TD(λ) Converges with Probability 1

Peter Dayan Terrence J Sejnowski
dayan@helmholtz.sdsc.edu tsejnowski@ucsd.edu

CNL, The Salk Institute
PO Box 85800,
San Diego, CA 92186-5800.

Running head: TD(λ) Converges with Probability 1
Keywords: reinforcement learning, temporal differences, Q-learning

Abstract

The methods of temporal differences (Samuel, 1959; Sutton 1984, 1988) allow agents to learn accurate predictions about stationary stochastic future outcomes. The learning is effectively stochastic approximation based on samples extracted from the process generating the agent's future.

Sutton (1988) proved that for a special case of temporal differences, the expected values of the predictions converge to their correct values, as larger samples are taken, and Dayan (1992) extended his proof to the general case. This paper proves the stronger result that the predictions of a slightly modified form of temporal difference learning converge with probability one, and shows how to quantify the rate of convergence.

1 Introduction

Temporal difference (TD) learning is a way of extracting information from observations of sequential stochastic processes so as to improve predictions of future outcomes. Its key insight is that estimates from successive states should be self-consistent – for instance, the prediction made at one state about a terminal outcome should be related to the prediction from the next state, since this transition is obviously one of the ways of getting to the termination point. The TD algorithm investigated here was invented by Sutton (1988), and uses the difference between such predictions to drive modifications to the parameters that generate them. In fact, Sutton defined a whole class of such TD algorithms, TD(λ), which look at these differences further and further ahead in time, weighted exponentially less according to their distance by the paramter λ.

TD(λ) algorithms have wide application, from modeling classical conditioning in animal learning (Sutton and Barto, 1987) to generating a prize-winning backgammon playing program (Tesauro, 1992) – however the theory underlying them is not

so well developed. Sutton (1988) proved the first theorem, which demonstrated under certain conditions that the mean TD estimates of the terminal reward or return from a particular form of absorbing Markov process converge to the appropriate values. Using some insightful analysis by Watkins (1989) about how TD methods for control relate to dynamic programming (Ross, 1983), Dayan (1992) generalised Sutton's theorem to cover the case where the differences between the predictions from many successive states are taken into account.

Unfortunately, mere convergence of the mean is a very weak criterion. For instance, consider a sequence $\mathcal{H}_n$ of independent, identically distributed, random variables with a finite expectation H. Trivially, $\lim_{n\to\infty} \mathcal{E}[\mathcal{H}_n] = H$, but in no useful sense do the $\mathcal{H}_n$ converge. Convergence *of* the mean is not the same as convergence *in* mean, which would require that $\lim_{n\to\infty} \mathcal{E}[|\mathcal{H}_n - H|] = 0$. The latter does not hold for the simple example. Convergence with probability one is one of the more desirable forms of stochastic convergence, and this note concentrates on proving that it holds for TD.

That the simplest form of TD converges with probability one in a very special case was shown in Dayan (1992). This paper pointed out the equivalence of TD and Watkins' (1989) $\mathcal{Q}$-learning stochastic control learning algorithm, in the case where at no stage is there any choice of possible action. Watkins (1989) and Watkins *et al* (1992) proved convergence of $\mathcal{Q}$-learning with probability one, and this assurance therefore extends to TD.

This paper applies some general theory by Kushner and Clark (1978), which was originally directed at the Robbins-Monro procedure (Robbins & Monro, 1951) to prove that TD(λ) with a linear representation converges with probability one. Kuan and White (1990; 1991) have applied the same theory to the stochastic convergence of static and dynamic backpropagation, provide a more easily approachable introduction to the methods, and also more directly applicable convergence conditions.

2 Definition of TD(λ)

We will treat the same Markov estimation problem that Sutton (1988) and, following him, Dayan (1992), used, and indeed will generally adopt their notation. Consider the case of an absorbing Markov chain with stochastic terminal returns, defined by sets and values:

$\mathcal{T}$		terminal states
$\mathcal{N}$		non-terminal states
$q_{ij} \in [0,1]$	$i \in \mathcal{N}, j \in \mathcal{N}$	transition probabilities between non-terminal states
$s_{ij} \in [0,1]$	$i \in \mathcal{N}, j \in \mathcal{T}$	transition probabilities to terminal states
$x_i \in \Re^c$	$i \in \mathcal{N}$	vectors representing non-terminal states
v_j	$j \in \mathcal{T}$	expected terminal return 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 estimation system is fed complete sequences $x_{i_1}, x_{i_2}, \ldots x_{i_m}$ of observation vectors, together with their scalar terminal return v. It has to generate for every non-terminal state $i \in \mathcal{N}$ a prediction of the expected value $\mathcal{E}[v|i]$ for starting from that state. If the transition matrix of the Markov chain were completely known, these predictions could be computed as:

$$\bar{e}^*_i \equiv \mathcal{E}[v|i] = \sum_{j\in\mathcal{T}} s_{ij}v_j + \sum_{j\in\mathcal{N}} q_{ij} \sum_{k\in\mathcal{T}} s_{jk}v_k + \sum_{j\in\mathcal{N}} q_{ij} \sum_{k\in\mathcal{N}} q_{jk} \sum_{l\in\mathcal{T}} s_{kl}v_l + \ldots . \qquad (1)$$

where $\bar{e}^*$ is the vector of correct predictions.

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

$$\mathcal{E}[v|i] = \left[\sum_{k=0}^{\infty} Q^k h\right]_i = \left[(I - Q)^{-1} h\right]_i \qquad (2)$$

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

During the learning phase, linear TD(λ) generates successive vectors $w_1, w_2, \ldots$, changing w after each complete observation sequence. Define $V^\lambda_n(i) = w_n.x_i$ as the prediction of the terminal return starting from state i, at stage n in learning. Then, during one such sequence, $V^\lambda_n(i_t)$ are the intermediate predictions of these terminal returns, and, abusing notation somewhat, define also $V^\lambda_n(i_{m+1}) = v$, the observed terminal return.[1] TD(λ) changes w according to:

$$w_{n+1} = w_n + \sum_{t=1}^{m} \left\{ \alpha_{n+1}[V^\lambda_n(i_{t+1}) - V^\lambda_n(i_t)] \sum_{k=1}^{t} \lambda^{t-k} \nabla_{w_n} V^\lambda_n(i_k) \right\}. \qquad (3)$$

where α_{n+1} is the learning rate for the n^{th} trial.

[1] Sutton used P^n_t for $V^\lambda_n(i_t)$.

Sutton proved the following theorem true for $\lambda = 0$; Dayan (1992) extended it to the case of general $0 < \lambda < 1$:

> **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 v_i, and for any linearly independent set of observation vectors $\{x_i | i \in \mathcal{N}\}$, there exists an $\epsilon > 0$ such that, if $\alpha_n = \alpha$ where $0 < \alpha < \epsilon$ 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 w_n denotes the weight vector after n sequences have been experienced, then
>
> $$\lim_{n\to\infty} \mathcal{E}[(w_n, x_i)] = \mathcal{E}[v|i] = \left[(I - Q)^{-1} h\right]_i, \forall i \in \mathcal{N}. \qquad (4)$$

In proving this for $\lambda < 1$, it turns out that there is a linear update matrix for the mean estimates of terminal return from each state after n sequences have been observed. If w_n are the actual weights after sequence n and $\bar{w}_{n+1}$ the expected weights after the next sequence, then:

$$X^T \bar{w}_{n+1} = X^T w_n - \alpha_{n+1} X^T X D \left[I - (1-\lambda)Q(I - \lambda Q)^{-1}\right](X^T w_n - \bar{e}^*), \qquad (5)$$

where X is the matrix whose columns are the vectors representing the non-terminal states; $[X]_{ab} = [x_a]_b$. Furthermore, the mean estimates converge appropriately because, if the vectors representing the states are independent (*ie* X is full rank), $-X^T X D \left[I - (1-\lambda)Q(I - \lambda Q)^{-1}\right]$ has a full set of eigenvalues all of whose real parts are negative. It turns out that these, together with some other conditions that are mainly guaranteed by the finiteness of the Markov chain, are just what justifies the use of the powerful stochastic convergence proof methods of Kushner and Clark (1978). The next two sections show how.

3 Convergence Proof

On pages 21-24 and 26-27 of their book, Kushner and Clark (1978) consider the following problem (changing the notation to fit with the above). $z_i \in \Re^c$ are random variables, $k : \Re^c \to \Re^c$ is a stochastic function whose mean for every $z \in \Re^c$ is $\bar{k}(z)$, $0 < \alpha_n < 1$ is a sequence of real numbers such that

$$\sum_{i=1}^{\infty} \alpha_n = \infty \text{ but } \sum_{i=1}^{\infty} \alpha_n^2 < \infty \qquad (6)$$

and

$$z_{n+1} = z_n + \alpha_{n+1}\bar{k}(z_n) + \alpha_{n+1}d_n, \quad \text{and} \tag{7}$$

$$d_n = \left[k(z_n) - \bar{k}(z_n)\right] \tag{8}$$

where d_n acts as a zero mean noise process.

Define a discrete 'time' as $t_n = \sum_{i=1}^{n} \alpha_i$ and a piecewise linear interpolation of the sequence $\{z_n\}$ as:

$$z^{\mathcal{L}}(t) = \frac{t_{n+1} - t}{\alpha_n} z_n + \frac{t - t_n}{\alpha_n} z_{n+1} \tag{9}$$

and functions that are left shifts of this as $z^n(t) \equiv z^{\mathcal{L}}(t + t_n)$. Then, as a consequence of their theorem 2.3.1, if the variance of the noise d_n is bounded and $\{z_n\}$ is bounded with probability 1, then, also with probability 1, the sequence of functions $\{z^n(t)\}$ has a convergent subsequence that tends to some function $z(t)$ which satisfies the equation:

$$\frac{dz}{dt} = \bar{k}(z) \tag{10}$$

Kushner and Clark also show that $z_n \to z_0$ as $n \to \infty$ *if* there is a particular constant solution z_0 of this differential equation which is asymptotically stable in the following way: for any bounded solution $z(t)$ and any $\epsilon > 0$ there is a $\delta > 0$ such that $|z(t) - z_0| < \epsilon$ for $t \geq 0$ if $|z(0) - z_0| < \delta$, and $z(t) \to z_0$ as $t \to \infty$.

Consider this in the context of TD(λ). Define $z_n \equiv X^T w_n$ as the vector containing all the predictions at the n^{th} trial. Then in the appropriate form of the update equation 7, the first two terms on the right hand side, which take care of the mean, are just equation 5. Therefore, the associated differential equation equivalent to expression 10 is:

$$\frac{dz}{dt} = \bar{k}(z) = -X^T X D \left[I - (1-\lambda)Q(I - \lambda Q)^{-1}\right](z - \bar{e}^*). \tag{11}$$

As mentioned above, from Sutton (1988) and Dayan (1992), if the vectors representing the states are independent (*ie* X is full rank), the negated growth matrix in equation 11 has a full set of eigenvalues all of whose real parts are negative. Therefore the differential equation is asymptotically stable about $\bar{e}^*$ in the above manner.

So, *if* the variance of the noise were bounded, and $\{z_n\}$ were bounded with probability 1, then, from the above, we would know that $z_n \to \bar{e}^*$ (equivalently $w_n \to w^*$ where $X^T w^* = \bar{e}^*$ as X is full rank) as $n \to \infty$, with probability 1, *ie* TD(λ) would converge to the right answer with probability 1.

Although far weaker conditions would probably suffice (Kushner, 1984), it is adequate to bound k using a projection technique. Choose a large real bound $\mathcal{B}$. If

z_{n+1}, updated according to equation 7, would lie outside the hypercube defined as $\{-\mathcal{B}, \mathcal{B}\}^c$, it is projected orthogonally back along the offending axes so that it lies on the surface. This makes both $\{z_n\}$ and the variance $\mathcal{V}[k(z)]$ bounded and so, since also the Markov chain is absorbing, Kushner and Clark's theorem 2.3.1 holds for TD(λ).

4 Rate of Convergence

Kushner and Clark go on to show how to quantify the rate of convergence of stochastic algorithms such as the Robbins-Monro procedure. In the notation of this paper, they consider setting $\alpha_n = \gamma(n+1)^{-r}$, for some $1 \geq r > 1/2$, define $d_n = z_n - \bar{e}^*$ and $u_n = (n+1)^s d_n$, and deem the rate of convergence the largest $s \in (0, 1)$ 'for which the asymptotic part of $\{u_n\}$ makes sense as a non-degenerate but "stable" process' (Kushner and Clark, 1978, p 233). r determines how fast learning can progress - the larger it is, the quicker.

If we define H as the Jacobian matrix of $\bar{k}()$ at $\bar{e}^*$, *ie*:

$$H = -X^T X D \left[I - (1-\lambda) Q (I - \lambda Q)^{-1} \right], \tag{12}$$

then, theorem 7.3.1 (Kushner and Clark, 1978, p 245), proves the following theorem (ignoring conditions that are clearly true in the present case):

Theorem P If

1. $\exists \bar{e}^*$ such that $z_n \to \bar{e}^*$ as $n \to \infty$, with probability 1, and $\bar{k}(\bar{e}^*) = 0$,
2. Either:
   - **a)** $r = 1, s = \frac{1}{2}$ and $\bar{H} \equiv \gamma H + sI$ has the real part of all its eigenvalues strictly less than 0, or
   - **b)** $r < 1, s = \frac{r}{2}$ and $\bar{H} \equiv \gamma H$ has the real part of all its eigenvalues strictly less than 0
3. There is a matrix M such that $\mathcal{E}\left[\{k(z_n) - \bar{k}(z_n)\}\{k(z_n) - \bar{k}(z_n)\}^T\right] \to M$ as $n \to \infty$.
4. There are real $\delta > 0$ and $\omega < \infty$ such that $\mathcal{E}[\left|k(z_n) - \bar{k}(z_n)\right|^{2+\delta}] \leq \omega$ for all n

then u_n converges in distribution to a normally distributed random variable.

Conditions 1 and 2b are guaranteed by the previous theorem (condition 2a is stronger and may hold in certain particular cases), and conditions 3 and 4 hold since the Markov chain is absorbing. Therefore, TD(λ) converges at least as fast as $r/2$, where r can be chosen to be 1 if condition 2a holds.

For case 2a, the asymptotic variance of the normal distribution of u_n is related to the covariance matrix M, which in turn is related to the variance of the update operator. Decreasing λ should decrease this variance (Watkins, 1989).[2] However, decreasing λ can also increase the bias and this slows optimal convergence. This trade-off between bias and variance (Watkins, 1989; Geman, Bienenstock & Doursat, 1992) is characteristic of such stochastic algorithms.

5 Discussion

The application of powerful stochastic approximation algorithms in this paper is neither the most general possible nor the most elegant. Kushner and Clark's work has been further extended (*eg* Kushner, 1984; Benveniste *et al*, 1990) and convergence with probability 1 could be proved under less restrictive conditions. Satinder Singh (personal communication) also has a more refined proof that generalises the Watkins *et al* (1992) result that $\mathcal{Q}-$learning, with synchronous updates, converges with probability 1.

Nevertheless, we have shown that TD(λ) converges with probability one, under the standard stochastic convergence constraints on the learning rate given in equation 6 and the other stated conditions. The maximal rate of convergence of this algorithm is determined by the eigenvalues of the update matrix. This gives for TD(λ) a similar assurance that other approximation algorithms enjoy.

Although these theorems provide mathematical assurance for the convergence of TD(λ), the actual rate of convergence can often be too slow for real-world problems, especially for state spaces of high dimensionality. Developing good representations of states is of critical importance in achieving good performance with this as well as other classes of reinforcement learning algorithm.

Acknowledgements

We are most grateful to Chung Ming Kuan and two anonymous reviewers for making detailed comments on the paper and improving the proof, and to Rich

[2] For case 2b, M becomes irrelevant.

Sutton and Halbert White for starting us on the course that led to it. Thanks also to Andy Barto, Vijaykumar Gullapalli, Satinder Singh, and Chris Watkins for invaluable discussions. Support was from SERC and the Howard Hughes Medical Institute.

References

[1] Benveniste, A, Métivier, M & Priouret, P (1990). *Adaptive Algorithms and Stochastic Approximation.* Berlin, Germany: Springer-Verlag.

[2] Dayan, P (1992). The convergence of TD(λ) for general λ. *Machine Learning,* **8**, pp 341-362.

[3] Geman, S, Bienenstock, E & Doursat, R (1991). Neural networks and the bias/variance dilemma. *Neural Computation,* **4**, pp 1-58.

[4] Kuan, CM & White, H (1990). *Recursive m-estimation, Non-linear Regression and Neural Network Learning with Dependent Observations.* Discussion paper, UCSD Department of Economics.

[5] Kuan, CM & White, H (1991). *Strong Convergence of Recursive m-estimators for Models with Dynamic Latent Variables.* Discussion paper 91-05, UCSD Department of Economics.

[6] Kushner, HJ (1984). *Approximation and Weak Convergence Methods for Random Processes, with applications to Stochastic Systems Theory.* Cambridge, MA: MIT Press.

[7] Kushner, HJ & Clark, D (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* Berlin, Germany: Springer-Verlag.

[8] Robbins, H & Monro, S (1951). A stochastic approximation method. *Annals of Mathematical Statistics,* **22**, pp 400-407.

[9] Ross, S (1983). *Introduction to Stochastic Dynamic Programming.* New York, Academic Press.

[10] Samuel, AL (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development,* **3**, pp 211-229.

[11] Sutton, RS (1984). *Temporal Credit Assignment in Reinforcement Learning.* PhD Thesis, University of Massachusetts, Amherst, MA.

[12] Sutton, RS (1988). Learning to predict by the methods of temporal difference. *Machine Learning,* **3**, pp 9-44.

[13] Sutton, RS & Barto, AG (1987). A temporal-difference model of classical conditioning. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society.* Seattle, WA.

[14] Tesauro, G (1992). Practical issues in temporal difference learning. *Machine Learning*, **8**, pp 257-278.

[15] Watkins, CJCH (1989). *Learning from Delayed Rewards.* PhD Thesis, University of Cambridge, England.

[16] Watkins, CJCH & Dayan, P (1992). $\mathcal{Q}$-learning. *Machine Learning*, **8**, pp 279-292.