An Analysis of non-Markov Automata Games: Implications for Reinforcement Learning

Mark D. Pendrith* and **Michael J. McGarity†**
*School of Computer Science and Engineering
†School of Electrical Engineering
The University of New South Wales
Sydney 2052 Australia
{pendrith,mikem}@cse.unsw.edu.au

Abstract

It has previously been established that for Markov learning automata games, the game equilibria are exactly the optimal strategies (Witten 1977; Wheeler & Narendra 1986). In this paper, we extend the game theoretic view of reinforcement learning to consider the implications for "group rationality" (Wheeler & Narendra 1986) in the more general situation of learning when the the Markov property cannot be assumed. We show that for a general class of non-Markov decision processes, if actual return (Monte Carlo) credit assignment is used with undiscounted returns, we are still guaranteed the optimal observation-based policies will be game equilibria when using the standard "direct" reinforcement learning approaches, but if either discounted rewards or a temporal differences style of credit assignment method is used, this is not the case.

Introduction

Reinforcement learning (RL) is a set of techniques that have been developed to effect unsupervised learning in agents interacting with a initially unknown and possibly changing environment. It is classically formulated in a table lookup form, where the agent can be in one of a finite number states at any time, and has the choice of finite number of actions to take from within each state. For this representation, powerful convergence and optimality results have been proven for a number of algorithms designed with the simplifying assumption that the environment is Markov, e.g. 1-step Q-learning (Watkins 1989). With this assumption, the problem of learning can be cast into the form of finding an optimal policy for a Markov decision process (MDP), and methods like 1-step Q-learning (QL) can be shown to be form of on-line asynchronous dynamic programming.

A Markov decision process consists of a set of states and a set of possible actions for the agent to choose from in each state. After the selection and execution of an action by the agent, a state transition occurs and the agent receives an immediate payoff (or reward). By *Markov*, it is meant that a decision process has state transition probabilities and immediate payoff (or reward) expectations dependent only upon the action taken within each state, and in particular is therefore *independent* of the history prior to arriving in that state.

In practice, however, RL techniques are routinely applied to many problem domains for which the Markov property does not hold. This might be because the environment is non-stationary, or is only partially observable; often the side-effects of state-space representation can lead to the domain appearing as non-Markov to a reinforcement learning agent.

In this paper, we examine various issues arising from applying standard RL algorithms to non-Markov decision processes (NMDPs). In particular, we are interested in the implications of using a "direct" (Singh, Jaakkola, & Jordan 1994) or *observation-based* method of RL for a non-Markov problem, i.e. where the problem is known to be non-Markov but partial or noisy state observations are presented directly to the RL algorithm without any attempt to identify a "true" Markov state.

The approach we take is to revisit the classic formulation of RL as as n-player learning automata game (Witten 1977; Wheeler & Narendra 1986).

Learning Automata Games

Wheeler and Narendra (1986) describe the learning automata game scenario as one of "myopic" local agents, unaware of the surrounding world, not even knowing that other agents exist. Each local agent, in attempting to maximise its own local payoff, simply chooses an action, waits for a response, and then updates its strategy on the basis of information accumulated to date. In this formulation, there is no explicit synchronisation of decision makers.

We can conceptually decompose a classic lookup-table representation RL system into such an automata game, with one automaton (player) for each system state, the policy action for state i becoming the local strategy for the i^{th} learning automaton. Indeed, this game theoretic view dates back to the earliest work in

RL, firstly in the motivation for the BOXES algorithm (Michie & Chambers 1968), and later more explicitly in Witten's analysis of his adaptive optimal controller for discrete-time Markov environments (Witten 1977).

Casting RL into an n-player game, it is convenient at times to translate the familiar MDP terminology into equivalent game theoretic terms. Instead of policy π we might refer to group or global strategy α. Instead of a deterministic policy, we refer to a *pure strategy*, and the term *mixed strategy* replaces stochastic policy. Finally, the optimality properties of standard RL methods like Q-learning for Markov systems corresponds to the notion of "group rationality" as described by Wheeler and Narendra (1986).

Fundamental to a game theoretic analysis is the notion of a *game equilibrium*. A Nash equilibrium is a global strategy that has the property that each component local strategy for a player is the best available play for that player assuming the other players play their local strategies consistent with that global strategy (Fudenberg & Tirole 1991).

In dynamic programming (DP) terms, a Nash equilibrium corresponds to a policy that is stable under policy iteration. It is well known (e.g. Puterman 1994) that for a MDP all suboptimal policies are unstable under policy iteration i.e. one step of the policy iteration process will result in a different policy. Moreover, the new policy will be a better policy; and so the process of policy iteration can be viewed as a hill-climbing process through the policy space of stationary policies, i.e. the result of each step in policy iteration results in a monotonic improvement in policy until an optimal policy is reached.

The special properties of a Markov domain ensure the strategy/policy space to be well-suited to a hill-climbing strategy; there are no "local maxima" or sub-optimal equilibrium points to contend with, and all the global maxima form a single connected "maxima plateau" that can be reached by starting a hill-climbing process from any point in the space.

It is also the case that a "partial" policy iteration, where only a subset of the states that would have policy changes under a full policy iteration step have their policy actions changed, will also monotonically improve the policy, and therefore result in effective hill-climbing. This is the key property that makes MDPs susceptible to RL techniques; it has become the convention to characterise RL in Markov domains as an asynchronous form of dynamic programming (Watkins 1989). If the RL method is a 1-step temporal differences (TD) method, like Watkins' 1-step Q-learning, the method resembles an on-line, asynchronous form of value iteration. If the RL method is an actual return or Monte Carlo based method, like P-Trace (Pendrith & Ryan 1996) the method resembles an on-line, asynchronous form of policy iteration.

So, for a Markov learning automata game, the optimal group strategies correspond to the equilibria for the game (Witten 1977; Wheeler & Narendra 1986). By way of contrast, for NMDPs and their corresponding learning automata games, it is straightforward to demonstrate that suboptimal equilibria are possible, and subsequently that policy iteration methods can fail by getting "stuck" in local maxima. Consider the NMDP in Figure 1.

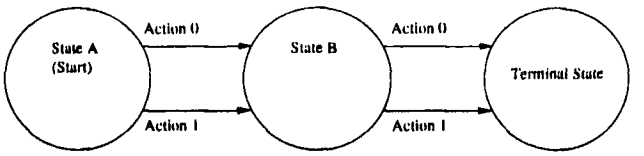

Figure 1: An NMDP with two actions available from starting state A, and two actions available from the successor state B. Both action 0 and action 1 from state A immediately lead to state B with no immediate reward. Action 0 and action 1 from state B both immediately leads to termination and a reward; the decision process is non-Markovian because the reward depends not only on the action selected form state B, but also on what action was previously selected from state A.

Figure 1 shows an NMDP with two actions available from starting state A, and two actions available from the successor state B. Both action 0 and action 1 from state A immediately lead to state B with no immediate reward. Action 0 and action 1 from state B both immediately leads to termination and a reward; the decision process is non-Markovian because the reward depends on what action was previously selected from state A, according to the schedule in Table 1.

	A action	B action	reward
π_0	0	0	1
π_1	0	1	-2
π_2	1	0	0
π_3	1	1	2

Table 1: Reward schedule for NMDP in Figure 1.

In the policy (strategy) space for this NMDP, the policy π_3 is clearly optimal, with a total reward of 2. Further, it is a game equilibrium: if states (players) A or B independently change policy (strategy), the total reward becomes -2 and 0 respectively. Notice that policy π_0 although clearly sub-optimal with a total reward of 1 is also a game equilibrium: if states (players) A or B independently change policy (strategy), the total reward becomes 0 and -2 respectively.

Although we have only explicitly considered deterministic policies (pure strategies) in the above discussion, we note that the result generalises straightforwardly to stochastic policies (mixed strategies).

In the case of the example above the optimal strategy was also a pure strategy. However, it is known that in general for games corresponding to NMDPs there may be be no pure strategy among the optimal group strategies, as will always be the case for MDPs (Singh, Jaakkola, & Jordan 1994).

Further, we show in this paper that if a TD method of credit assignment is used, or the rewards are discounted, the optimal global strategies may not be equilibrium points in the strategy space, even if an optimal pure strategy exists. This means that even if the problems of local maxima are overcome, the optimal policies may not be attractive under some standard RL techniques.

It turns out the key property of optimal policies being stable under RL is only preserved if the additional restrictions of using undiscounted rewards and using actual return credit assignment methods are imposed.

Learning Equilibria

For the analysis of standard RL algorithms for NMDPs, it is useful for us to introduce the notion of a *learning equilibrium*, a type of Nash equilibrium which is relative to a particular learning method.

So just as we can talk about a policy that is stable under policy-iteration, we might talk about a policy that is stable under 1-step Q-learning, for example.

A learning equilibrium has the property that if you replace the current state (or state/action) value estimates with the expected value of the those estimates given the current policy and the the learning method being used, then the policy remains unchanged.

For any MDP with a total discounted reward optimality criterion, the only learning equilibria for any of the RL or DP methods discussed so far will be policy maxima. A policy that is stable under policy-iteration is also stable under value-iteration, or under 1-step Q-learning according to our definition above.

Clearly, having a global maximum in policy space which is also a learning equilibrium is a necessary condition for convergence to an optimal policy under a given learning method.

This basic idea provides the motivation for the form of analysis that follows.

hPOMDPs

The essence of an NMDP is that the history of states and actions leading to the present state may in some way influences the expected outcome of taking an action within that state. When applying a standard RL method like 1-step Q-learning to an NMDP, the history is not used even if available — this is what (Singh, Jaakkola, & Jordan 1994) call *direct* RL for NMDPs. Therefore, one potentially useful approach to modelling a general class of NMDPs is by considering a process that becomes Markov when the full history of states and actions leading to the present state is known, but only *partially observable* if this history is not available or only partially available, i.e. the history provides the missing state information. This property defines a class of partially observable Markov Decision Process (POMDP) we will call hPOMDPs (with h for history).

Before we proceed further, a technical change in terminology used up to this point is called for. Although we have been referring to "states" of a NMDP, hereafter we will generally be referring to the *observations* of an hPOMDP. This brings our terminology into line with the POMDP literature, and thereby avoids a possible source of confusion.

hPOMDPs capture nicely the sort of non-Markovianness that is encountered when state aggregation due to state-space representation or other forms of state-aliasing occur; usually, in cases like these, history can make the observation less ambiguous to some extent, and the more history you have the more precisely you can determine the true state. In control theory, this coincides with the important notion of an *observable system.*

In (Singh, Jaakkola, & Jordan 1994) is discussed a POMDP class similar to hPOMDPs in several important respects. The authors of that paper felt it was difficult in their approach to give a meaningful definition of optimality using a discounted reward framework in the context of POMDPs. The stated difficulty was that it is not guaranteed for a POMDP that there exists an observation-based policy (OBP) that simultaneously maximises the value of each observation; for MDPs, an optimal policy has the property that all state values are maximal.

In the framework we propose, we avoid this problem by adopting an alternative "first principles" definition of optimality for observation-based policies (refer to Equation (3)). Using this definition, the criterion of optimality used in (Singh, Jaakkola, & Jordan 1994) becomes merely a property of optimal policies for MDPs — one that just happens not to generalise to NMDPs.

The other important difference is that Singh et al. limited their formal analysis and results to ergodic systems and gain-optimal average reward RL. The framework proposed here extends to non-ergodic as well as to discounted reward systems, leading to a much more direct understanding of the full implications of applying standard discounted reward RL methods like 1-step Q-learning to the sort of non-Markov environments that are commonly encountered.

A Discounted Reward Framework for NMDPs

Because we are interested in what happens when applying standard discounted reward RL methods like QL to NMDPs, we restrict our attention to the class of *finite* hPOMDPs (i.e., a hPOMDP such that the observation/action space $S \times A$ is finite).[1] This effectively models the RL table-lookup representation for which

[1]Note that this does *not* imply there are only a finite number of states in the underlying MDP. (Singh, Jaakkola, & Jordan 1994) considered a class of POMDPs for which the underlying MDPs had only finite states.

all the strong convergence results have been proven in the context of MDPs.

Summing Over Histories

We consider a *total path* or *trace* through a finite hPOMDP which can be written as a sequence of observation/action pairs

$$(\langle s_0, a_0\rangle, \langle s_1, a_1\rangle, \ldots, \langle s_i, a_i\rangle, \ldots)$$

where $\langle s_i, a_i\rangle$ is the pair associated with the i^{th} time-step of this path through the system. For any finite or infinite horizon total path t there is an associated expected total discounted reward $R(t)$.

We can express the probability P_s^π of a particular observation s being visited under policy π as

$$P_s^\pi = \int_{t \in T_s} p(t, \pi)$$

where the set T_s is the set of possible traces that includes s, and $p(t, \pi)$ as the probability of that trace occurring under policy π. We can also write

$$P_{\bar{s}}^\pi = (1 - P_s^\pi) = \int_{t \in T_{\bar{s}}} p(t, \pi)$$

where $P_{\bar{s}}^\pi$ is the complementary probability of state s not being visited, $T_{\bar{s}}$ being the set of traces that do not include s.

We note that in general, e.g. if the process is non-absorbing, a trace may be of infinite length, and therefore $p(t, \pi)$ may be infinitesimal and T_s uncountable, in which case working directly with the above integrals would require the techniques of measure theory (Billingsley 1986), where $p(t, \pi)$ would actually be a measure on the space of traces. However, we can avoid these complications by observing that executing a trace that involves one or more visits s is logically equivalent to executing a trace that involves a first visit to s, and therefore

$$\int_{t \in T_s} p(t, \pi) = \sum_{h \in H_s} p(h, \pi) \qquad (1)$$

where H_s is the set of finite length *first-visit histories*, which are the possible chains of observation/action pairs leading to a first visit to observation s, and $p(h, \pi)$ is associated probability of a first visit occurring by that history under policy π. Because $h \in H_s$ are of finite length, $p(h, \pi)$ is finite and H_s is countable,[2] and therefore we can express the value as a straightforward sum rather than an integral, simplifying matters considerably. The approach we take in the following is to define the values of these integrals in terms of equivalent sums; whenever an integral of the sort above is used, it can be treated as a place-holder or shorthand for a value that will be defined in terms of these sums.

[2]Consider that the histories could be arranged into classes by length, and could be sorted by some arbitrary lexicographic ordering within each length class, to enable a mapping onto the natural numbers. Note that each length class would have to be finite since we are dealing with finite observations and finite actions within each observation; therefore we can start counting in the zero length class (which contains only the null history $\emptyset$), moving on to classes of length 1,2,3 ... in turn.

Defining Analogs of Q-value and State Value for hPOMDPs

We denote the utility of taking action a from observation s with history h and following π thereafter as

$$U^\pi(s, a, h)$$

and is well-defined by the definition of an hPOMDP; it can be considered the "Q-value" of the underlying (possibly infinite state) MDP where the action a is taken from "true" state $s + h$.

If we were to consider a to be a probabilistic action, which would be the case for stochastic policies, we can generalise the above definition as follows:

$$U^\pi(s, a, h) = \sum_{b \in A} Pr(b|a) U^\pi(s, b, h)$$

where A is the set of available primitive actions in observation s, and $Pr(b|a)$ is the probability of primitive action $b \in A$ being executed under probabilistic action a.

A value that is of interest if we are considering what can be learned applying standard RL methods directly to hPOMDPs is the following weighted average of the above defined utilities

$$Q^\pi(s, a) = \begin{cases} \sum_{h \in H_s} \frac{p(h,\pi)}{P_s^\pi} U^\pi(s, a, h) & \text{if } P_s^\pi > 0 \\ \text{undefined if } P_s^\pi = 0 \end{cases}$$

$Q^\pi(s, a)$ is what might be called the "observation first-visit Q-value"; we observe it is the value a first-visit Monte Carlo method will associate with taking action a from observation s in the hPOMDP. Using this value, we define the value of an observation to be

$$V^\pi(s) = Q^\pi(s, \pi_s)$$

where π_s is the policy action for observation s under policy π.

We note that the values of both $Q^\pi(s, a)$ and $V^\pi(s)$ are undefined for s if $P_s = 0$ (i.e., s is unreachable) under π. This is because, unlike the case for MDPs, it is difficult to assign a sensible meaning to the notion of the value of taking an action from an unreachable observation. For an MDP, even if the state is not reachable under policy π, it is still possible to consider what the expected reward would be artificially starting the MDP from that state; but this idea doesn't work for hPOMDPs, precisely because the path by which it arrived at the observation potentially affects the value

(i.e. the Markov assumption does not hold.) In short, the notion of an "observation first-visit Q-value" is fairly empty if a first visit simply isn't possible.

Policy Values for hPOMDPs

A direct analog of the MDP definition for the value of a policy using a general discounted reward structure is

$$J(\pi) = \sum_{s \in S} \sigma_s U^\pi(s, \pi_s, \emptyset) \tag{2}$$

where σ_s is the probability of starting in observation s, and $\emptyset$ is the trivial history of no observations or actions preceding observation s. From the definition of U above, $J(\pi)$ is well-defined; we define an optimal observation-based policy π^* simply by

$$J(\pi^*) = \max_\pi J(\pi) \tag{3}$$

An interesting alternative way to express the value of a policy π for an hPOMDP is

$$J(\pi) = \int_{t \in T} R(t)\, p(t, \pi) \tag{4}$$

using the above definitions and the idea of integrating over total paths. We could further decompose the total expectation into a component that involves observation s and another that is independent of change to the policy for observation s in the following expression:

$$J(\pi) = \int_{t \in T_s} R(t)\, p(t, \pi) + \int_{t \in T_{\bar{s}}} R(t)\, p(t, \pi) \tag{5}$$

Note that for a general discounted reward structure we can write

$$\int_{t \in T_s} R(t)\, p(t, \pi) \stackrel{def}{=} \sum_{h \in H_s} p(h, \pi)[R(h) + \gamma^{l_h} U^\pi(s, \pi_s, h)] \tag{6}$$

where $0 \leq \gamma \leq 1$ is the discount factor, l_h is the length of history h, and $R(h)$ is the expectation of truncated return associated with history h, and therefore the first term of the RHS of (5) is well-defined. Since the value of $J(\pi)$ in equation (2) is also well-defined, we can write

$$\int_{t \in T_{\bar{s}}} R(t)\, p(t, \pi) \stackrel{def}{=} \sum_{x \in S} \sigma_x U^\pi(x, \pi_x, \emptyset) - \sum_{h \in H_s} p(h, \pi)[R(h) + \gamma^{l_h} U^\pi(s, \pi_s, h)] \tag{7}$$

making both parts of the RHS of equation (5) well-defined.

These definitions provide a framework for analysing hPOMDPs using a total future discounted reward criterion, applying equally well to both ergodic and non-ergodic systems.

Analysis of Observation-Based Policy Learning Methods for hPOMDPs

The first results we present are two lemmas useful in the proof of the Theorem 1, and in discussion of Theorem 2.

Lemma 1 *If two observation-based policies π and $\overset{s}{\pi}$ for a hPOMDP differ only in policy for one observation s, then the difference in values between the policies π and $\overset{s}{\pi}$ can be expressed as*

$$J(\overset{s}{\pi}) - J(\pi) = \sum_{h \in H_s} p(h, \pi)\gamma^{l_h}[U^{\overset{s}{\pi}}(s, \overset{s}{\pi}_s, h) - U^\pi(s, \pi_s, h)] \tag{8}$$

where γ is the discount factor and l_h is the length of history h.

Proof From equation (5) we can write the difference in value between policies $\overset{s}{\pi}$ and π as

$$\begin{aligned} J(\overset{s}{\pi}) - J(\pi) \;=\; & \int_{t \in T_s} R(t)\, p(t, \overset{s}{\pi}) + \int_{t \in T_{\bar{s}}} R(t)\, p(t, \overset{s}{\pi}) - \\ & \int_{t \in T_s} R(t)\, p(t, \pi) - \int_{t \in T_{\bar{s}}} R(t)\, p(t, \pi) \end{aligned}$$

which simplifies to

$$J(\overset{s}{\pi}) - J(\pi) = \int_{t \in T_s} R(t)\, p(t, \overset{s}{\pi}) - \int_{t \in T_s} R(t)\, p(t, \pi)$$

considering that $\int_{t \in T_{\bar{s}}} R(t)\, p(t, \overset{s}{\pi})$ must be equal to $\int_{t \in T_{\bar{s}}} R(t)\, p(t, \pi)$ as the policies are only different in s, and the traces $t \in T_{\bar{s}}$ by definition do not involve s. Similarly we note that $P^{\overset{s}{\pi}}_{\bar{s}} = P^\pi_{\bar{s}}$, and therefore $P^{\overset{s}{\pi}}_s = P^\pi_s$. Using equation (1), we can rewrite the above as

$$\begin{aligned} J(\overset{s}{\pi}) - J(\pi) \;=\; & \sum_{h \in H_s} p(h, \overset{s}{\pi})[R(h) + \gamma^{l_h} U^{\overset{s}{\pi}}(s, \overset{s}{\pi}_s, h)] - \\ & \sum_{h \in H_s} p(h, \pi)[R(h) + \gamma^{l_h} U^\pi(s, \pi_s, h)] \end{aligned}$$

Since $\overset{s}{\pi}$ is only different to π in observation s, the distribution of histories leading to the first visit to s are not affected. Therefore, $p(h, \overset{s}{\pi}) = p(h, \pi)$ for all $h \in H_s$, and we can write

$$\begin{aligned} & J(\overset{s}{\pi}) - J(\pi) = \\ & \quad \sum_{h \in H_s} p(h, \pi)[R(h) + \gamma^{l_h} U^{\overset{s}{\pi}}(s, \overset{s}{\pi}_s, h)] - \\ & \quad \sum_{h \in H_s} p(h, \pi)[R(h) + \gamma^{l_h} U^\pi(s, \pi_s, h)] \\ & = \sum_{h \in H_s} p(h, \pi)\gamma^{l_h}[U^{\overset{s}{\pi}}(s, \overset{s}{\pi}_s, h) - U^\pi(s, \pi_s, h)] \end{aligned}$$

□

Lemma 2 *If two observation-based policies π and $\overset{s}{\pi}$ for an undiscounted hPOMDP differ only in policy for one observation s, then the difference in values between the policies π and $\overset{s}{\pi}$ can be expressed as*

$$J(\overset{s}{\pi}) - J(\pi) = P_s^{\pi}[V^{\overset{s}{\pi}}(s) - V^{\pi}(s)] \qquad (9)$$

Proof Using equation (8) from Lemma 1 (omitting the γ^{l_h} factor since $\gamma = 1$ for an undiscounted hPOMDP), we can derive the difference in policy values as follows (note that stepping from the second to the third line assumes the equivalence of $p(h,\pi)$ and $p(h,\overset{s}{\pi})$, and also of P_s^{π} and $P_s^{\overset{s}{\pi}}$, as discussed in the proof of Lemma 1):

$$\begin{aligned} J(\overset{s}{\pi}) - J(\pi) &= \sum_{h \in H_s} p(h,\pi)[U^{\overset{s}{\pi}}(s, \overset{s}{\pi}_s, h) - U^{\pi}(s, \pi_s, h)] \\ &= P_s^{\pi}\left[\sum_{h \in H_s} \frac{p(h,\pi)}{P_s^{\pi}} U^{\overset{s}{\pi}}(s, \overset{s}{\pi}_s, h) - \right. \\ &\qquad \left. \sum_{h \in H_s} \frac{p(h,\pi)}{P_s^{\pi}} U^{\pi}(s, \pi_s, h)\right] \\ &= P_s^{\pi}\left[\sum_{h \in H_s} \frac{p(h,\overset{s}{\pi})}{P_s^{\overset{s}{\pi}}} U^{\overset{s}{\pi}}(s, \overset{s}{\pi}_s, h) - \right. \\ &\qquad \left. \sum_{h \in H_s} \frac{p(h,\pi)}{P_s^{\pi}} U^{\pi}(s, \pi_s, h)\right] \\ &= P_s^{\pi}[Q^{\overset{s}{\pi}}(s, \overset{s}{\pi}_s) - Q^{\pi}(s, \pi_s)] \\ &= P_s^{\pi}[V^{\overset{s}{\pi}}(s) - V^{\pi}(s)] \end{aligned}$$

□

Lemma 2 has a strong intuitive basis, suggesting its applicability to a very general class of decision processes including but not limited to hPOMDPs. Equation (9) corresponds to the straightforward observation that for an undiscounted reward process, by changing policy in exactly one reachable state under policy π, the change in value of the expected total reward for the new policy is equal to the change in first-visit expected value for the changed state multiplied by the *a priori* probability that state will have a first-visit under policy π.

We emphasize the generality of the result because otherwise it might be misconstrued that the next result we prove (Theorem 1) is somehow tied strongly to the hPOMDP formalisation, when in fact the result is quite general. The proof of Theorem 1 is a simple and generalisable argument which indicates an analogue of Theorem 1 is true for any class of decision process for which Equation (9) holds true.

Theorem 1 *If a first-visit Monte Carlo method of credit assignment is used for a hPOMDP where $\gamma = 1$, then the optimal observation-based policies will be learning equilibria.*

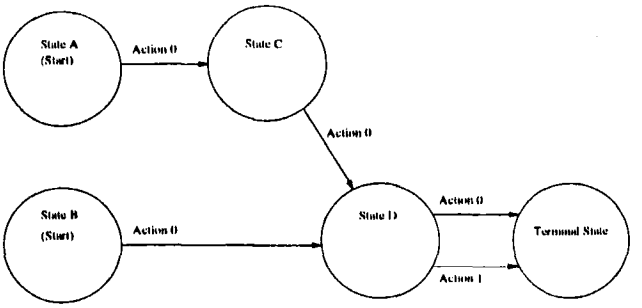

Figure 2: An NMDP with one action available from the two equiprobable starting states A and B; one action available and from intermediate state C; and two actions available from the penultimate state D. An action from state A leads to state C without reward; actions from states B and C lead to state D without reward. Both action 0 and action 1 from state D immediately lead to termination and a reward; the decision process is non-Markovian because the reward depends not only on the action taken from state D, but also on the starting state.

Proof Suppose an optimal observation-based policy π is not a learning equilibrium under a first-visit Monte Carlo credit assignment method; then there must exist a observation s such that $V^{\overset{s}{\pi}}(s) > V^{\pi}(s)$ for some policy $\overset{s}{\pi}$ that is different to π only in observation s. By Lemma 2, the difference in policy values is

$$J(\overset{s}{\pi}) - J(\pi) = P_s^{\pi}[V^{\overset{s}{\pi}}(s) - V^{\pi}(s)]$$

Since $V^{\overset{s}{\pi}}(s) > V^{\pi}(s)$ and $P_s^{\pi} > 0$ (i.e. observation s is reachable under π),[3] then $J(\overset{s}{\pi}) > J(\pi)$. But this is not possible since π is an optimal policy; hence an optimal policy is a learning equilibrium. □

Theorem 1 is a positive result: it shows that, at least under certain restricted conditions, an optimal observation-based policy is also guaranteed to represent a game equilibrium for a direct RL style learner.

The next question is whether we can generalise the result. Does the result hold for general γ? Does the the result hold for TD returns instead of Monte Carlo style "roll-outs"?

The next result addresses the issue of using discounted returns for general γ:

Theorem 2 *Theorem 1 does not generalise to $\gamma \in [0,1)$.*

Proof We prove this by providing a counter-example. We consider the NMDP in Figure 2.

We assume that $\gamma < 1$ for this discounted reward decision process; suppose the reward schedule is as follows:

[3]Note that observation s must be reachable under both π and $\overset{s}{\pi}$ otherwise both $V^{\pi}(s)$ and $V^{\overset{s}{\pi}}(s)$ would be undefined, which is incompatible with the hypothesis $V^{\overset{s}{\pi}}(s) > V^{\pi}(s)$.

Start state	action D	reward
A	0	r_1
A	1	r_2
B	0	r_3
B	1	r_4

Let π_0 and π_1 be the group strategies (policies) that correspond to 0 and 1 being the policy action from D. We set $r_1 \ldots r_4$ such that $Q^\pi(D,0) > Q^\pi(D,1)$ for arbitrary π (i.e. $(r_1 + r_3)/2 > (r_2 + r_4)/2$), but also so that $J(\pi_0) < J(\pi_1)$ (i.e. $(\gamma r_1 + r_3)/2 < (\gamma r_2 + r_4)/2$). For example, let $r_2 = 0$, $r_3 = 1$, $r_4 = 2$, and select r_1 such that $\gamma r_1 < 1 < r_1$.

In such a case, D will see action 0 as preferable, which appears locally optimal even though the choice results in sub-optimal group strategy π_0. Thus the sole optimal group strategy π_1 does not represent a learning equilibrium for this game. □

The basis of the problems for discounted returns with $\gamma < 1$ can be seen in Lemma 1: the γ^{t_h} weights are visible to the total discounted return, but not to the observation first-visit estimator. This is why "group rationality" in this instance breaks down; in the special case of $\gamma = 1$, however, the individual and the group interests suddenly become aligned, as shown in Theorem 1.

Next we examine the case where TD style returns are used; we used 1-step Q-learning in the example that follows:

Theorem 3 *If a 1-step Q-learning method of credit assignment is used for direct RL of a NMDP, then it is not guaranteed there exists an optimal observation-based policy representing a learning equilibrium.*

Proof We prove this by providing an example of an NMDP where the optimal policy is not a learning equilibrium under 1-step Q-learning. In this case we can consider the NMDP in Figure 1 and associated reward schedule in Table 1.

The key to our analysis is to note that a TD-based method like 1-step Q-learning which estimates $Q^\pi(A,0)$ and $Q^\pi(A,1)$ for any policy π will evaluate these actions as of equal utility; therefore, a stochastic action selector will tend to select these actions with equal probability in the limit.

If $\langle A,0\rangle$ and $\langle A,1\rangle$ are being selected with approximately equal probability, then State B will favour action 0 with an expected reward value of $(1+0)/2 = 0.5$ over action 1 with an expected reward value of $(-2+2)/2 = 0$. This implies policies π_1 and π_3 are both unstable in the limit since they both require that $\langle B,1\rangle$ to be the local strategy for state B; but for the reasons given above $\langle B,0\rangle$ will always become inevitably more attractive as state A becomes agnostic about $\langle A,0\rangle$ versus $\langle A,1\rangle$. Even if 1-step Q-learning is initially set with the optimal policy, it will eventually diverge away from it to a situation where it fluctuates between π_0 and π_2 as the learning equilibria.

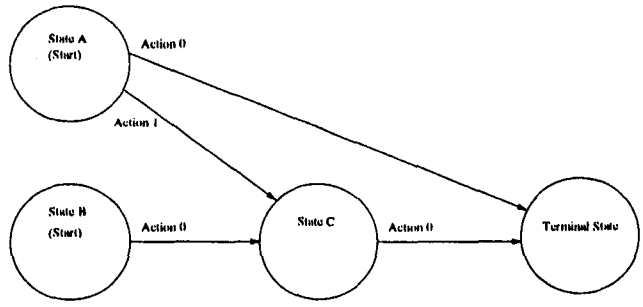

Figure 3: An NMDP with two equiprobable starting states A and B. There are two actions available from state A, but only one action available from B and C. Action 1 from state A leads to state C without reward, as does action 0 from state B. Action 0 from both states A and C immediately leads to termination and a reward; the decision process is non-Markovian because the reward received by C depends not only on the immediate action taken, but also on the starting state.

Finally, we note the above analysis holds true for arbitrary discount factor $\gamma \in [0,1]$. □

In the above analysis, we have represented 1-step Q-learning as a consensus or distributed learner, behaving more like an economy of selfish agents rather than a single learning agent. The easy and natural reasoning suggests the power of the game theoretic analytic framework: for example, we note Theorem 3 also settles an important conjecture in (Singh, Jaakkola, & Jordan 1994) regarding the optimality of QL for observation-based policies of POMDPs. The authors of that paper conjectured that QL in general might not be able to find the optimal deterministic observation-based policy for POMDPs; this result follows directly from the proof of Theorem 3.

Finally, we note that the proof of Theorem 3 can be extended straightforwardly from 1-step to multi-step corrected truncated returns (CTRs). For the case of n-step CTRs, we simply have to add an extra $n-1$ states between state A and state B in Figure 1.

Corollary 1 *Theorem 3 can be generalised to n-step corrected truncated return methods for general n.*

While the proof of Theorem 3 also directly pertains to $TD(\lambda)$ returns (Sutton 1988) for the special case where $\lambda = 0$, to generalise the result for $0 \leq \lambda < 1$ we take a slightly different approach:

Theorem 4 *If a $TD(\lambda)$ credit-assignment method is used for direct RL of a NMDP, then for $\lambda < 1$ it is not guaranteed there exists an optimal observation-based policy representing a learning equilibrium.*

Proof Consider the NMDP is Figure 3. States A and B are the equiprobable starting states. We note all the transitions are deterministic, and that state A has two actions to select from while states B and C have one. Action 0 from state C leads directly to termination with an immediate reward; if the starting state is A,

the immediate reward is 1, but if the starting state is B, the immediate reward will be zero. Action 0 from state A also has a termination and a non-zero immediate reward associated with it, the exact value of which we will discuss in a moment. All other transitions have a zero immediate reward associated with them.

The expected value of $\langle C, 0\rangle$ for an observation based policy π depends upon the relative frequency of the transitions $A \rightarrow C$ and $B \rightarrow C$; this in turn depends upon how often state A selects action 1 for the sake of active exploration. We only assume the relative frequencies of action 0 and action 1 selections from state A are both non-zero; hence $Q^\pi(C, 0) \in (0, 0.5)$.

Assuming $\gamma \in [0, 1]$, from the rules of TD updates we can derive that $Q^\pi(A, 1) = \gamma(\lambda.1 + (1-\lambda)Q^\pi(C, 0))$. This interests us, because $Q^\pi(A, 1)$ would equal γ under a Monte Carlo method of credit assignment, but for $TD(\lambda)$ returns $Q^\pi(A, 1) < \gamma$ for all $\lambda < 1$.

Therefore, if the value of the immediate reward for $\langle A, 0\rangle$ is such that $Q^\pi(A, 1) < Q^\pi(A, 0) < \gamma$, then state A would prefer action 0 over action 1, even though the global optimal strategy corresponds to selecting action 1. In such a case, the global strategy for this NMDP does not represent a learning equilibrium if $TD(\lambda)$ returns are used with $\lambda < 1$. □

Taken together, these results show that the key property of optimal observation-based policies being stable in non-Markov domains for direct RL methods is only preserved if the additional restrictions of using undiscounted rewards and using actual return credit assignment methods are imposed.

From Discounted to Undiscounted to Average Rewards

A move from discounted to undiscounted rewards naturally suggests a closer look at average reward RL methods for equilibrium properties in non-Markov environments. Some steps in this direction have already been made in (Singh, Jaakkola, & Jordan 1994) and (Jaakkola, Singh, & Jordan 1995); the results presented above add weight to arguments that this is indeed the right direction to be heading.

In moving to average reward criteria for NMDPs, an interesting set of open questions remain for future investigation. In particular, Theorem 2 may point to subtle problems translating "transient reward" sensitive metrics such as Blackwell optimality (Mahadevan 1996), from MDPs to NMDPs. Investigations are continuing in this direction.

Conclusions

A game theoretic approach has proven to be an aid to understanding the theoretical implications of applying standard discounted reward RL methods to non-Markov environments. Complementary to earlier work, the framework we present extends to non-ergodic as well as discounted reward NMDPs, facilitating a much more direct understanding of the issues involved.

Our analysis starts with the simple observation that having a global maximum in policy space which is also a learning equilibrium is a necessary condition for convergence to an optimal policy under a given learning method. We discover that for an important general class of non-Markov domains, undiscounted, actual return methods have significant theoretical advantages over discounted returns and TD methods of credit-assignment. This has potentially major implications for RL as it is currently practiced.

References

Billingsley, P. 1986. *Probability and measure.* John Wiley & Sons.

Fudenberg, D., and Tirole, J. 1991. *Game Theory.* MIT Press.

Jaakkola, T.; Singh, S.; and Jordan, M. 1995. Reinforcement learning algorithm for partially observable Markov decision problems. In *Advances in Neural Information Processing Systems 7.* Morgan Kaufmann.

Mahadevan, S. 1996. Sensitive discount optimality: Unifying discounted and average reward reinforcement learning. In L.Saitta., ed., *Machine Learning: Proc. of the Thirteenth Int. Conf.* Morgan Kaufmann.

Michie, D., and Chambers, R. 1968. BOXES: An experiment in adaptive control. In E.Dale, and D.Michie., eds., *Machine Intelligence 2,* 137–152. Edinburgh: Edinburgh Univ. Press.

Pendrith, M., and Ryan, M. 1996. Actual return reinforcement learning versus Temporal Differences: Some theoretical and experimental results. In L.Saitta., ed., *Machine Learning: Proc. of the Thirteenth Int. Conf.* Morgan Kaufmann.

Puterman, M. 1994. *Markov decision processes : Discrete stochastic dynamic programming.* New York: John Wiley & Sons.

Singh, S.; Jaakkola, T.; and Jordan, M. 1994. Learning without state-estimation in partially observable Markovian decision processes. In W.Cohen, and H.Hirsh., eds., *Machine Learning: Proc. of the Eleventh Int. Conf.* New Brunswick, New Jersey: Morgan Kaufmann.

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

Watkins, C. 1989. *Learning from Delayed Rewards.* Ph.D. Thesis, King's College, Cambridge.

Wheeler, Jr., R. M., and Narendra, K. S. 1986. Decentralized learning in finite Markov chains. *IEEE Trans. on Automatic Control* AC-31(6):519–526.

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