Proceedings of the 2007 IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning (ADPRL 2007)

Convergence of Model-Based Temporal Difference Learning for Control

Hado van Hasselt and Marco A. Wiering
Intelligent Systems Group, Department of Information and Computing Sciences
Utrecht University
Padualaan 14, 3508 TB Utrecht, The Netherlands
Telephone: +31 - 30 - 251 9372
Fax: +31 - 30 - 251 3791
Email: {hado,marco}@cs.uu.nl

***Abstract*— A theoretical analysis of Model-Based Temporal Difference Learning for Control is given, leading to a proof of convergence. This work differs from earlier work on the convergence of Temporal Difference Learning by proving convergence to the optimal value function. This means that not the values of the current policy are found, but instead the policy is updated in such a manner that ultimately the optimal policy is guaranteed to be reached.**

I. Introduction

Reinforcement Learning (RL) uses a notion of value to guide the behavior of an agent. This value usually represents the future discounted reward the agent is expected to receive from a certain situation onward. If these values are known, the agent can maximize its received rewards by selecting the action corresponding to the maximal value. Unfortunately, values are usually not known a priori and must thus be learned.

There are several different RL algorithms that attempt to solve the same problem. Temporal Difference Learning (TD-Learning) [1], Q-Learning [2] and SARSA [3] are the best known examples. Here, TD-Learning can be used to find values of states, given a certain policy of the agent, while Q-Learning and SARSA find values for state-action pairs. Q-Learning and SARSA have been proven to converge to the optimal policies under certain assumptions. TD-Learning only learns the values of the current policy, and not what the optimal policy is, or what its values are. In this paper we do not discuss eligibility traces and therefore whenever we refer to TD, we in fact mean TD(0).

Model-Based TD-Learning uses TD-Learning to update the current approximation of the value function. A model of the dynamics of the environment is assumed to be known. Problems where this is typically the case include games, such as chess and go. Also maze problems and other settings where the dynamics of the environment are known fall into the scope of this approach. Note that it is not required that the environment is deterministic. Stochastic environments, for example in the game of backgammon, can also be solved with Model-Based TD-Learning. In this paper, we show that convergence to the optimal policy and corresponding value function can be guaranteed under similar conditions as needed for the convergence of SARSA.

In the next section we will first give a short summary of RL and present the notation we will use in this article. Then we present the convergence proof for general Model-Based TD-Learning for Control. This proof will require that the policy is greedy in the limit with infinite exploration.

II. Reinforcement Learning

An agent is assumed to learn from interaction with its environment. An underlying Markov Decision Process (MDP) is assumed. An MDP can be viewed as a tuple (S, A, R, T) where:

- S is the set of all states and $s_t \in S$ is the state the agent is in at time t.
- A is the set of all possible actions and $a_t \in A$ is the action the agent performs at time t.
- $R : S \times A \times S \to \mathbb{R}$ is the reward function that maps a state s_t, an action a_t and the resulting state s_{t+1} into a reward $R(s_t, a_t, s_{t+1})$. The reward r_t is known to the agent when reaching the state s_{t+1} and is sampled from a distribution with expected value $R(s_t, a_t, s_{t+1})$.
- $T : S \times A \times S \to [0, 1]$ is the transition function, where $T(s, a, s')$ gives the probability of arriving in state s' when performing action a in state s.

We assume S and A to be discrete finite sets.

A. Values and Q-Functions

An agent can learn by storing values for each state or for each state-action pair. State values represent the discounted cumulative reward that the agent expects to receive in the future when reaching that state. State-action values represent the discounted cumulative reward it expects to receive after performing that specific action in that state. The goal for the agent is to learn an action selection policy $\pi : S \times A \to [0, 1]$ that optimizes the cumulative reward. Here $\pi_t(s, a)$ gives the probability of selecting action a in state s at time t. Formally, starting at time t, we want the agent to optimize the total discounted cumulative reward:

$$r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots = \sum_{i=0}^{\infty} \gamma^i r_{t+i}$$

$0 \leq \gamma < 1$ is a discount factor that determines the relative impact of immediate rewards compared to more distant rewards. It also ensures that the sum of discounted rewards, and therefore the value, is finite. The value of a state s is denoted by $V(s)$. The value of a state-action pair (s, a) is denoted by $Q(s, a)$. Let Q^π and V^π denote the Q-function and state value function corresponding to some policy π. By definition, we get:

$$\begin{aligned} Q^\pi(s,a) &= E\{\sum_{i=0}^{\infty} \gamma^i r_{t+i} | s_t = s, a_t = a, \pi\} \\ V^\pi(s) &= E\{\sum_{i=0}^{\infty} \gamma^i r_{t+i} | s_t = s, \pi\} \end{aligned}$$

We denote the optimal policy by π^* and its corresponding state and state-action functions by V^* and Q^*. There is always at least one optimal policy. We know that the value function corresponding to the optimal policy will have the following property:

$$V^*(s) = \max_a \sum_{s'} T(s,a,s')(R(s,a,s') + \gamma V^*(s'))$$

which is called the Bellman optimality equation for V^* [4], [5]. Similarly, for Q we get:

$$Q^*(s,a) = \sum_{s'} T(s,a,s')(R(s,a,s') + \gamma \max_{a'} Q^*(s',a'))$$

When Q^* is known, the optimal policy can be found simply by selecting the action with the highest value given the current state. The following properties hold for the optimal values:

$$\begin{aligned} \forall s \in S : \pi^* &= \arg\max_\pi V^\pi(s) \\ \max_\pi V^\pi(s) &= V^*(s) = \max_a Q^*(s,a) \end{aligned}$$

Below we give a short introduction in Reinforcement Learning, where the values of states are approximated during learning. T and R are presumed given, though they could also be approximated by past experiences.

B. Learning the Values

The values of states can be updated using Temporal Difference (TD) learning [1]:

$$V_{t+1}(s_t) = V_t(s_t) + \alpha_t \delta_t \tag{1}$$

Where δ_t is the TD-error, defined as $r_t + \gamma V_t(s_{t+1}) - V_t(s_t)$ and $0 \leq \alpha_t \leq 1$ is a learning rate. It should be noted that this equation converges to the values of states given a certain policy and does not necessarily learn the optimal values. It has been proven that when these values are stored in a table, using update (1) will result in convergence of the values to the actual expected returns for the current policy [1], [6].

Convergence to the optimal policy has been proven under certain conditions for variants of this update using Q-values instead of state values, such as Q-Learning [7]–[10] and SARSA [3], [11]. For Actor-Critic Systems also a proof of convergence to the optimal policy exists, where update (1) is used to update their values and a separate actor determines the action [12]. This proof requires updating the critic and actor on two time scales, with specific restrictions on the update of the actor.

When the state space is continuous, parametrized function approximators (FAs) can be used to store the value of observed states and generalize to unseen states. The update is then performed on the parameters of the FA. For instance, a neural network can be used. The parameters then are the weights of the network. Let θ^V denote the parameters of the FA. The update rule corresponding to Temporal Difference (TD) learning is derived from update (1) and then becomes:

$$\theta^V_{i,t+1} = \theta^V_{i,t} + \alpha \delta_t \frac{\partial V_t(s_t)}{\partial \theta^V_{i,t}} \tag{2}$$

Here $\theta^V_{i,t}$ is the i^{th} component of the parameter vector θ^V at time t and $V_t(s)$ is the output of the FA at time t with state s as input. The update rules corresponding to Q-learning and SARSA are similar. These methods have been extensively studied. See for instance the book by Bertsekas and Tsitsiklis (1996).

In the rest of this paper, our analysis will be focused on the discrete update (1). A logical next step for further work will be to extend this analysis to include the use of linear function approximators to store the value function.

Another possibility to approximate V is to use Least-Squares TD-Learning (LSTD) [13]. LSTD has been proven to be equivalent to standard, non-incremental least-squared linear regression for $\lambda = 1$ [14], where λ is the eligibility trace parameter that determines how far rewards are propagated back to the features that determine the value function. A setting of $\lambda = 1$ corresponds to updating V towards the full Monte-Carlo return for each state. Conversely, a setting of $\lambda = 0$ corresponds to the sampled one-step lookahead. For simplicity, in this paper we will only consider values updated with TD(0), indicating that $\lambda = 0$ and the update used is (1).

C. Selecting Actions

In the case of Model-Based TD-Learning, we can use the current approximation of the value function to determine which action to choose. Using the model, for all actions a we can determine $E\{r_t + \gamma V_t(s_{t+1}) | a_t = a\}$, which is equivalent to $\sum_{s'} T(s_t, a, s')(R(s_t, a, s') + \gamma V_t(s'))$. For instance we can then use a ϵ-greedy exploration scheme, selecting the greedy action $\max_a \sum_{s'} T(s_t, a, s')(R(s_t, a, s') + \gamma V_t(s'))$ with probability $(1 - \epsilon)$ and selecting a random action otherwise. Of course, there are other possibilities for action selection and exploration, but for now we limit ourselves to this approach.

In the next section we will show that as long as our policy is greedy in the limit with infinite exploration (GLIE) [15] in terms of $E\{r_t + \gamma V_t(s_{t+1})\}$, the value function V will converge to the optimal value function V^* and therefore, the greedy policy in terms of this function will in fact be the optimal policy. This means that our convergence proof does not strictly require a model for convergence, though one-step model-based lookahead may be the most apparent application.

D. Dynamic Programming and Model-Based TD-Learning

If T and R are completely known, it is possible to determine the V and Q values of any policy through dynamic programming (DP) [4]. This also allows the optimal policy to be found. However, in most real-world settings DP requires extreme amounts of computation. This is because DP computes the value functions for all states, and does not consider that promising policies only bring the agent in a small part of the state space.

Instead of using one-step lookahead to select an action and update the value function using temporal difference learning, it is also possible to backup the state value using the one-step lookahead. This algorithm is called real-time dynamic programming [16]. For deterministic environments this algorithm would be the same as the method using temporal difference learning when we set α to 1. However, for stochastic environments such as the game of backgammon, computing an action based on a one-step lookahead is not the same as backing up a state value using one-step lookahead. The difference is that when selecting an action, the dice have already been rolled, while for backing up a state value we have to compute transitions based on all possible dice rolls. Therefore it is logical that Tesauro used TD-learning in his famous TDGammon program for learning to play backgammon [17].

On the other hand, even when the MDP is known, we could also use for example Q-learning to learn the optimal policy. However, for particular problems where there are many possible actions and different actions in different states lead to the same consecutive states, learning action values is much less efficient than only learning state values.

Therefore we believe that the model-based TD algorithm for learning an optimal policy is a promising algorithm for particular types of problems. We will next show that this algorithm converges to the optimal policy under particular conditions. It should be noted that instead of using a model, it would also be possible to train an actor using Actor-Critic algorithms. The proof below is also applicable in such cases as long as the actor is updated such that it is GLIE with respect to $E\{r_t + \gamma V_t(s_{t+1})\}$.

III. Convergence

In this section we prove that under certain conditions TD-Learning converges with probability one to the optimal value function. Further we show what the conditions for convergence are.

First a small note on rewards. In this section we will first consider an alternative value $\bar{V}$ instead of the normal V. The difference between these values is that $\bar{V}$ is updated with the non-stochastic reward $R(s_t, a_t, s_{t+1})$ instead of the usual r_t, resulting in the following update:

$$\bar{V}_{t+1}(s_t) = (1-\alpha_t)\bar{V}_t(s_t) + \alpha_t(R(s_t, a_t, s_{t+1}) + \gamma \bar{V}_t(s_{t+1})) \quad (3)$$

This is equivalent to assuming that $r_t = r_n$ when $s_t = s_n$, $a_t = a_n$ and $s_{t+1} = s_{n+1}$. We do this to make the proof shorter and easier to read. However, the proof can easily be extended to include stochastic reward functions as we will show further on. Based on the definitions of V and $\bar{V}$, we can already conclude that the optimal values are equal: $V^* = \bar{V}^*$.

A. G Values

For our proof we define a new value function $G : S \times A \times S \rightarrow \mathbb{R}$. We initialize the values of G by means of the initial values of $\bar{V}$ as follows:

$$G_0(s, a, s') \stackrel{def}{=} R(s, a, s') + \gamma \bar{V}_0(s') \quad (4)$$

The G values are never actually stored or used by any algorithm. Rather they serve as theoretical values. We will show a connection between the G values and the $\bar{V}$ values as would be observed when updating the critic of an Actor Critic System using equation (3). Then we use the G values to prove convergence of the $\bar{V}$ values to the optimal values $\bar{V}^*$.

We can view the G values as the target for updates on another alternative value function V^G. We define the initial values as $\forall s : V_0^G(s) = \bar{V}_0(s)$. Given an experience consisting of a state s_t, an action a_t and a consecutive state s_{t+1}, V^G and G are updated by means of the following update rules:

$$V_{t+1}^G(s_t) = (1-\alpha_t)V_t^G(s_t) + \alpha_t G_t(s_t, a_t, s_{t+1}) \quad (5)$$

$$\forall s, a : G_{t+1}(s, a, s_t) = (1-\alpha_t)G_t(s, a, s_t) + \alpha_t\Big(R(s, a, s_t) + \gamma G_t(s_t, a_t, s_{t+1})\Big) \quad (6)$$

The latter update bears some similarity to the SARSA update, with the important difference that each time step a whole range of G values is updated instead of a single Q value as in the case of SARSA.

We will first show that by construction, for each state and time step $V_t^G(s) = \bar{V}_t(s)$. Then we will show that G converges to the optimal values G^* which by definition fulfill the following equality:

$$\begin{aligned} & G^*(s, a, s') \\ = \; & R(s, a, s') + \gamma \bar{V}^*(s') \\ = \; & R(s, a, s') + \\ & \gamma \max_{a'} \sum_{s''} T(s', a', s'')\Big(R(s', a', s'') + \gamma \bar{V}^*(s'')\Big) \\ = \; & R(s, a, s') + \gamma \max_{a'} \sum_{s''} T(s', a', s'') G^*(s', a', s'') \end{aligned}$$

In order to prove that for all states and all time steps we have $V_t^G(s) = \bar{V}_t(s)$, we first prove the following lemma:

Lemma 1: Consider values G that are initialized by definition (4) and updated by update (6). Then for all time steps t, all states s and s' and all actions a the following equation is valid:

$$\forall t, s, a, s' : R(s, a, s') + \gamma \bar{V}_t(s') = G_t(s, a, s') \quad (7)$$

Proof: This proof will be by induction on t. By definition (4) we have:

$$G_0(s,a,s') = R(s,a,s') + \gamma \bar{V}_0(s') \quad ,$$

which is our induction basis. We assume

$$G_t(s,a,s') = R(s,a,s') + \gamma \bar{V}_t(s') \quad ,$$

for some t and all s, a and s'. We show that it follows from updates (3) and (6) that then

$$G_{t+1}(s,a,s') = R(s,a,s') + \gamma \bar{V}_{t+1}(s') \quad ,$$

for all states s, s' and actions a. For all $s' \neq s_t$, from updates (3) and (6) it immediately follows that:

$$G_{t+1}(s,a,s') = G_t(s,a,s') = R(s,a,s') + \gamma \bar{V}_t(s') = R(s,a,s') + \gamma \bar{V}_{t+1}(s')$$

Now we consider all G values that are updated at time t:

$$\begin{aligned}
& G_{t+1}(s,a,s_t) \\
\stackrel{def}{=} \; & (1-\alpha_t) G_t(s,a,s_t) + \\
& \alpha_t \Big(R(s,a,s_t) + \gamma G_t(s_t,a_t,s_{t+1}) \Big) \\
= \; & (1-\alpha_t) \Big(R(s,a,s_t) + \gamma \bar{V}_t(s_t) \Big) + \\
& \alpha_t \Big(R(s,a,s_t) + \gamma G_t(s_t,a_t,s_{t+1}) \Big) \\
= \; & R(s,a,s_t) + \gamma \Big((1-\alpha_t) \bar{V}_t(s_t) + \\
& \alpha_t G_t(s_t,a_t,s_{t+1}) \Big) \\
= \; & R(s,a,s_t) + \gamma \Big((1-\alpha_t) \bar{V}_t(s_t) + \\
& \alpha_t \Big(R(s_t,a_t,s_{t+1}) + \gamma \bar{V}_t(s_{t+1}) \Big) \Big) \\
\stackrel{def}{=} \; & R(s,a,s_t) + \gamma \bar{V}_{t+1}(s_t)
\end{aligned}$$

The second and fourth equalities hold because of the induction hypothesis. This proves that property (7) holds and therefore proves Lemma 1. ■

Now we will use Lemma 1 to prove another lemma, stating that for all time steps and states $V_t^G(s) = \bar{V}_t(s)$.

Lemma 2: Given a sequence $\{\bar{V}_0, s_1, a_1, r_1, \alpha_1, s_2, \ldots, s_t\}$, for each time t and each state s the value of $\bar{V}_t(s)$ as updated by update (3) is equal to the value of $V_t^G(s)$ as updated by update (5).

Proof: We will also prove this by induction on t. Clearly, by definition, $\forall s : \bar{V}_0(s) = V_0^G(s)$. This is our induction basis. Now assume that $\forall s : \bar{V}_t(s) = V_t^G(s)$. Because updates (5) and (3) only update the values of state s_t, it follows that:

$$\forall s \neq s_t : \bar{V}_{t+1}(s) = \bar{V}_t(s) = V_t^G(s) = V_{t+1}^G(s)$$

Now we have to show that $\bar{V}_{t+1}(s_t) = V_{t+1}^G(s_t)$. For clarity, we repeat updates (3) and (5) in our present notation:

$$\bar{V}_{t+1}(s_t) = (1-\alpha_t)\bar{V}_t(s_t) + \alpha_t(R(s_t,a_t,s_{t+1}) + \gamma \bar{V}_t(s_{t+1}))$$

$$V_{t+1}^G(s_t) = (1-\alpha_t)V_t^G(s_t) + \alpha_t(G_t(s_t,a_t,s_{t+1}))$$

Clearly, the required result is reached if we can show that $R(s_t,a_t,s_{t+1}) + \gamma \bar{V}_t(s_{t+1}) = G_t(s_t,a_t,s_{t+1})$ for all time steps. This follows from Lemma 1, proving Lemma 2. ■

To prove convergence of G to G^*, we use a similar construction as Singh et al. used to prove the convergence of SARSA [15]. For clarity we repeat the lemma given as Lemma 1 in that paper. We use $||\cdot||$ to denote a maximum norm.

Lemma 3: Consider a stochastic process $(\alpha_t, \Delta_t, F_t)$, $t \geq 0$, where $\alpha_t, \Delta_t, F_t : X \rightarrow \mathbb{R}$ satisfy the equations

$$\Delta_{t+1}(x) = (1-\alpha_t(x))\Delta_t(x) + \alpha_t(x)F_t(x),$$

where $x \in X$ and $t = 0, 1, 2, \ldots$. Let P_t be a sequence of increasing σ-fields such that α_0 and Δ_0 are P_0-measurable and α_t, Δ_t and F_{t-1} are P_t-measurable, $t = 1, 2, \ldots$. Assume that the following hold:

1) the set X is finite.
2) $\alpha_t(x) \in [0,1]$, $\sum_t \alpha_t(x) = \infty$, $\sum_t (\alpha_t(x))^2 < \infty$ w.p.1.
3) $||E\{F_t(\cdot)|P_t\}|| \leq \kappa ||\Delta_t|| + c_t$, where $\kappa \in [0,1)$ and c_t converges to zero w.p.1.
4) $\mathrm{Var}\{F_t(x)|P_t\} \leq K(1+\kappa||\Delta_t||)^2$, where K is some constant.

Then, Δ_t converges to zero with probability one.

For a proof of this lemma we refer to the work by Singh et al. (2000). Now we will use this lemma to prove convergence of G. We assume the MDPs we discuss are unichain in order to fulfill the infinite exploration condition. An MDP is called unichain when each policy results in an ergodic Markov chain.

Theorem 1: In a finite unichain state-action MDP, consider a policy π that ensures non-zero probabilities for every action in every state. Assume a_t is chosen according to π_t and assume π is updated such that it is greedy in the limit with infinite exploration (GLIE) in terms of G. Further assume that G is updated by update (6). Then G converges to G^* and π converges to the optimal policy π^* under the following assumptions:

1) The values G are stored in a lookup table.
2) The learning rates satisfy $\alpha_t(s,a,s_t) \in [0,1]$, $\sum_t \alpha_t(s,a,s_t) = \infty$, $\sum_t (\alpha_t(s,a,s_t))^2 < \infty$ and $\alpha_t(s,a,s') = 0$ unless $s' = s_t$.
3) $\forall s,a,s' : \mathrm{Var}\{R(s,a,s')\} < \infty$.

Proof: We define $X = S \times A \times S$. Then the iterative process presented in Lemma 3 reduces to:

$$\Delta_{t+1}(s,a,s_t) = (1-\alpha_t(s,a,s_t))\Delta_t(s,a,s_t) + \alpha_t(s,a,s_t)F_t(s,a,s_t)$$

Now we choose:

$$\Delta_t(s,a,s_t) = G_t(s,a,s_t) - G^*(s,a,s_t),$$

and

$$F_t(s,a,s_t) = R(s,a,s_t) + \gamma G_t(s_t,a_t,s_{t+1}) - G^*(s,a,s_t).$$

The first two conditions of Lemma 3 hold by definition of the state and action spaces and the learning rates. The last condition holds because we define the reward function to be bounded. This means that we only have to show that the third condition of Lemma 3 holds to prove convergence of G_t to G^*.

Let P_t denote the past $\{\bar{V}_0,\ s_1,\ \alpha_1,\ a_1,\ r_1, \ldots,\ r_{t-1},\ s_t,\ \alpha_t\}$ up until reaching state s_t. Note that $G_0, \ldots, G_t$ are P_t-measurable, thus so are Δ_t and F_{t-1}. Then F_t satisfies the following property, where we denote $R(s,a,s_t)$ as r to save space:

$$
\begin{aligned}
& ||E\Big\{F_t(s,a,s_t)|P_t\Big\}|| \\
\stackrel{def}{=}\ & ||E\Big\{r+\gamma G_t(s_t,a_t,s_{t+1}) - G^*(s,a,s_t)|P_t\Big\}|| \\
\stackrel{def}{=}\ & ||E\Big\{r+\gamma G_t(s_t,a_t,s_{t+1}) - \\
& \Big(r+\gamma \max_{a'}\sum_{s''} T(s_t,a',s'')G^*(s_t,a',s'')\Big)|P_t\Big\}|| \\
=\ & \gamma||E\Big\{G_t(s_t,a_t,s_{t+1}) - \\
& \max_{a'}\sum_{s''} T(s_t,a',s'')G^*(s_t,a',s'')|P_t\Big\}|| \\
=\ & \gamma||\sum_{a'}\pi_t(s_t,a')\sum_{s''}T(s_t,a',s'')G_t(s_t,a',s'') - \\
& \max_{a'}\sum_{s''} T(s_t,a',s'')G^*(s_t,a',s'')|| \\
\leq\ & \gamma||\max_{a'}\sum_{s''}T(s_t,a',s'')G_t(s_t,a',s'') - \\
& \max_{a'}\sum_{s''} T(s_t,a',s'')G^*(s_t,a',s'')|| + \\
& \gamma||\sum_{a'}\pi_t(s_t,a')\sum_{s''}T(s_t,a',s'')G_t(s_t,a',s'') - \\
& \max_{a'}\sum_{s''} T(s_t,a',s'')G_t(s_t,a',s'')|| \\
\leq\ & \gamma||\Delta_t|| + \\
& \gamma||\sum_{a'}\pi_t(s_t,a')\sum_{s''}T(s_t,a',s'')G_t(s_t,a',s'') - \\
& \max_{a'}\sum_{s''} T(s_t,a',s'')G_t(s_t,a',s'')||
\end{aligned}
$$

The second term corresponds to c_t in the lemma. This term converges to zero, based on the assumption that the policy is GLIE and the conditions on the reward function and the learning rates. This requires that G is bounded, which follows from the definition of G and the assumption that R is bounded. Therefore G converges to G^*. Because the policy is greedy in the limit, this automatically means π converges to π^*. ■

Now we prove the convergence of $\bar{V}$ and π to $\bar{V}^*$ and π^*.

Theorem 2: In a finite unichain state-action MDP, consider a policy π that ensures non-zero probabilities for every action in every state. Assume a_t is chosen according to π_t and assume π is updated such that it is GLIE with regard to G. Further assume that $\bar{V}$ is updated by update (3). Then $\bar{V}$ converges to $\bar{V}^*$ and π converges to the optimal policy π^* under the following assumptions:

1) The values $\bar{V}$ are stored in a lookup table.
2) The learning rates satisfy $\alpha_t(s_t) \in [0,1]$, $\sum_t \alpha_t(s_t) = \infty$, $\sum_t(\alpha_t(s_t))^2 < \infty$ and $\alpha_t(s) = 0$ unless $s = s_t$.
3) $\forall s,a,s' : \text{Var}\{R(s,a,s')\} < \infty$.

Proof: By construction of G, an optimal policy with respect to G must also be an optimal policy with respect to $\bar{V}$. Note the following - perhaps somewhat unexpected - side effect of Lemma 1:

$$\forall t,s',a',s : \bar{V}_t(s) = \frac{1}{\gamma}\Big(G_t(s',a',s) - R(s',a',s)\Big)$$

This property allows us to make the following derivation concerning $\bar{V}$:

$$
\begin{aligned}
& \lim_{t\to\infty}\bar{V}_t(s) \\
=\ & \lim_{t\to\infty}\frac{1}{\gamma}\Big(G_t(s',a',s) - R(s',a',s)\Big) \\
=\ & \frac{1}{\gamma}\Big(G^*(s',a',s) - R(s',a',s)\Big) \\
\stackrel{def}{=}\ & \bar{V}^*(s)
\end{aligned}
$$

This proves that $\bar{V}$ as updated by update (3) converges to the optimal value $\bar{V}^*$. Because we consider a policy that is greedy in the limit, automatically the policy then converges to the optimal policy π^*. ■

B. Including stochastic rewards

When we consider stochastic rewards, it is possible that $r_t \neq r_n$, while $(s_t,a_t,s_{t+1}) = (s_n,a_n,s_{n+1})$. Now we will show that the convergence proof above also holds for the value V as updated by update (1). To show this, we will prove that the difference between V and $\bar{V}$ converges to zero with probability one. Again, we use Lemma 3.

Theorem 3: The difference between $\bar{V}$ as updated by update (3) and V as updated by update (1) converges to zero with probability one if $\text{Var}\{r\} < \infty$.

Proof: To apply Lemma 3, we define $X = S$ and require $\alpha_t(s) = 0$ when $s \neq s_t$. Further, we have $\Delta_t^V(s_t) = V_t(s_t) - \bar{V}_t(s_t)$ and therefore the following definition for $F_t^V(s_t)$:

$$
\begin{aligned}
& F_t^V(s_t) \\
=\ & \Big(r_t+\gamma V_t(s_{t+1})\Big) - \Big(R(s_t,a_t,s_{t+1}) + \gamma\bar{V}_t(s_{t+1})\Big) \\
=\ & \gamma\Big(V_t(s_{t+1}) - \bar{V}_t(s_{t+1})\Big) + \Big(r_t - R(s_t,a_t,s_{t+1})\Big) \\
=\ & \gamma\Delta_t^V(s_{t+1}) + \Big(r_t - R(s_t,a_t,s_{t+1})\Big)
\end{aligned}
$$

Clearly, when $\text{Var}\{r\} < \infty$ and the usual restrictions on α apply, we have conditions 1,2 and 4 of Lemma 3. Since $E\{r_t - R(s_t,a_t,s_{t+1})\} = 0$, it is also easy to show that condition 3 holds in the present case:

$$
\begin{aligned}
& ||E\{F_t^V(\cdot)|P_t\}|| \\
=\ & ||E\Big\{\gamma\Delta_t^V(\cdot) + \Big(r_t - R(s_t,a_t,s_{t+1})\Big)|P_t\Big\}|| \\
=\ & \gamma||\Delta_t^V(\cdot)||
\end{aligned}
$$

Because all conditions specified in Lemma 3 are fulfilled, we have the desired result. ■

This leads us to our main Theorem, proving convergence of Model-Based TD for Control with minimal restrictions on the updates.

Theorem 4: V as updated by update (1) converges to V^* with probability one under the following assumptions:

1) The values V are stored in a lookup table.
2) The learning rates satisfy $\alpha_t(s_t) \in [0,1]$, $\sum_t \alpha_t(s_t) = \infty$, $\sum_t (\alpha_t(s_t))^2 < \infty$ and $\alpha_t(s) = 0$ unless $s = s_t$.
3) Var$\{r\} < \infty$.
4) The policy followed is greedy in the limit with respect to $E\{r_t + \gamma V_t(s_{t+1})\}$ with infinite exploration.

Proof: V converges to $\bar{V}$ by Theorem 3 and $\bar{V}$ converges to $\bar{V}^*$ by Theorem 2. This immediately implies convergence of V to $\bar{V}^*$ with the restrictions as given above. Since $\bar{V}^*$ and V^* are equal, this implies V as updated by update (1) converges to V^*, concluding our proof. ■

IV. A Failure of Direct Convergence of V

In this section we attempt to directly prove convergence of V to V^* by simple application of Lemma 3 with $X = S$. The proof will fail, allowing us to locate the exact problems with this approach. The approach above, using G values, was constructed to avoid exactly those obstacles.

Theorem 5: In a finite unichain state-action MDP, consider a policy π that ensures non-zero probabilities for every action in every state. Assume a_t is chosen according to π_t and assume π is GLIE. Further assume that V is updated by update (1). Then V converges to V^* and π converges to the optimal policy π^* under the following assumptions:

1) The values V are stored in a lookup table.
2) The learning rates satisfy $\alpha_t(s_t) \in [0,1]$, $\sum_t \alpha_t(s_t) = \infty$, $\sum_t (\alpha_t(s_t))^2 < \infty$ and $\alpha_t(s) = 0$ unless $s = s_t$.
3) Var$\{r\} < \infty$.

Attempted Proof of Theorem 5:

We define $X = S$ and require $\alpha_t(s) = 0$ when $s \neq s_t$. Then the iterative process presented in Lemma 3 becomes

$$\Delta_{t+1}(s_t) = (1 - \alpha_t(s_t))\Delta_t(s_t) + \alpha_t(s_t)F_t(s_t),$$

where we choose:

$$\begin{aligned} \Delta_t(s_t) &= V_t(s_t) - V^*(s_t) \\ F_t(s_t) &= r_t + \gamma V_t(s_{t+1}) - V^*(s_t) \end{aligned}$$

Once again, the only condition of real interest is condition 3 of Lemma 3. The goal is to find a way to show that F_t is less or equal to $\kappa||\Delta_t|| + c_t$ for some κ and c_t. We will be able to derive this. However, knowledge about either π^* or V^* will be needed for c_t to converge to zero, as we will see in the following derivation:

$$\begin{aligned} & ||E\{F_t(s_t)|P_t\}|| \\ = \; & ||E\{r_t + \gamma V_t(s_{t+1}) - V^*(s_t)|P_t\}|| \\ = \; & ||\sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s')) - \\ & V^*(s_t)|| \\ = \; & ||\sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s')) - \\ & \sum_a \pi^*(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s'))|| \end{aligned}$$

There are a few options to continue from here. We note that for any policy π and any transition function T by definition we have:

$$||\sum_a \pi(s,a) \sum_{s'} T(s,a,s')(V_t(s') - V^*(s'))|| \leq ||\Delta_t|| \quad (8)$$

To make use of this fact, we must replace the sum over the current policy with a sum over the optimal policy, or vice versa. This is exactly what we will do in the following two options for the continued derivation.

a) Option 1:

$$\begin{aligned} & ||\sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s')) - \\ & \sum_a \pi^*(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s'))|| \\ \leq \; & ||\sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s')) - \\ & \sum_a \pi^*(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s'))|| + \\ & ||\sum_a \pi^*(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s')) - \\ & \sum_a \pi^*(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s'))|| \\ = \; & ||\sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s')) - \\ & \sum_a \pi^*(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s'))|| + \\ & ||\gamma \sum_a \pi^*(s_t,a) \sum_{s'} T(s_t,a,s')(V_t(s') - V^*(s'))|| \\ \leq \; & ||\sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s')) - \\ & \sum_a \pi^*(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s'))|| + \\ & \gamma||\Delta_t|| \end{aligned}$$

This last result is of the required form. We have $\kappa = \gamma$ and:

$$\begin{aligned} c_t = ||\sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s')) - \\ \sum_a \pi^*(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s'))|| \end{aligned}$$

However, for convergence, c_t should go to zero in the limit, which is only possible if we can get π_t to converge to π^*. This is of course not an option, since this would require knowledge of the optimal policy before actually finding it. Therefore, we explore another option:

b) Option 2:

$$\begin{aligned}
& || \sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s')) - \\
& \sum_a \pi^*(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s'))|| \\
\leq\ & || \sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s')) - \\
& \sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s'))|| + \\
& || \sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s')) - \\
& \sum_a \pi^*(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s'))|| \\
=\ & ||\gamma \sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(V_t(s') - V^*(s'))|| + \\
& || \sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s')) - \\
& \sum_a \pi^*(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s'))|| \\
\leq\ & \gamma ||\Delta_t|| + \\
& || \sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s')) - \\
& \sum_a \pi^*(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s'))|| \\
\leq\ & \gamma ||\Delta_t|| + \\
& || \sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s')) - \\
& \max_a \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s'))||
\end{aligned}$$

Again $\kappa = \gamma$. And this time:

$$c_t = || \sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s')) - \max_a \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s'))||$$

This time, the only way to get c_t to converge to zero is by updating the present policy π_t towards the maximum a. Though this looks similar to just requiring the policy to be greedy in the limit, in this case unfortunately it is not the same. When examining the term carefully, we see that the requirement here is to make the policy greedy in terms of V^* and not V_t as would be the case with normal GLIE policies. Of course, since V^* is not known, once again the proof is stuck.

We show one final attempt for convergence by first finding a term that will converge to zero and then trying to establish a connection of the remaining terms to $||\Delta_t||$.

c) Option 3:

$$\begin{aligned}
& || \sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s')) - \\
& \sum_a \pi^*(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s'))|| \\
=\ & || \sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s')) - \\
& \max_a \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s'))|| \\
\leq\ & || \sum_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s')) - \\
& \max_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s'))|| + \\
& || \max_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s')) - \\
& \max_a \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s'))|| \\
=\ & c_t + \\
& || \max_a \pi_t(s_t,a) \sum_{s'} T(s_t,a,s')(r_t + \gamma V_t(s')) - \\
& \max_a \sum_{s'} T(s_t,a,s')(r_t + \gamma V^*(s'))||
\end{aligned}$$

This time we succeed in finding a term that converges to zero and can be used as c_t. However, the second term does not necessarily converge to zero and also cannot be guaranteed to be smaller than $||\Delta_t||$. To show this, consider a deterministic MDP with two possible actions a and b in state s_t that correspond to two different consecutive states x and y such that $T(s_t,a,x) = T(s_t,b,y) = 1$ and $T(s_t,a,y) = T(s_t,b,x) = 0$. Further, assume the following values hold at time t:

$V_t(s_t)$	$V_t(x)$	$V_t(y)$	$V^*(s_t)$	$V^*(x)$	$V^*(y)$
2	1	0	2	2	0

Finally, assume that $\gamma = 0.9$, $E\{r_t|a_t = a\} = 0.2$ and $E\{r_t|a_t = b\} = 0$. Obviously, we have $||\Delta_t|| = |V_t(x) - V^*(x)| = 1$. However, the second term above will reduce to:

$$|\pi_t(s_t,a)(0.2 + \gamma) - 2.0|$$

This term is larger than $||\Delta_t||$ whenever $\pi_t(s_t,a) < 1/(0.2 + 0.9)$, which is of course by no means impossible. Therefore, also in this case convergence cannot be guaranteed.

The analysis above shows why a direct attempt to prove convergence of TD-Learning to the optimal values does not work. Luckily by using G values as defined above, convergence in fact can be guaranteed. However, we showed that to simply apply Lemma 3 on V knowledge of V^* or π^* is required for convergence, thus showing the need of the detour through G values.

V. Conclusion

We have shown how convergence to the optimal value function V^* and the optimal policy π^* can be guaranteed

when using Temporal Difference Learning to update V and using a model-based one-step lookahead action selection procedure. Also we have shown where a direct attempt to prove convergence of V can fail.

The only restriction for convergence is a policy that is greedy in the limit with infinite exploration (GLIE) with respect to the expected value of the reward and discounted value of the next state. Model-Based TD-Learning fulfills this restriction. However, other action selection methods are also possible, as long as the GLIE restriction is met. Application of this restriction to variants of general Reinforcement Learning algorithms is a subject for future research. Most notably, successful application to variants of model-free Actor Critic Systems seems possible. It will also be interesting to extend the proof to parametrized value functions, for instance using linear function approximators.

REFERENCES

[1] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3, pp. 9–44, 1988.

[2] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. dissertation, King's College, Cambridge, England, 1989.

[3] R. S. Sutton, "Generalization in reinforcement learning: Successful examples using sparse coarse coding," in *Advances in Neural Information Processing Systems 8*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds. MIT Press, Cambridge MA, 1996, pp. 1038–1045.

[4] R. E. Bellman, *Dynamic Programming.* Princeton University Press., 1957.

[5] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT press, Cambridge MA, A Bradford Book, 1998.

[6] P. Dayan, "The convergence of TD(λ) for general lambda," *Machine Learning*, vol. 8, pp. 341–362, 1992.

[7] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, pp. 279–292, 1992.

[8] T. Jaakkola, M. I. Jordan, and S. P. Singh, "On the convergence of stochastic iterative dynamic programming algorithms," *Neural Computation*, vol. 6, pp. 1185–1201, 1994.

[9] J. N. Tsitsiklis, "Asynchronous stochastic approximation and Q-learning," *Machine Learning*, vol. 16, pp. 185–202, 1994.

[10] C. Szepesvári and M. Littman, "A unified analysis of value-function-based reinforcement-learning algorithms," *Neural Computation*, vol. 11, no. 8, pp. 2017–2059, 1999.

[11] G. Rummery and M. Niranjan, "On-line Q-learning using connectionist systems," Cambridge University, UK, Tech. Rep. CUED/F-INFENG-TR 166, 1994.

[12] V. R. Konda and J. N. Tsitsiklis, "On actor-critic algorithms," *SIAM Journal on Control and Optimization*, vol. 42, no. 4, pp. 1143–1166, 2003. [Online]. Available: http://epubs.siam.org/sam-bin/dbq/article/38569

[13] S. J. Bradtke and A. G. Barto, "Linear least-squares algorithms for temporal difference learning," *Machine Learning*, vol. 22, pp. 33–57, 1996.

[14] J. A. Boyan, "Technical update least squares temporal difference learning," *Machine Learning*, vol. 49, no. 2-3, pp. 233–246, 2002.

[15] S. Singh, T. Jaakkola, M. Littman, and C. Szepesvári, "Convergence results for single-step on-policy reinforcement-learning algorithms," *Machine Learning*, vol. 38, no. 3, pp. 287–308, 2000.

[16] A. G. Barto, S. J. Bradtke, and S. P. Singh, "Learning to act using real-time dynamic programming," *Artificial Intelligence*, vol. 72, pp. 81–138, 1995.

[17] G. Tesauro, "Practical issues in temporal difference learning," in *Advances in Neural Information Processing Systems 4*, D. S. Lippman, J. E. Moody, and D. S. Touretzky, Eds. San Mateo, CA: Morgan Kaufmann, 1992, pp. 259–266.