_2^\lambda$, ...,[1] changing $\mathbf{w}^\lambda$ after each complete observation sequence. Define $V_n^\lambda(i) = \mathbf{w}_n^\lambda \cdot \mathbf{x}_i$ as the prediction of the terminal value starting from state i, at stage n in learning. Then, during one such sequence, $V_n^\lambda(i_t)$ are the intermediate predictions of these terminal values, and, abusing notation somewhat, define also $V_n^\lambda(i_{m+1}) = z$, the observed terminal value. Note that in Sutton (1988), Sutton used P_t^n for $V_n^\lambda(i_t)$. TD(λ) changes $\mathbf{w}^\lambda$ according to:

$$\mathbf{w}_{n+1}^\lambda = \mathbf{w}_n^\lambda + \sum_{t=1}^{m} \left\{ \alpha[V_n^\lambda(i_{t+1}) - V_n^\lambda(i_t)] \sum_{k=1}^{t} \lambda^{t-k} \nabla_{\mathbf{w}_n} V_n^\lambda(i_k) \right\} \tag{3}$$

where α is the learning rate.

Sutton showed that TD(1) is just the normal LMS estimator (Widrow & Stearns, 1985), and also proved that the following theorem:

Theorem T For any absorbing Markov chain, for any distribution of starting probabilities μ_i such that there are no inaccessible states, for any outcome distributions with finite expected values $\bar{z}_j$, and for any linearly independent set of observation vectors $\{\mathbf{x}_i\,|\,i \in \mathcal{N}\}$, there exists an $\epsilon > 0$ such that, for all positive $\alpha < \epsilon$ and for any initial weight vector, the predictions of linear TD(λ) (with weight updates after each sequence) converge in expected value to the ideal predictions (2); that is, if $\mathbf{w}_n^\lambda$ denotes the weight vector after n sequences have been experienced, then

$$\lim_{n\to\infty} \mathbb{E}[\mathbf{w}_n^\lambda \cdot \mathbf{x}_i] = \mathbb{E}[z\,|\,i] = [(I - Q)^{-1}\mathbf{h}]_i, \ \forall i \in \mathcal{N}.$$

is true in the case that $\lambda = 0$. This paper proves theorem T for general λ.

2.2. Localist representation

Equation (3) conflates two issues; the underlying TD(λ) algorithm and the representation of the prediction functions V_n^λ. Even though these will remain tangled in the ultimate proof of convergence, it is beneficial to separate them out, since it makes the operation of the algorithm clearer.

Consider representing V_n^λ as a look-up table, with one entry for each state. This is equivalent to choosing a set of vectors $\mathbf{x}_i$ for which just one component is 1 and all the others are 0 for each state, and no two states have the same representation. These trivially satisfy the conditions of Sutton's theorem, and also make the $\mathbf{w}_n$ easy to interpret, as each component is the prediction for just one state. Using them also prevents generalization. For this representation, the terms $\nabla_{\mathbf{w}_n} V_n^\lambda(i_k)$ in the sum

$$\sum_{k=1}^{t} \lambda^{t-k} \nabla_{\mathbf{w}_n} V_n^\lambda(i_k)$$

just reduce to counting the number of times the chain has visited each state, exponentially weighted in recency by λ. In this case, as in the full linear case, these terms do not depend on n, only on the states the chain visits. Define a characteristic function for state j:

$$\chi_j(k) = \begin{cases} 1 & \text{if } i_k = j \\ 0 & \text{otherwise} \end{cases}$$

and the prediction function $V_n^\lambda(i)$ as the entry in the look-up table for state i at stage n during learning. Then equation (3) can be reduced to its elemental pieces

$$V_{n+1}^\lambda(i) = V_n^\lambda(i) + \sum_{t=1}^{m} \left\{ \alpha[V_n^\lambda(i_{t+1}) - V_n^\lambda(i_t)] \sum_{k=1}^{t} \lambda^{t-k} \chi_i(k) \right\} \tag{4}$$

in which the value for each state is updated separately.

To illustrate this process, take the punctate representation of the states B, C, D, E and F in figure 1 to be:[2]

$$\overset{\mathbf{x}_B}{\begin{pmatrix}1\\0\\0\\0\\0\end{pmatrix}} \quad \overset{\mathbf{x}_C}{\begin{pmatrix}0\\1\\0\\0\\0\end{pmatrix}} \quad \overset{\mathbf{x}_D}{\begin{pmatrix}0\\0\\1\\0\\0\end{pmatrix}} \quad \overset{\mathbf{x}_E}{\begin{pmatrix}0\\0\\0\\1\\0\end{pmatrix}} \quad \overset{\mathbf{x}_F}{\begin{pmatrix}0\\0\\0\\0\\1\end{pmatrix}}$$

If the observed sequence is D, C, D, E, F, E, F, G, then the sums

$$\sum_{k=1}^{t} \lambda^{t-k} \nabla_{\mathbf{w}_n} V_n^\lambda(i_k)$$

after each step are:

$$\begin{array}{ccccccc} D & C & D & E & F & E & F \\ \begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix} & \begin{bmatrix}0\\1\\\lambda\\0\\0\end{bmatrix} & \begin{bmatrix}0\\\lambda\\1+\lambda^2\\0\\0\end{bmatrix} & \begin{bmatrix}0\\\lambda^2\\\lambda+\lambda^3\\1\\0\end{bmatrix} & \begin{bmatrix}0\\\lambda^3\\\lambda^2+\lambda^4\\\lambda\\1\end{bmatrix} & \begin{bmatrix}0\\\lambda^4\\\lambda^3+\lambda^5\\1+\lambda^2\\\lambda\end{bmatrix} & \begin{bmatrix}0\\\lambda^5\\\lambda^4+\lambda^6\\\lambda+\lambda^3\\1+\lambda^2\end{bmatrix} \end{array}$$

and component i in this sum is clearly

$$\sum_{k=1}^{t} \lambda^{t-k}\chi_i(k)$$

and time t.

2.3. Contraction mappings

Watkins (1989) showed that a fruitful way of looking at TD estimators is through dynamic programming and its associated contraction mappings. The method starts from the current prediction function $V_n(i)$, $\forall i \in \mathfrak{N}$, shows how to define a whole collection of statistically better estimators $V_{n+1}(i)$, $\forall i \in \mathfrak{N}$ based on an observed sequence, and then uses for TD(λ) a linear combination of these estimators. Except where explicitly noted, this section follows Watkins exactly—the equations developed have their exact analogues for the linear representation, as will be seen in section 2.5.

Imagine the chain starts at some state i_0, and runs forward through states i_1, i_2, $\ldots$, ultimately absorbing. Define the r-step estimate of i_0 as *either* the estimate $V_n(i_r)$ of state i_r, if the chain is not absorbed after r steps and so $i_r \in \mathfrak{N}$, *or* the observed terminal value z of the sequence, if the chain has absorbed before this time.

Formally, define the random variables

$$\begin{aligned} V^1_{n,i_0} &= \begin{cases} V_n(i_1) & \text{if } i_1 \in \mathfrak{N} \\ z & \text{otherwise} \end{cases} \\ V^2_{n,i_0} &= \begin{cases} V_n(i_2) & \text{if } i_2 \in \mathfrak{N} \\ z & \text{otherwise} \end{cases} \\ &\vdots \\ V^r_{n,i_0} &= \begin{cases} V_n(i_r) & \text{if } i_r \in \mathfrak{N} \\ z & \text{otherwise} \end{cases} \\ &\vdots \end{aligned} \tag{5}$$

where i_1 is the first state accessed by the Markov chain in one particular sequence starting from i_0; i_2 is the second, and so on, and z is the observed terminal value if the chain gets absorbed before time step r is reached. These are random variables, since they depend on the particular sequence of states which will be observed. Naturally, they also depend on the values $V_n(i)$.

The point of these is that if the chain absorbs after completing s steps from i_0, then each of the V^r_{n,i_0}, for $r \geq s$ will be based on the terminal value provided by the world, rather than one derived from the 'bootstraps' $V_n(i)$. V^r_{n,i_0} should therefore on average be more accurate than V_n and so can be used incrementally to improve it. This can be shown by looking at the difference between $\mathbb{E}[V^r_{n,i_0}]$ and $\mathbb{E}[z \mid i_0]$, the ideal predictions.

Here:

$$\mathbb{E}[V^r_{n,i_0}] = \sum_{t_1 \in \mathfrak{T}} q_{i_0 t_1} \bar{z}_{t_1} + \sum_{i_1 \in \mathfrak{N}} Q_{i_0 i_1} \sum_{t_2 \in \mathfrak{T}} q_{i_1 t_2} \bar{z}_{t_2} + \dots + \sum_{i_{r-1} \in \mathfrak{N}} Q^{r-1}_{i_0 i_{r-1}} \sum_{t_r \in \mathfrak{T}} q_{i_{r-1} t_r} \bar{z}_{t_r} + \sum_{i_r \in \mathfrak{N}} Q^r_{i_0 i_r} V_n(i_r) \tag{6}$$

whereas it can easily be shown that

$$\mathbb{E}[z \mid i_0] = \sum_{t_1 \in \mathfrak{T}} q_{i_0 t_1} \bar{z}_{t_1} + \sum_{i_1 \in \mathfrak{N}} Q_{i_0 i_1} \sum_{t_2 \in \mathfrak{T}} q_{i_1 t_2} \bar{z}_{t_2} + \dots + \tag{7}$$

$$\sum_{i_{r-1} \in \mathfrak{N}} Q^{r-1}_{i_0 i_{r-1}} \sum_{t_r \in \mathfrak{T}} q_{i_{r-1} t_r} \bar{z}_{t_r} + \sum_{i_r \in \mathfrak{N}} Q^r_{i_0 i_r} \mathbb{E}[z \mid i_r] \tag{8}$$

Therefore,

$$\mathbb{E}[V^r_{n,i_0}] - \mathbb{E}[z \mid i_0] = \sum_{i_r \in \mathfrak{N}} Q^r_{i_0 i_r} (V_n(i_r) - \mathbb{E}[z \mid i_r]) \tag{9}$$

Watkins actually treated a slightly different case, in which the target values of the predictors are based on *discounted* future values whose contribution diminishes exponentially with the time until they happen. In this case it is easier to see how the reduction in error is brought about. His analogue of equation (9) was:

$$\mathbb{E}[V^r_{n,i_0}] - \mathbb{E}[z \mid i_0] = \gamma^r \sum_{i_r \in \mathfrak{N}} Q^r_{i_0 i_r} (V_n(i_r) - \mathbb{E}[z \mid i_r])$$

where $\gamma < 1$ is the discount factor. Since

$$\sum_{i_r \in \mathfrak{N}} Q^r_{i_0 i_r} \leq 1,$$

as Q is the matrix of a Markov chain, Watkins could guarantee that

$$\max_{i_0} |\mathbb{E}[V^r_{n,i_0}] - \mathbb{E}[z | i_0]| \leq \gamma^r \max_{i_r} |V_n(i_r) - \mathbb{E}[z | i_r]|$$

which provides a (weak) guarantee that the error of V^r_n will be less than that of V_n.

The nondiscounted case, when $\gamma = 1$, is somewhat different. Here, for some initial states i_0, there is a nonzero probability that the chain will absorb before finishing r steps. In this case, the value of V^r_{n,i_0}, being z, will be unbiased, and so should provide for error reduction. Even if the chain does not absorb, its value can be no farther from what it should be than is the most inaccurate component of V_n. Although there is no error reduction due to γ, it is guaranteed that

$$\sum_{i_r \in \mathfrak{N}} Q^r_{i_0 i_r} \leq 1$$

with inequality for all those states from which it is possible to absorb within r steps.

This does not ensure that

$$\max_{i_0} |\mathbb{E}[V^r_{n,i_0}] - \mathbb{E}[z | i_0]| < \max_{i_r} |V_n(i_r) - \mathbb{E}[z | i_r]|$$

since the maximum could be achieved, pathologically, at a state from which it is impossible to absorb in only r steps. However, the estimates for the states that are within r steps of absorption will, on average, improve, and this should, again on average, filter back to the other states.

Watkins demonstrated that TD(λ) is based on a weighted average of the V^a_{n,i_0}. Consider

$$V^\lambda_{n,i_0} = (1 - \lambda) \sum_{a=1}^{\infty} \lambda^{a-1} V^a_{n,i_0} \tag{10}$$

which is also a valid estimator of the terminal value starting at i_0. He points out that in choosing the value of λ, there is a tradeoff between the bias caused by the error in V_n, and the variance of the real terminal value z. The higher λ, the more significant are the V^r_n for higher values of r, and the more effect the unbiased terminal values will have. This leads to higher variance and lower bias. Conversely, the lower λ, the less significant are the contributions from higher values of r, and the less the effect of the unbiased terminal values. This leads to smaller variance and greater bias.

It remains to be shown that TD(λ) is indeed based on this combination estimator. Expanding out the sum in equation (10).

$$V^\lambda_{n,i_0} - V_n(i_0) = [V_n(i_1) - V_n(i_0)] + \lambda[V_n(i_2) - V_n(i_1)] + \lambda^2[V_n(i_3) - V_n(i_2)] + \ldots \tag{11}$$

defining $V_n(i_s) = z$ for $s > \max\{t | i_t \in \mathfrak{N}\}$.

The whole point of defining V^{λ}_{n,i_0} is so that it can be used to make V more accurate. The obvious incremental update rule to achieve this has

$$V_{n+1}(i_0) = V_n(i_0) + \alpha[V^{\lambda}_{n,i_0} - V_n(i_0)]. \tag{12}$$

From equation (11) it is apparent that the changes to $V_n(i_0)$ involve summing future values of $V_n(i_{t+1}) - V_n(i_t)$ weighted by powers of λ. Again following Watkins, these differences can be calculated through an activity trace based on the characteristic functions $\chi_i(t)$ that were defined earlier as a way of counting how often and how recently the chain has entered particular states. Using index t for the members of the observed sequence, the on-line version of the TD(λ) rule has

$$V_{t+1}(i) = V_t(i) + \alpha[V_t(i_{t+1}) - V_t(i_t)] \sum_{k=1}^{t} \lambda^{t-k}\chi_i(k). \tag{13}$$

For the problem that Sutton treated, the change to V_n is applied off-line, after a complete sequence through the chain. Therefore, if the states through which the chain passes on one sequence are $i_0, i_1, \ldots, i_{m-1} \in \mathfrak{N}$, and $i_m \in \mathfrak{T}$, it absorbs with terminal value $V_n(i_m) \equiv z$, and V_{n+1} is the new estimator after experiencing the sequence, then

$$\begin{aligned}
V_{n+1}(i_0) &= V_n(i_0) \quad + \sum_{t=1}^{m} \alpha[V_n(i_{t+1}) - V_n(i_t)] \sum_{k=1}^{t} \lambda^{t-k}\chi_{i_0}(k) \\
V_{n+1}(i_1) &= V_n(i_1) \quad + \sum_{t=2}^{m} \alpha[V_n(i_{t+1}) - V_n(i_t)] \sum_{k=1}^{t} \lambda^{t-k}\chi_{i_1}(k) \\
&\ \ \vdots \\
V_{n+1}(i_{m-1}) &= V_n(i_{m-1}) + \qquad \alpha[z - V_n(i_{m-1})] \sum_{k=1}^{m} \lambda^{t-k}\chi_{i_{m-1}}(k),
\end{aligned}$$

summing over terms where $i_a = i_b$ (so $\chi_{i_a} \equiv \chi_{i_b}$). Note that these expressions are exactly the same as the TD(λ) weight change formula in equation (4).

Thus, the actual TD(λ) algorithm is based on the exponentially weighted sum defined in equation (10) of the outcomes of the V^r_i random variables. The mean contraction properties of these variables will therefore determine the mean contraction properties of the overall TD(λ) estimator.

2.4. Linear representation

The previous subsection considered the TD(λ) algorithm isolated from the representation Sutton used. Although a number of different representations might be employed, the simplest is the linear one he adopted. Identifying the vectors **x** with the states they represent gives

$$V_n(\mathbf{x}) = \mathbf{w}_n \cdot \mathbf{x}$$

where $\mathbf{w}_n$ is the weight vector at stage n of learning.

The basic algorithm is concerned with the V_n^λ predictor random variables rather than how their values can be used to change the initial predictor V_n. Under the new representation, equation (12) no longer makes sense since the states cannot be separated in the appropriate manner. Rather, the information about the error has to be used to update all the weights on which it depends. The appropriate formula, derived from the delta-rule is

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \alpha[V_{n,i_0}^\lambda - V_n(i_0)]\, \nabla_{\mathbf{w}_n} V_n(i_0)$$

weighting the error due to state i_0 by the vector representation of i_0. Then the equivalent of equation (13) is just Sutton's main TD(λ) equation (3).

More sophisticated representations such as kd-trees (see Omohundro (1987) for a review) or CMACs (Albus, 1975) may lead to faster learning and better generalization, but each requires a separate convergence proof. Dayan (1991) compares the qualities of certain different representations for Barto, Sutton and Watkins' grid task (Barto, Sutton & Watkins, 1990).

2.5. *The proof of theorem T*

The strategy for proving theorem T is to follow Sutton (1988) in considering the expected value of the new prediction weight vector given the observation of a complete sequence, and to follow Watkins in splitting this change into the components due to the equivalents of the V^r random variables, and then summing them. Mean error reduction over iterations will be assured by the equivalent of equation (9) for the linear representation.

Define the $V_{\cdot,\cdot}^r$ random variables as in equation (5) as

$$V_{n,i_0}^r = \begin{cases} \mathbf{w}_n^r \cdot \mathbf{x}_{i_r} & \text{if } \mathbf{x}_{i_r} \in \mathfrak{N} \\ z & \text{otherwise} \end{cases}$$

where $\mathbf{x}_i$ are identified with the states in the observed sequence, $\mathbf{w}_n^r$ is the current weight vector defining the estimated terminal values, and z is the actual value. Then, after observing the whole sequence, $\mathbf{w}_n^r$ is updated as:

$$\begin{aligned} \mathbf{w}_{n+1}^r &= \mathbf{w}_n^r + \alpha \sum_{\mathbf{x}_i \in \mathfrak{N} \text{ visited}} [V_{n,i}^r - V_n(i)]\, \nabla_{\mathbf{w}_n} V_n(i) \\ &= \mathbf{w}_n^r + \alpha \sum_{\mathbf{x}_i \in \mathfrak{N} \text{ visited}} [V_{n,i}^r - \mathbf{w}_n \cdot \mathbf{x}_i]\, \mathbf{x}_i. \end{aligned} \tag{14}$$

An exact parallel of Sutton's proof procedure turns out to apply to $\mathbf{w}^r$. Define η_{ij}^s as the number of times the s-step transition

$$\mathbf{x}_i \rightarrow \mathbf{x}_{k_1} \rightarrow \mathbf{x}_{k_2} \ldots \rightarrow \mathbf{x}_{k_{s-1}} \rightarrow \mathbf{x}_j$$

occurs, for any intermediate states $\mathbf{x}_{k_t} \in \mathfrak{N}$.

The sum in equation (14) can be regrouped in terms of source and destination states of the transitions:

$$
\begin{aligned}
\mathbf{w}_{n+1}^r = \mathbf{w}_n^r &+ \alpha \sum_{i \in \mathfrak{N}} \sum_{j_r \in \mathfrak{N}} \eta_{ij_r}^r [\mathbf{w}_n^r \cdot \mathbf{x}_{j_r} - \mathbf{w}_n^r \cdot \mathbf{x}_i]\, \mathbf{x}_i \\
&+ \alpha \sum_{i \in \mathfrak{N}} \sum_{j_r \in \mathfrak{T}} \eta_{ij_r}^r [z_{j_r} - \mathbf{w}_n^r \cdot \mathbf{x}_i]\, \mathbf{x}_i \\
&+ \alpha \sum_{i \in \mathfrak{N}} \sum_{j_{r-1} \in \mathfrak{T}} \eta_{ij_{r-1}}^{r-1} [z_{j_{r-1}} - \mathbf{w}_n^r \cdot \mathbf{x}_i]\, \mathbf{x}_i \qquad (15) \\
&\;\vdots \\
&+ \alpha \sum_{i \in \mathfrak{N}} \sum_{j_1 \in \mathfrak{T}} \eta_{ij_1}^1 [z_{j_1} - \mathbf{w}_n^r \cdot \mathbf{x}_i]\, \mathbf{x}_i
\end{aligned}
$$

where z_j indicates that the terminal value is generated from the distribution due to state j, and the extra terms are generated by the possibility that, from visiting any $\mathbf{x}_i \in \mathfrak{N}$, the chain absorbs before taking r further steps.

Taking expected values over sequences, for $i \in \mathfrak{N}$

$$
\begin{aligned}
\mathbb{E}[\eta_{ij}^r] &= d_i Q_{ij}^r && \text{for } j \in \mathfrak{N} \\
\mathbb{E}[\eta_{ij}^r] &= \sum_{k \in \mathfrak{N}} d_i Q_{ik}^{r-1} q_{kj} && \text{for } j \in \mathfrak{T} \\
\mathbb{E}[\eta_{ij}^{r-1}] &= \sum_{k \in \mathfrak{N}} d_i Q_{ik}^{r-2} q_{kj} && \text{for } j \in \mathfrak{T} \\
&\;\vdots \\
\mathbb{E}[\eta_{ij}^1] &= d_i q_{ij} && \text{for } j \in \mathfrak{T}
\end{aligned}
$$

where d_i is the expected number of times the Markov chain is in state i in one sequence. For an absorbing Markov chain, it is known that the dependency of this on the probabilities μ_i of starting in the various states is:

$$
d_i = \sum_{j \in \mathfrak{N}} \mu_j (I - Q)_{ji}^{-1} = [\mu^T (I - Q)^{-1}]_i \qquad (16)
$$

Substituting into equation (15), after taking expectations on both sides, noting that the dependence of $\mathbb{E}[\mathbf{w}_{n+1}^r \,|\, \mathbf{w}_n^r]$ on $\mathbf{w}_n^r$ is linear, and using $\bar{\mathbf{w}}$ to denote expected values, a close relation of equation (6) emerges for the linear representation:

$$\begin{aligned}\bar{\mathbf{w}}_{n+1}^r = \bar{\mathbf{w}}_n^r + \alpha \sum_{i\in\mathfrak{N}} d_i\mathbf{x}_i \Bigg[& \sum_{j_r\in\mathfrak{N}} Q_{ij_r}^r(\mathbf{x}_{j_r} \cdot \bar{\mathbf{w}}_n^r) \\ & - (\mathbf{x}_i \cdot \bar{\mathbf{w}}_n^r) \left\{ \sum_{j_r\in\mathfrak{N}} Q_{ij_r}^r + \sum_{\substack{j_r\in\mathfrak{I},\\ k\in\mathfrak{N}}} Q_{ik}^{r-1}q_{kj_r} + \ldots + \sum_{j_1\in\mathfrak{I}} q_{ij_1} \right\} \\ & + \sum_{\substack{j_r\in\mathfrak{I},\\ k\in\mathfrak{N}}} Q_{ik}^{r-1}q_{kj_r}\bar{z}_{j_r} + \sum_{\substack{j_{r-1}\in\mathfrak{I},\\ k\in\mathfrak{N}}} Q_{ik}^{r-2}q_{kj_{r-1}}\bar{z}_{j_{r-1}} + \ldots + \sum_{j_1\in\mathfrak{I}} q_{ij_1}\bar{z}_{j_1} \Bigg] .\end{aligned}$$

Define X to be the matrix whose columns are $\mathbf{x}_i$, so $[X]_{ab} = [\mathbf{x}_a]_b$, and D to be the diagonal matrix $[D]_{ab} = \delta_{ab}d_a$, where δ_{ab} is the Kronecker delta. Remembering that $h_i = \Sigma_{j\in\mathfrak{I}}\, q_{ij}\bar{z}_j$, and converting to matrix form

$$\bar{\mathbf{w}}_{n+1}^r = \bar{\mathbf{w}}_n^r + \alpha XD[Q^rX^{\mathrm{T}}\bar{\mathbf{w}}_n^r - X^{\mathrm{T}}\bar{\mathbf{w}}_n^r + (Q^{r-1} + Q^{r-2} + \ldots + I)\mathbf{h}] \tag{17}$$

since

$$\sum_{j_r\in\mathfrak{N}} Q_{ij_r}^r + \sum_{\substack{j_r\in\mathfrak{I},\\ k\in\mathfrak{N}}} Q_{ik}^{r-1}q_{kj_r} + \ldots + \sum_{j_1\in\mathfrak{I}} q_{ij_1} = 1$$

as this covers all the possible options for r-step moves from state i.

Define the correct predictions $[\bar{\mathbf{e}}^*]_i = \mathbb{E}[z\,|\,i]$; then also, from equation (2),

$$\begin{aligned}\bar{\mathbf{e}}^* &= [\mathbb{E}[z\,|\,i]] \\ &= \mathbf{h} + Q\mathbf{h} + Q^2\mathbf{h} + \ldots \\ &= (I + Q + Q^2 + \ldots + Q^{r-1})\mathbf{h} + Q^r(I + Q + Q^2 + \ldots + Q^{r-1})\mathbf{h} + \ldots \\ &= \sum_{k=0}^{\infty} [Q^r]^k(I + Q + Q^2 + \ldots + Q^{r-1})\mathbf{h} \\ &= (I - Q^r)^{-1}(I + Q + Q^2 + \ldots + Q^{r-1})\mathbf{h}\end{aligned} \tag{18}$$

where the sum converges since the chain is absorbing. This is another way of writing equation (7).

Multiplying equation (17) on the left by X^{T},

$$\begin{aligned}X^{\mathrm{T}}\bar{\mathbf{w}}_{n+1}^r &= X^{\mathrm{T}}\bar{\mathbf{w}}_n^r + \alpha X^{\mathrm{T}}XD[(I + Q + Q^2 + \ldots + Q^{r-1})\mathbf{h} + Q^rX^{\mathrm{T}}\bar{\mathbf{w}}_n^r - X^{\mathrm{T}}\bar{\mathbf{w}}_n^r] \\ &= [I - \alpha X^{\mathrm{T}}\, XD(I - Q^r)]X^{\mathrm{T}}\bar{\mathbf{w}}_n^r + \alpha X^{\mathrm{T}}XD(I + Q + Q^2 + \ldots + Q^{r-1})\mathbf{h}\end{aligned}$$

Subtracting $\bar{\mathbf{e}}^*$ from both sides of the equation, and noting that from equation (18) $(I - Q^r)\bar{\mathbf{e}}^* = (I + Q + Q^2 + \ldots + Q^{r-1})\mathbf{h}$, this gives the update rule, which is the equivalent of equation (9):

$$[X^{\mathrm{T}}\bar{\mathbf{w}}_{n+1}^r - \bar{\mathbf{e}}^*] = [I - \alpha X^{\mathrm{T}}XD(I - Q^r)]X^{\mathrm{T}}\bar{\mathbf{w}}_n^r + \alpha X^{\mathrm{T}}XD(I - Q^r)\bar{\mathbf{e}}^* - \bar{\mathbf{e}}^*$$

$$= [I - \alpha X^{\mathrm{T}}XD(I - Q^r)][X^{\mathrm{T}}\bar{\mathbf{w}}_n^r - \bar{\mathbf{e}}^*].$$

The Watkins construction of TD(λ) developed in equation (10) in the previous section reveals that, starting from $\mathbf{w}_n^r = \mathbf{w}_n^\lambda$, $\forall r$,

$$\mathbf{w}_{n+1}^\lambda = (1 - \lambda) \sum_{r=1}^{\infty} \lambda^{r-1}\mathbf{w}_{n+1}^r$$

Therefore, since for $0 < \lambda < 1$, $(1 - \lambda)\Sigma_{r=1}^{\infty} \lambda^{r-1} = 1$,

$$[X^{\mathrm{T}}\bar{\mathbf{w}}_{n+1}^\lambda - \bar{\mathbf{e}}^*] = \left\{(1 - \lambda) \sum_{r=1}^{\infty} \lambda^{r-1}[I - \alpha X^{\mathrm{T}}XD(I - Q^r)]\right\} [X^{\mathrm{T}}\bar{\mathbf{w}}_n^\lambda - \bar{\mathbf{e}}^*]$$

$$= \{I - \alpha X^{\mathrm{T}}XD(I - (1 - \lambda)Q[I - \lambda Q]^{-1})\}\ [X^{\mathrm{T}}\bar{\mathbf{w}}_n^\lambda - \bar{\mathbf{e}}^*]$$

where $\bar{\mathbf{w}}^\lambda$ are the expected weights from the TD(λ) procedure. The sum

$$(1 - \lambda) \sum_{r=1}^{\infty} \lambda^{r-1}Q^r = (1 - \lambda)Q[I - \lambda Q]^{-1} \tag{19}$$

converges since $0 < \lambda < 1$.

Define

$$\Delta_\lambda = I - \alpha X^T XD(I - (1 - \lambda)Q[I - \lambda Q]^{-1})$$

then the truth of theorem T will be shown if it can be demonstrated that $\exists\epsilon > 0$ such that for $0 < \alpha < \epsilon$, $\lim_{n\to\infty}\Delta_\lambda^n = 0$. For then $[X^{\mathrm{T}}\bar{\mathbf{w}}_n^r - \bar{\mathbf{e}}^*] \to \mathbf{0}$ as $n \to \infty$, and all the estimates will tend to be correct.

Almost all of Sutton's (1988) proof of this applies *mutatis mutandis* to the case that $\lambda \neq 0$, always provided the crucial condition holds that X has full rank. For completeness, the entire proof is given in the appendix. Overall it implies that the expected values of the estimates will converge to their desired values as more sequences are observed under the conditions stated in theorem T.

2.6 Non-independence of the $\mathbf{x}_i$

In moving from Watkins' representation-free proof to Sutton's treatment of the linear case, one assumption was that the $\mathbf{x}_i$, the vectors representing the states, were independent. If they are not, so that matrix X does not have full rank, the proof breaks down. $D(I - (1 - \lambda)Q[I - \lambda Q]^{-1})$ is still positive, however $X^{\mathrm{T}}XD(I - (1 - \lambda)Q[I - \lambda Q]^{-1})$ will no longer have a full set of eigenvalues with positive real parts, since the null subspace

$$Y = \{\mathbf{y} \,|\, XD(I - (1 - \lambda)Q[I - \lambda Q]^{-1})\mathbf{y} = \mathbf{0}\} \neq \{\mathbf{0}\}$$

is not empty. Any nonzero member of this is an eigenvector with eigenvalue 0 of $X^{\mathrm{T}}XD(I - (1 - \lambda)Q[I - \lambda Q]^{-1})$.

Saying what will happen to the expected values of the weights turns out to be easier than understanding it. Choose a basis:

$$\{\mathbf{b}_1, \ldots, \mathbf{b}_p, \mathbf{b}_{p+1}, \ldots, \mathbf{b}_n\} \text{ for } \mathcal{R}^n,$$

with $\mathbf{b}_i \in Y$, for $1 \leq i \leq p$ being a basis for Y.

Then the proof in the appendix applies exactly to $\mathbf{b}_{p+1}, \ldots, \mathbf{b}_n$; that is there exists some $0 < \epsilon < 1$ such that:

$$\lim_{n\to\infty} [I - \alpha X^{\mathrm{T}}XD(I - (1 - \lambda)Q[I - \lambda Q]^{-1})]^n \, \mathbf{b}_i = \mathbf{0}, \text{ for } p < i \leq n, \text{ and } 0 < \alpha < \epsilon.$$

Also,

$$[I - \alpha X^{\mathrm{T}}XD(I - (1 - \lambda)Q[I - \lambda Q]^{-1})]^n \, \mathbf{b}_i = \mathbf{b}_i, \text{ for } 1 \leq i \leq p$$

by the definition of Y.

Writing

$$X^{\mathrm{T}}\bar{\mathbf{w}}_0^{\lambda} - \bar{\mathbf{e}}^* = \sum_{i=1}^{n} \beta_i \mathbf{b}_i,$$

then

$$\begin{aligned} X^{\mathrm{T}}\bar{\mathbf{w}}_n^{\lambda} - \bar{\mathbf{e}}^* &= [I - \alpha X^{\mathrm{T}}XD(I - Q^r)]^n[X^{\mathrm{T}}\bar{\mathbf{w}}_0{}^{\lambda} - \bar{\mathbf{e}}^*] \\ &= [I - \alpha X^{\mathrm{T}}XD(I - Q^r)]^n \left[\sum_{i=1}^{n} \beta_i \mathbf{b}_i \right] \\ &\to \sum_{i=1}^{p} \beta_i \mathbf{b}_i, \text{ as } n \to \infty \end{aligned}$$

and so

$$XD(I - (1 - \lambda)Q[I - \lambda Q]^{-1})[X^{\mathrm{T}}\bar{\mathbf{w}}_n^{\lambda} - \bar{\mathbf{e}}^*] \to \mathbf{0} \text{ as } n \to \infty. \tag{20}$$

To help understand this result, consider the equivalent for the LMS rule, TD(1). There

$$XD[X^{\mathrm{T}}\bar{\mathbf{w}}_n^{1} - \bar{\mathbf{e}}^*] \to \mathbf{0} \text{ as } n \to \infty. \tag{21}$$

and so, since D is symmetric,

$$\frac{\partial}{\partial \bar{\mathbf{w}}_n^{1}} [X^{\mathrm{T}}\bar{\mathbf{w}}_n^{1} - \bar{\mathbf{e}}^*]^{\mathrm{T}} D [X^{\mathrm{T}}\bar{\mathbf{w}}_n^{1} - \bar{\mathbf{e}}^*] = X(D + D^{\mathrm{T}})[X^{\mathrm{T}}\bar{\mathbf{w}}_n^{1} - \bar{\mathbf{e}}^*] \tag{22}$$

$$= 2XD[X^{\mathrm{T}}\bar{\mathbf{w}}_n^{1} - \bar{\mathbf{e}}^*] \tag{23}$$

$$\to \mathbf{0} \text{ as } n \to \infty, \tag{24}$$

by equation (21). For weights $\mathbf{w}$, the square error for state i is $|[X^{\mathrm{T}}\mathbf{w} - \bar{\mathbf{e}}^*]|_i^2$, and the expected number of visits to i in one sequence is d_i. Therefore the quadratic form

$$[X^T\mathbf{w} - \bar{\mathbf{e}}^*]^T D [X^T\mathbf{w} - \bar{\mathbf{e}}^*]$$

is just the loaded square error between the predictions at each state and their desired values, where the loading factors are just the expected frequencies with which the Markov chain hits those states. The condition in equation (24) implies that the expected values of the weights tend to be so as to minimize this error.

This does not happen in general for $\lambda \neq 1$. Intuitively, bias has returned to haunt. For the case where X is full rank, Sutton shows that it is harmless to use the inacurrate estimates from the next state $\mathbf{x}_{i_{t+1}}\cdot\mathbf{w}$ to criticize the estimates for the current state $\mathbf{x}_{i_t}\cdot\mathbf{w}$. Where X is not full rank, these successive estimates become biased on account of what might be deemed their 'shared' representation. The amount of extra bias is then related to the amount of sharing and the frequency with which the transitions happen from one state to the next.

Formalizing this leads to a second issue; the interaction between the two statistical processes of calculating the mean weight and calculating the expected number of transitions. Comparing equations (20) and (21), one might expect

$$\lim_{n \to \infty} \frac{\partial}{\partial \bar{\mathbf{w}}_n^{\lambda}} [X^{\mathrm{T}}\bar{\mathbf{w}}_n^{\lambda} - \bar{\mathbf{e}}^*]^{\mathrm{T}} D(I - (1 - \lambda)Q[I - \lambda Q]^{-1})[X^{\mathrm{T}}\bar{\mathbf{w}}_n^{\lambda} - \bar{\mathbf{e}}^*] = \mathbf{0} \tag{25}$$

However, the key step in proving equation (24) was the transition between equations (22) and (23), which relied on the symmetry of D. Since Q is not in general symmetric, this will not happen.

Defining

$$\begin{aligned} \mathbf{g}(\mathbf{w}') &= \frac{\partial}{\partial \mathbf{w}} [X^{\mathrm{T}}\mathbf{w} - \bar{\mathbf{e}}^*]^{\mathrm{T}} D(I - (1 - \lambda)Q[I - \lambda Q]^{-1})[X^{\mathrm{T}}\mathbf{w}' - \bar{\mathbf{e}}^*] \\ &= XD(I - (1 - \lambda)Q[I - \lambda Q]^{-1})[X^{\mathrm{T}}\mathbf{w}' - \bar{\mathbf{e}}^*] \end{aligned} \tag{26}$$

all that will actually happen is that $\mathbf{g}(\bar{\mathbf{w}}_n^\lambda) \rightarrow \mathbf{0}$ as $n \rightarrow \infty$.

Although the behavior described by equation (25) is no more satisfactory than that described by equation (26), it is revealing to consider what happens if one attempts to arrange for it to hold. This can be achieved by 'completing' the derivative, i.e., by having a learning rule whose effect is

$$[X^{\mathrm{T}}\bar{\mathbf{w}}_{n+1}^\lambda - \bar{\mathbf{e}}^*] =$$

$$\left\{ I - \alpha X^T X \left(D - \frac{1 - \lambda}{2} [DQ(I - \lambda Q)^{-1} + (I - \lambda Q^{\mathrm{T}})^{-1} Q^{\mathrm{T}} D^{\mathrm{T}}] \right) \right\} [X^{\mathrm{T}}\bar{\mathbf{w}}_n^\lambda - \bar{\mathbf{e}}^*]$$

The Q^T term effectively arranges for backwards as well as forwards learning to occur, so that not only would state i_t adjust its estimate to make it more like state i_{t+1}, but also state i_{t+1} would adjust its estimate to make it more like state i_t.

Werbos (1990) and Sutton (personal communication) both discussed this point in the context of the gradient descent of TD(λ) rather than its convergence for non-independent $\mathbf{x}_i$. Werbos presented an example based on a learning technique very similar to TD(0), in which completing the derivative in this manner makes the rule converge away from the true solution. He faulted this procedure for introducing the unhelpful correlations between the learning rule and the random moves from one state to the next which were mentioned above. He pointed out the convergence in terms of functions $\mathbf{g}$ in equation (26) in which the $\mathbf{w}'$ weights are fixed.

Sutton presented an example to help explain the result. At first sight, augmenting TD(λ) seems quite reasonable; after all it could quite easily happen by random chance of the training sequences that the predictions for one state are more accurate than the predictions for the next at some point. Therefore, training the second to be more like the first would be helpful. However, Sutton pointed out that time and choices always move forward, not backwards. Consider the case shown in figure 2, where the numbers over the arrows represent the transition probabilities, and the numbers at the terminal nodes represent terminal absorbing values.

Here, the value of state A is reasonably 1/2, as there is 50% probability of ending up at either Y or Z. The value of state B, though, should be 1, as the chain is certain to end up at Y. Training forwards will give this, but training backwards too will make the value of B tend to 3/4. In Werbos' terms, there are correlations between the weights and the possible transitions that count against the augmented term. Incidentally, this result does not affect TD(1), because the training values, being just the terminal value for the sequence, bear no relation to the transitions themselves, just the number of times each state is visited.

Coming back to the case where X is not full rank. TD(λ) for $\lambda \neq$ will still converge, but away from the 'best' value, to a degree that is determined by the matrix

$$(I - (1 - \lambda)Q[I - \lambda Q]^{-1}).$$

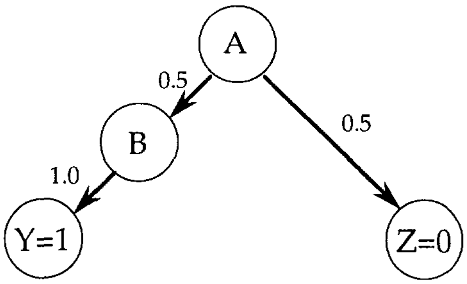

Figure 2. Didactic example of the pitfalls of backwards training. If Y and Z are terminal states with values 1 and 0 respectively, what values should be assigned to states A and B respectively.

3. Convergence with probability one

Sutton's proof and the proofs in the previous section accomplish only the nadir of stochastic convergence, *viz* convergence of the mean, rather than the zenith, *viz* convergence with probability one. Watkins (1989) proved convergence with probability one for a form of prediction and action learning he called $\mathcal{Q}$-learning. This section shows that this result can be applied almost directly to the discounted predictive version of TD(0), albeit without the linear representations, and so provides the first strong convergence proof for a temporal difference method.

Like dynamic programming (DP), $\mathcal{Q}$-learning combines prediction and control. Consider a controlled, discounted, nonabsorbing Markov-process, i.e., one in which at each state $i \in \mathfrak{N}$ there is a finite set of possible actions $a \in \mathcal{A}$. Taking one action leads to an immediate reward, which is a random variable $r_i(a)$ whose distribution depends on both i and a, and a stochastic transition according to a Markov matrix $\mathcal{P}_{ij}(a)$ for $j \in \mathfrak{N}$. If an agent has some policy $\pi(i) \in \mathcal{A}$, which determines which action it would perform at state i, then, defining the value of i under π as $V^\pi(i)$, this satisfies:

$$V^\pi(i) = \mathbb{E}[r_i(\pi(i))] + \gamma \sum_{j \in \mathfrak{N}} \mathcal{P}_{ij}(\pi(i)) V^\pi(j),$$

where γ is the discount factor. Define the $\mathcal{Q}$ value of state i and action a under policy π as:

$$\mathcal{Q}^\pi(i, a) = \mathbb{E}[r_i(a)] + \gamma \sum_{j \in \mathfrak{N}} \mathcal{P}_{ij}(a) V^\pi(j),$$

which is the value of taking action a at i followed by policy π thereafter. Then the theory of DP (Bellman & Dreyfus, 1962) implies that a policy which is at least as good as π is to take the action a^* at state i where $a^* = \text{argmax}_b\{\mathcal{Q}^\pi(i, b)\}$, and to follow π in all other states. In this fact lies the utility of the $\mathcal{Q}$ values. For discounted problems, it turns out that there is at least one optimal policy π^*; define $\mathcal{Q}^*(i, a) = \mathcal{Q}^{\pi^*}(i, a)$.

$\mathcal{Q}$-learning is a method of determining $\mathcal{Q}^*$, and hence an optimal policy, based on exploring the effects of actions at states. Consider a sequence of observations (i_n, a_n, j_n, z_n), where the process at state i_n is probed with action a_n, taking it to state j_n and giving reward z_n. Then define recursively:

$$\mathcal{Q}_{n+1}(i, a) = \begin{cases} (1 - \alpha_n)\mathcal{Q}_n(i, a) + \alpha_n(z_n + \gamma U_n(j_n)) & \text{if } i = i_n \text{ and } a = a_n, \\ \mathcal{Q}_n(i, a) & \text{otherwise} \end{cases} \tag{27}$$

for any starting values $\mathcal{Q}_0(i, a)$, where $U_n(j_n) = \max_b\{\mathcal{Q}_n(j_n, b)\}$. The α_n are a set of learning rates that obey standard stochastic convergence criteria:

$$\sum_{k=1}^{\infty} \alpha_{n^k(i,a)} = \infty, \sum_{k=1}^{\infty} \alpha^2_{n^k(i,a)} < \infty, \forall i \in \mathfrak{N}, a \in \mathfrak{A} \tag{28}$$

where $n^k(i, a)$ is the k^{th} time that $i_n = i$ and $a_n = a$. Watkins (1989) proved that if, in addition, the rewards are all bounded, then, with probability one:

$$\lim_{n\to\infty} \mathcal{Q}_n(i, a) = \mathcal{Q}^*(i, a),$$

Consider a degenerate case of a controlled Markov process in which there is only one action possible from every state. In that case, the $\mathcal{Q}^\pi$, V^π, and the (similarly defined) U^π values are exactly the same and equal to $\mathcal{Q}^*$, and equation (27) is exactly the on-line form of TD(0) for the case of a nonabsorbing chain in which rewards (i.e., the terminal values discussed above in the context of absorbing Markov chains) arrive from every state rather than just some particular set of absorbing states. Therefore, under the conditions of Watkins' theorem, the on-line version of TD(0) converges to the correct predictions, with probability one.

Although clearly a TD procedure, there are various differences between this and the one described in the previous section. Here, learning is on-line, that is the $V(=\mathcal{Q})$ values are changed for every observation. Also, learning need not proceed along an observed sequence—there is no requirement that $j_n = i_{n+1}$, and so uncoupled or disembodied moves can be used.[3] The conditions in equation (28) have as a consequence that every state must be visited infinitely often. Also note that Sutton's proof, since it is confined to showing convergence in the mean, works for a fixed learning rate α, whereas Watkins', in common with other stochastic convergence proofs, requires α_n to tend to 0.

Also, as stated, the $\mathcal{Q}$-learning theorem only applies to discounted, nonabsorbing, Markov chains, rather than the absorbing ones with γ=1 of the previous section. γ<1 plays the important rôle in Watkins' proof of bounding the effect of early $\mathcal{Q}_n$ values. It is fairly easy to modify his proof to the case of an absorbing Markov chain with γ=1, as the ever increasing probability of absorption achieves the same effect. Also, the conditions of Sutton's theorem imply that every nonabsorbing state will be visited infinitely often, and so it suffices to have one set of α_i that satisfy the conditions in (28) and apply them sequentially for each visit to each state in the normal running of the chain.

4. Conclusions

This paper has used Watkins' analysis of the relationship between temporal difference (TD) estimation and dynamic programming to extend Sutton's theorem that TD(0) prediction converges in the mean, to the case of theorem T; TD(λ) for general λ. It also demonstrated that if the vectors representing the states are not linearly independent, then TD(λ) for $\lambda \neq 1$ converges to a different solution from the least mean squares algorithm.

Further, it has applied a special case of Watkins' theorem that $\mathfrak{Q}$-learning, his method of incremental dynamic programming, converges with probability one, to show that TD(0) using a localist state representation, also converges with probability one. This leaves open the question of whether TD(λ), with punctate or distributed representations, also converges in this manner.

Appendix: Existence of appropriate α

Defining

$$\Delta_\lambda = I - \alpha X^{\mathrm{T}} XD(I - (1 - \lambda)Q[I - \lambda Q]^{-1}),$$

it is necessary to show that there is some ϵ such that for $0 < \alpha < \epsilon$, $\lim_{n\to\infty} \Delta_\lambda^n = 0$. In the case that $\lambda = 0$ (for which this formula remains correct), and X has full rank, Sutton proved this on pages 26–28 of (Sutton, 1988), by showing successively that $D(I - Q)$ is positive, that $X^{\mathrm{T}}XD(I - Q)$ has a full set of eigenvalues all of whose real parts are positive, and finally that α can thus be chose such that all eigenvalues of $I - \alpha X^T XD(I - Q)$ are less than 1 in modulus. This proof requires little alteration to the case that $\lambda \neq 0$, and its path will be followed exactly.

The equivalent of $D(I - Q)$ is $D(I - (1 - \lambda)Q[I - \lambda Q]^{-1})$. This will be positive definite, according to a lemma by Varga (1962) and an observation by Sutton, if

$$S = D(I - (1 - \lambda)Q[I - \lambda Q]^{-1}) + \{D(I - (1 - \lambda)Q[I - \lambda Q]^{-1})\}^{\mathrm{T}}$$

can be shown to be strictly diagonally dominant with positive diagonal entries. This is the part of the proof that differs from Sutton, but even here, its structure is rather similar.

Define

$$S_r = D(I - Q^r) + \{D(I - Q^r)\}^{\mathrm{T}}.$$

Then

$$\begin{aligned}[S_r]_{ii} &= [D(I - Q^r)]_{ii} + [\{D(I - Q^r)\}^{\mathrm{T}}]_{ii} \\ &= 2d_i[I - Q^r]_{ii} \\ &= 2d_i(1 - [Q^r]_{ii}) \\ &> 0,\end{aligned}$$

since Q is the matrix of an absorbing Markov chain, and so Q^r has no diagonal elements ≥ 1. Therefore S_r has positive diagonal elements.

Also, for $i \neq j$,

$$\begin{aligned}[S_r]_{ij} &= d_i[I - Q^r]_{ij} + d_j[I - Q^r]_{ji} \\ &= -d_i[Q^r]_{ij} - d_j[Q^r]_{ji} \\ &\leq 0\end{aligned}$$

since all the elements of Q, and hence also those of Q^r, are positive.

In this case, S_r will be strictly diagonally dominant if, and only if, $\Sigma_j[S_r]_{ij} \geq 0$, with strict inequality for some i.

$$\begin{aligned}\sum_j [S_r]_{ij} &= \sum_j d_i[I - Q^r]_{ij} + \sum_j d_j[I - Q^r]_{ji} \\ &= d_i \sum_j [I - Q^r]_{ij} + [\mathbf{d}^T(I - Q^r)]_i \\ &= d_i \left[1 - \sum_j [Q^r]_{ij}\right] + [\mu^T(I - Q)^{-1}(I - Q^r)]_i \qquad (29) \\ &= d_i \left[1 - \sum_j [Q^r]_{ij}\right] + [\mu^T(I + Q + Q^2 + \ldots + Q^{r-1})]_i \qquad (30) \\ &\geq 0 \qquad (31)\end{aligned}$$

where equation (29) follows from equation (16), equation (30) holds since

$$I - Q^r = (I - Q)(I + Q + Q^2 + \ldots + Q^{r-1})$$

and equation (31) holds since $\Sigma_j[Q^r]_{ij} \leq 1$, as the chain is absorbing, and $[Q^s]_{ij} \geq 0$, $\forall s$. Also, there exists at least one i for which $\mu_i > 0$, and the inequality is strict for that i.

Since S_r is strictly diagonally dominant for all $r \geq 1$,

$$\begin{aligned}S_\lambda &= (1 - \lambda) \sum_{r=1}^{\infty} \lambda^{r-1} S_r \\ &= D(I - (1 - \lambda)Q[I - \lambda Q]^{-1}) + \{D(I - (1 - \lambda)Q[I - \lambda Q]^{-1})\}^{\mathrm{T}} \\ &= S\end{aligned}$$

is strictly diagonally dominant too, and therefore $D(I - (1 - \lambda)Q[I - \lambda Q]^{-1})$ is positive definite.

The next stage is to show that $X^{\mathrm{T}}XD(I - (1 - \lambda)Q[I - \lambda Q]^{-1})$ has a full set of eigenvalues all of whose real parts are positive. $X^{\mathrm{T}}X$, D and $(I - (1 - \lambda)Q[I - \lambda Q]^{-1})$ are all nonsingular, which ensures that the set is full. Let ψ and $\mathbf{u}$ be any eigenvalue-eigenvector pair, with $\mathbf{u} = \mathbf{a} + \mathbf{b}i$ and $\mathbf{v} = (X^{\mathrm{T}}X)^{-1}\mathbf{u} \neq \mathbf{0}$, so $\mathbf{u} = (X^{\mathrm{T}}X)\mathbf{v}$. Then

$$\begin{aligned}\mathbf{u}^{*}D(I - (1 - \lambda)Q[I - \lambda Q]^{-1})\mathbf{u} &= \mathbf{v}^{*}X^{\mathrm{T}}XD(I - (1 - \lambda)Q[I - \lambda Q]^{-1})\mathbf{u}\\ &= \mathbf{v}^{*}\psi\mathbf{u}\\ &= \psi\mathbf{v}^{*}(X^{\mathrm{T}}X)\mathbf{v}\\ &= \psi(X\mathbf{v})^{*}X\mathbf{v}\end{aligned}$$

where '*' indicates conjugate transpose. This implies that

$$\mathrm{Re}[\mathbf{u}^{*}D(I - (1 - \lambda)Q[I - \lambda Q]^{-1})\mathbf{u}] = \mathrm{Re}(\psi(X\mathbf{v})^{*}X\mathbf{v})$$

or equivalently,

$$\begin{aligned}\{(X\mathbf{v})^{*}X\mathbf{v}\}\mathrm{Re}[\psi] &= \mathbf{a}^{\mathrm{T}}D(I - (1 - \lambda)Q[I - \lambda Q]^{-1})\mathbf{a}\\ &\quad + \mathbf{b}^{\mathrm{T}}D(I - (1 - \lambda)Q[I - \lambda Q]^{-1})\mathbf{b}.\end{aligned}$$

Since the right side (by positive definiteness) and $(X\mathbf{v})^{*}X\mathbf{v}$ are both strictly positive, the real part of ψ must be strictly positive too.

Furthermore, $\mathbf{u}$ must also be an eigenvector of

$$I - \alpha X^{\mathrm{T}}XD(I - (1 - \lambda)Q[I - \lambda Q]^{-1})$$

since

$$\begin{aligned}[I - \alpha X^{\mathrm{T}}XD(I - (1 - \lambda)Q[I - \lambda Q]^{-1}]\mathbf{u} &= \mathbf{u} - \alpha\psi\mathbf{u}\\ &= (1 - \alpha\psi)\mathbf{u}.\end{aligned}$$

Therefore, all the eigenvalues of $I - \alpha X^{\mathrm{T}}XD(I - (1 - \lambda)Q[I - \lambda Q]^{-1})$ are of the form $1 - \alpha\psi$ where $\psi \equiv \nu + \phi i$ has positive ν. Take

$$0 < \alpha < \frac{2\nu}{\nu^{2} + \phi^{2}},$$

for all eigenvalues ψ, and then all the eigenvalues $1 - \alpha\psi$ of the iteration matrix are guaranteed to have modulus less than one. By another theorem of Varga (1962)

$$\lim_{n\to\infty} [I - \alpha X^{\mathrm{T}}XD(I - (1 - \lambda)Q[I - \lambda Q]^{-1})]^{n} = 0.$$

Acknowledgments

This paper is based on a chapter of my thesis (Dayan, 1991). I am very grateful to Andy Barto, Steve Finch, Alex Lascarides, Satinder Singh, Chris Watkins, David Willshaw, the large number of people who read drafts of the thesis, and particularly Rich Sutton and two anonymous reviewers for their helpful advice and comments. Support was from SERC. Peter Dayan's current address is CNL, The Salk Institute, P.O. Box 85800, San Diego, CA 92186-5800.

Notes

1. Here and subsequently, a superscript λ is used to indicate a TD(λ)-based estimator.
2. States A and G are absorbing and so are not represented.
3. This was one of Watkins' main motivations, as it allows his system to learn about the effect of actions it believes to be suboptimal.

References

Albus, J.S. (1975). A new approach to manipulator control: The Cerebellar Model Articulation Controller (CMAC). *Transactions of the ASME: Journal of Dynamical Systems, Measurement and Control, 97,* 220–227.

Barto, A.G., Sutton, R.S. & Anderson, C.W. (1983). Neuronlike elements that can solve difficult learning problems. *IEEE Transactions on Systems, Man, and Cybernetics, 13,* 834–846.

Barto, A.G., Sutton, R.S. & Watkins, C.J.C.H. (1990). Learning and sequential decision making. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks.* Cambridge, MA: MIT Press, Bradford Books.

Bellman, R.E. & Dreyfus, S.E. (1962). *Applied dynamic programming.* RAND Corporation.

Dayan, P. (1991). *Reinforcing connectionism: Learning the statistical way.* Ph.D. Thesis, University of Edinburgh, Scotland.

Hampson, S.E. (1983). *A neural model of adaptive behavior.* Ph.D. Thesis. University of California, Irvine, CA.

Hampson, S.E. (1990). *Connectionistic problem solving: computational aspects of biological learning.* Boston, MA: Birkhäuser Boston.

Holland, J.H. (1986). Escaping brittleness: The possibilities of general-purpose learning algorithms applied to parallel rule-based systems. In R.S. Michalski, J.G. Carbonell & T.M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach, 2.* Los Altos, CA: Morgan Kaufmann.

Klopf, A.H. (1972). *Brain function and adaptive systems—A heterostatic theory.* Air Force Research Laboratories Research Report, AFCRL-72-0164. Bedford, MA.

Klopf, A.H. (1982). *The hedonistic neuron: A theory of memory, learning, and intelligence.* Washington, DC: Hemisphere.

Michie. D. & Chambers, R.A. (1968). BOXES: An experiment in adaptive control. *Machine Intelligence, 2,* 137–152.

Moore, A.W. (1990). *Efficient memory-based learning for robot control.* Ph.D. Thesis, University of Cambridge Computer Laboratory, Cambridge, England.

Omohundro, S. (1987). Efficient algorithms with neural network behaviour. *Complex Systems, 1,* 273–347.

Samuel, A.L. (1959). Some studies in machine learning using the game of checkers. Reprinted in E.A. Feigenbaum & J. Feldman (Eds.) (1963). *Computers and thought.* McGraw-Hill.

Samuel, A.L. (1967). Some studies in machine learning using the game of checkers II: Recent progress. *IBM Journal of Research and Development, 11,* 601–617.

Sutton, R.S. (1984). *Temporal credit assignment in reinforcement learning.* Ph.D. Thesis, University of Massachusetts, Amherst, MA.

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

Varga, R.S. (1962). *Matrix iterative analysis.* Englewood Cliffs, NJ: Prentice-Hall.

Watkins, C.I.C.H. (1989). *Learning from delayed rewards.* Ph.D. Thesis. University of Cambridge, England.

Werbos, P.J. (1990). Consistency of HDP applied to a simple reinforcement learning problem. *Neural Networks, 3,* 179–189.

Widrow, B. & Stearns, S.D. (1985). *Adaptive signal processing.* Englewood Cliffs, NJ: Prentice-Hall.

Witten, I.H. (1977). An adaptive optimal controller for discrete-time Markov environments. *Information and Control, 34,* 286–295.