Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

On Q-learning Convergence for Non-Markov Decision Processes

Sultan Javed Majeed[1], **Marcus Hutter**[2]
[1,2] Research School of Computer Science, Australian National University, Australia
[1] http://www.sultan.pk, [2] http://www.hutter1.net

Abstract

Temporal-difference (TD) learning is an attractive, computationally efficient framework for model-free reinforcement learning. Q-learning is one of the most widely used TD learning technique that enables an agent to learn the *optimal* action-value function, i.e. Q-value function. Contrary to its widespread use, Q-learning has only been proven to converge on Markov Decision Processes (MDPs) and Q-uniform abstractions of finite-state MDPs. On the other hand, most real-world problems are inherently non-Markovian: the full true state of the environment is not revealed by recent observations. In this paper, we investigate the behavior of Q-learning when applied to non-MDP and non-ergodic domains which may have infinitely many underlying states. We prove that the convergence guarantee of Q-learning can be extended to a class of such non-MDP problems, in particular, to some non-stationary domains. We show that state-uniformity of the optimal Q-value function is a necessary and sufficient condition for Q-learning to converge even in the case of infinitely many internal states.

1 Introduction

Temporal-difference learning [Sutton, 1988] is a well-celebrated model-free learning framework in machine learning. In TD, an agent learns the optimal action-value function (also known as the Q-value function) of the underlying problem without explicitly building or learning a model of the environment. The agent can learn the optimal behavior from the learned Q-value function: the optimal action maximizes the Q-value function. It is generally assumed that the environment is Markovian and ergodic for a TD agent to converge [Tsitsiklis, 1994; Bertsekas and Tsitsiklis, 1995].

The TD agents, apart from a few restrictive cases[1], are not proven to learn[2] non-Markovian environments, whereas most real-world problems are inherently non-Markovian: the full true state of the environment is not revealed by the last observation, and the set of true states can be infinite, e.g. as effectively in non-stationary domains. Therefore, it is important to know if the agent performs well in such non-Markovian domains to work with a broad range of real-world problems.

In this work, we investigate convergence of one of the most widely used TD learning algorithms, Q-learning [Watkins and Dayan, 1992]. Q-learning has been shown to converge in MDP domains [Tsitsiklis, 1994; Bertsekas and Tsitsiklis, 1995], whereas there are empirical observations that Q-learning sometimes also work in some non-MDP domains [Sutton and Barto, 1984]. First non-MDP convergence of Q-learning has been reported by Li *et al.* [2006] for the environments that are Q-uniform abstractions of finite-state MDPs. The recent results in Extreme State Aggregation (ESA) [Hutter, 2016] indicate that under some conditions there exists a deterministic, near-optimal policy for non-MDP environments which are not required to be abstractions of any finite-state MDP. These positive results motivated this work to extend the non-MDP convergence proof of Q-learning to a larger class of infinite internal state non-MDPs.

The most popular extension of MDP is a finite-state partially observable Markov decision process (POMDP). In a POMDP the environment has a *hidden* true state, and the observations from the environment, generally, do not reveal the true state. Therefore, the agent either has to keep a full interaction history, estimate the true state or maintain a belief over the possible true states. In our formulation, we use an even more general class of processes, history-based decision process (HDP): the history-based process is equivalent to an *infinite-state* POMDP [Leike, 2016]. We provide a simple proof of Q-learning convergence to a class of domains that encompasses significantly more domains than MDP and intersects with POMDP and HDP classes. We name this class Q-value uniform Decision Process (QDP) and show that Q-learning converges in QDPs. Moreover, we show that QDP is the largest class where Q-learning can converge, i.e. QDP provides the necessary and sufficient conditions for Q-leaning convergence.

Apart from a few toy problems, it is always a leap of faith to treat real-world problems as MDPs. An MDP *model* of the underlying true environment is implicitly *assumed* even for model-free algorithms. Our result helps to relax this assump-

[1] See Section 5 for exceptions.

[2] In this work we use the term "learn a domain" in the context of learning to act optimally and not to learn a model/dynamics of the domain.

tion: rather assuming the domain being a *finite-state* MDP, we can suppose it to be a QDP, which is a much weaker implicit assumption. The positive result of this paper can be interpreted in a couple of ways; **a)** as discussed above, it provides theoretical grounds for Q-learning to be applicable in a much broader class of environments or **b)** if the agent has access to a QDP aggregation map as a potential model of the true environment or the agent has a companion map learning/estimation algorithm to build such a model, then this combination of the aggregation map with Q-learning converges. It is an interesting topic to learn such maps, but is beyond the scope this work.

The rest of the paper is structured as follows. In Section 2 we set up the framework. Section 3 drafts the QDP class. Section 4 gives a preview of our main convergence result. Section 5 provides a context of our work in the literature. Section 6 contains the proof of the main results. In Section 7 we numerically evaluate Q-learning on a few non-MDP toy domains. Section 8 concludes the paper.

2 Setup

We use the general history-based agent-environment reinforcement learning framework [Hutter, 2005; Hutter, 2016]. The agent and the environment interact in cycles. At the beginning of a cycle $t \in \mathbb{N}$ the agent takes an action a_t from a finite action-space $\mathcal{A}$. The environment dispenses a real-valued reward r_{t+1} from a set $\mathcal{R} \subset \mathbb{R}$ and an observation o_{t+1} from a finite set of observations $\mathcal{O}$. However, in our setup, we assume that the agent does not directly use this observation, e.g. because $\mathcal{O}$ maybe too huge to learn from. The agent has access to a map/model ϕ of the environment that takes in the observation, reward and previous interaction history and provides the same reward r_{t+1} and a mapped state s_{t+1} from a finite set of states $\mathcal{S}$; and the cycle repeats. Formally, this agent-environment interaction generates a growing history $h_{t+1} := h_t a_t o_{t+1} r_{t+1}$ from a set of histories $\mathcal{H}^t := (\mathcal{A} \times \mathcal{O} \times \mathcal{R})^t$. The set of all finite histories is denoted by $\mathcal{H}^* := \bigcup_t \mathcal{H}^t$. The map ϕ is assumed to be a surjective mapping from $\mathcal{H}^*$ to $\mathcal{S}$. We use $h_t := \epsilon$ to denote the empty history and $:=$ to express an equality by definition. In general (non-MDPs), at any time-instant t the transition probability to the next observation $o' := o_{t+1}$ and reward $r' := r_{t+1}$ is a function of the history-action $(h, a) := (h_t, a_t)$-pair and not of the state-action $(s, a) := (s_t, a_t)$-pair. The true environment as is a history-based (decision) process.

Definition 1 (History-based Decision Process (HDP)) *A history-based decision process P is a stochastic mapping from a history-action pair to observation-reward pairs. Formally, $P : \mathcal{H}^* \times \mathcal{A} \rightsquigarrow \mathcal{O} \times \mathcal{R}$, where $\rightsquigarrow$ denotes a stochastic mapping.*

We use Q to denote an action-value function of a HDP, and Q^* denotes the optimal Q-value function.

$$Q^*(h,a) := \sum_{o'r'} P(o'r'|ha) \left(r' + \gamma \max_{\tilde{a}} Q^*(hao'r', \tilde{a}) \right) \tag{1}$$

where $\gamma \in [0, 1)$ is a discount factor. In our setup, the agent does not maintain a complete history and effectively experiences the agent-environment interaction as an action-state-reward sequence $(a_t, s_t, r_t)_{t\in\mathbb{N}}$. We call it a *state-process* [3] induced by the map ϕ.

Definition 2 (State-process) *For a history h that is mapped to a state s, a state-process p_h is a stochastic mapping from a state-action pair with the fixed state s to state-reward pairs. Formally, $p_h : \{s\} \times \mathcal{A} \rightsquigarrow \mathcal{S} \times \mathcal{R}$.*

The relationship between the underlying HDP and the induced state-process for an $s = \phi(h)$ is formally defined as:

$$p_h(s'r'|sa) := \sum_{o':\phi(hao'r')=s'} P(o'r'|ha). \tag{2}$$

We denote the action-value function of the state-process by q, and the optimal Q-value function is given by q^*:

$$q^*(s,a;h) := \sum_{s'r'} p_h(s'r'|sa) \left(r' + \gamma \max_{\tilde{a}} q^*(s', \tilde{a}; h) \right) \tag{3}$$

It is clear that $p_h(s'r'|sa)$ may not be same as $p_{\dot{h}}(s'r'|sa)$ for any two histories h and $\dot{h}$ mapped to a same state s. If the state-process is an MDP, then p_h is independent of history and so is q^*, and convergence of Q-learning follows from this MDP condition [Bertsekas and Tsitsiklis, 1995]. However, we do not assume such a condition and go beyond MDP mappings. We later show — by constructing examples — that q^* can be made independent of history while the state-process is still history dependent, i.e. non-MDP.

Now we formally define Q-learning: At each time-step t the agent maintains an action-value function estimate q_t. The agent in a state $s := s_t$ takes an action $a := a_t$ and receives a reward $r' := r_{t+1}$ and the next state $s' := s_{t+1}$. Then the agent performs an action-value update to the (s, a)-estimate with the following Q-learning update rule:

$$q_{t+1}(s,a) := \\ q_t(s,a) + \alpha_t(s,a) \left(r' + \gamma \max_{\tilde{a}} q_t(s', \tilde{a}) - q_t(s,a) \right) \tag{4}$$

where $(\alpha_t)_{t\in\mathbb{N}}$ is a learning rate sequence.

3 Q-Value Uniform Decision Process (QDP)

In this section we formulate a class of environments called Q-value uniform decision processes, i.e. QDP class. This class is substantially larger than MDP and has a non-empty intersection with POMDP and HDP (Figure 1). In a state-aggregation context, a model is a QDP if it satisfies the following state-uniformity condition.

Definition 3 (State-uniformity condition) *For any action a, if any two histories h and $\dot{h}$ map to the same state s, then the optimal Q-values of the underlying HDP of these histories are the same, i.e. state-uniform; $Q^*(h, a) = Q^*(\dot{h}, a)$. It is easy to see that in this case $q^*(s, a) = Q^*(h, a)$ [Hutter, 2016].*

[3] It is technically a state and reward process, but since rewards are not affected by the mapping ϕ, we suppress the reward part to put more emphasis on the contrast between history and state dependence.

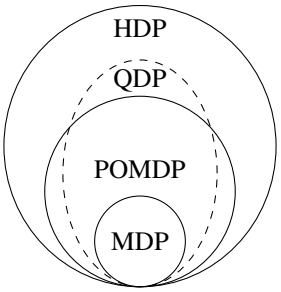

Figure 1: QDP in the perspective of other decision problem classes.

The state-uniformity condition is weaker than the MDP condition: the latter implies the former but not the other way around [Hutter, 2016]. Therefore, trivially MDP $\subseteq$ QDP $\subseteq$ HDP. The example in Figure 2 shows that QDP $\cap$ POMDP$\backslash$MDP $\neq \emptyset$. The example in Figure 5 proves that QDP $\backslash$POMDP $\neq \emptyset$. Moreover, it is easy to find examples in POMDP$\backslash$QDP and HDP$\backslash$QDP.

It is easy to see that QDP is a much larger class than MDP since the former allows non-stationary domains: it is possible to have $p_{h_t}(\cdot|sa) \neq p_{h_\tau}(\cdot|sa)$ for some histories $\phi(h_t) = \phi(h_\tau) = s$ at two different time-steps but still maintaining $Q^*(h_t, a) = Q^*(h_\tau, a)$ (Figure 5). Moreover, the definition of QDP enables us to approximate most if not all problems as a QDP model: any number of *similar* Q^*-value histories can be merged into a single QDP state [Hutter, 2016]. In particular, a QDP model of the environment can provide more compression in terms of state space size as compared to an MDP model: multiple MDP states with the same/similar Q-value can be merged into a single QDP state but not necessarily the other way around. Thus, QDP allows for more compact models for an environment than its MDP counterparts.

In general, we can say that a POMDP has both the dynamics and Q-values as functions of history. Whereas, the definition of QDP provides models where only the dynamics can be history-dependent. Therefore, QDP captures a subset of POMDP models that have history-independent Q-values.

4 Main Result

We assume that the state-process is ergodic — i.e. all states are reachable under any policy from the current state after sufficiently many steps.

Assumption 4 (Ergodicity) *The state-process is ergodic.*

Because of the ergodicity assumption we can suppose the following standard stochastic approximation conditions on each state-action (s, a)-pair's learning rate sequence $(\alpha_t)_{t\in\mathbb{N}}$[4].

$$\sum_{t=0}^{\infty} \alpha_t(s, a) = \infty, \qquad \sum_{t=0}^{\infty} \alpha_t^2(s, a) < \infty. \tag{5}$$

The above conditions on the learning rate ensure that the agent asymptotically decreases the learning rate to converge to a fixed point but never stops learning to avoid local maxima [Bertsekas and Tsitsiklis, 1995]. It is critical to note that we assume ergodicity of the state-process but not of the underlying HDP: a state can be reached multiple times from different histories but any history is only reached once. We assume that the state-process is a QDP.

Assumption 5 (QDP) *The state-process is a QDP.*

It is important to consider that we only assume the optimal action-value to be a function of states. We do not suppose any structure on the intermediate action-value estimates *cf.* $q_t(s, a) \neq Q_t(h, a)$. We also assume that rewards are bounded. This is a standard condition for Q-learning convergence.

Assumption 6 (Bounded rewards) *The rewards are bounded, i.e.* $r \in [r_{\min}, r_{\max}]$.

We have all the components in place to extend Q-learning convergence in QDP.

Theorem 7 (Q-learning Convergence in QDP) *Under Assumptions 4, 5 and 6, and with a learning rate sequence* $(\alpha_t)_{t\in\mathbb{N}}$ *satisfying* (5)*, the sequence* $(q_t)_{t\in\mathbb{N}}$ *generated by the iteration* (4) *converges to* $q^* = Q^*$.

Hence, the agent learns the optimal action-value function of a QDP state-process.

5 Related Work

A similar result was first reported by Li *et al.* [2006] in a finite state MDP setting. We confirm and extend the findings by considering a more general class of environments, i.e. HDP. Also, we do not assume a weighting function to define the state-process *cf.* [Li *et al.*, 2006, Definition 1], and our proof is based on time-dependent contraction mappings (see Section 6 for the details).

A finite-state POMDP is the most commonly used extension of an MDP. It is well-known that the class of finite-state POMDPs is a subset of HDP class [Leike, 2016]. One prevalent approach to handle the non-Markovian nature of a POMDP is to estimate a Markovian model with a state estimation method [Whitehead and Lin, 1995; Lin and Mitchell, 1992; Cassandra *et al.*, 1994; Cassandra, 1994] or use a finite subset of the recent history as a state [McCallum, 1995] to form a k-order MDP (k-MDP). Then Q-learning is applied to learn this resultant state-based MDP. This is a different approach to ours. We do not try to estimate an MDP or k-MDP representation of the underlying HDP.

Singh et. al. [Singh *et al.*, 1994] investigate a direct application of model-free algorithms to POMDPs without a state estimation step akin to our setup but limited to finite state POMDPs only. They show that an optimal policy in a POMDP may be non-stationary and stochastic. But the learned policy in direct model-free algorithms, such as Q-learning, is generally stationary and deterministic by design. Moreover, they also show that the optimal policy of POMDP can be arbitrarily better than the optimal stationary, deterministic policy of the corresponding MDP. These negative results of [Singh *et al.*, 1994] are based on counter-examples that violate the state-uniformity assumption of the optimal

[4]Note that $\alpha_t(s, a) := 0, \forall (s, a) \neq (s_t, a_t)$.

Q-value function. Similar negative findings are reported by Littman [1994]. Our positive convergence result holds for a subset of POMDPs that respects the state-uniformity condition.

Pendrith and McGarity [1998] show that a direct application of standard reinforcement learning methods, such as Q-learning, to non-Markovian domains can have a stationary optimal policy, if undiscounted return is used as a performance measure. Perkins and Pendrith [2002] prove existence of a fixed point in POMDPs for continuous behavior policies. However, this fixed point could be significantly worse than the average reward achieved by an oscillating discontinuous behavior policy. It signifies the effect of behavior policy on the learning outcome. On the other hand, our convergence result is valid as long as all the state-action pairs are visited infinitely often. The nature of the behavior policy does not directly affect our convergence result. A comprehensive survey of solution methods for POMDPs is provided by Murphy [2000] and more recently by Thrun *et al.* [2005].

6 Convergence Proof

We provide a proof of Theorem 7 in this section. Superficially, the proof looks similar to a standard MDP proof [Bertsekas and Tsitsiklis, 1995], however, a subtlety is involved in the definition of the contraction map: the contraction map is a function of history. This history-dependence lets the proof scale to non-MDP domains, especially to non-stationary domains.

Proof of Theorem 7. At a time-instant t with a history h_t we rewrite (4) in terms of an operator and a noise term.

$$q_{t+1}(s,a) = \\ (1-\alpha_t(s,a))\, q_t(s,a) + \alpha_t(s,a)\,(F_{h_t} q_t(s,a) + w_{h_t}(s,a)) \quad (6)$$

where, the noise term is defined as follows,

$$w_{h_t}(s,a) := r' + \gamma \max_{\tilde{a}} q_t(s',\tilde{a}) - F_{h_t} q_t(s,a). \quad (7)$$

Since the agent samples from the underlying HDP, the operator F_{h_t} is a history-based operator.

$$F_{h_t} q_t(s,a) := \mathbb{E}_{p_{h_t}}\left[r' + \gamma \max_{\tilde{a}} q_t(s',\tilde{a}) | \mathcal{F}_t\right] \quad (8)$$

where $\mathcal{F}_t$ is a complete history of the algorithm up to time-step t that signifies all information including h_t, $(\alpha_k)_{k\le t}$ and the state sequence $(s_k)_{k\le t}$. We use $\mathbb{E}_{p_{h_t}}$ as an expectation operator with respect to p_{h_t}.

Noise is bounded. Now we show that the noise term is not a significant factor that affects the convergence of Q-learning. By construction it has a zero mean value:

$$\mathbb{E}_{p_{h_t}}\left[w_{h_t}(s,a)|\mathcal{F}_t\right] = \\ \mathbb{E}_{p_{h_t}}\left[r' + \gamma \max_{\tilde{a}} q_t(s',\tilde{a}) - F_{h_t} q_t(s,a) | \mathcal{F}_t\right] = 0. \quad (9)$$

Due to the bounded reward assumption the variance of the noise term is also bounded.

$$\begin{aligned}
\mathbb{E}_{p_{h_t}}\left[w^2_{h_t}(s,a)|\mathcal{F}_t\right] &= \mathbb{E}_{p_{h_t}}\left[\left(r' + \gamma \max_{\tilde{a}} q_t(s',\tilde{a})\right)^2 |\mathcal{F}_t\right] \\
&\quad - \mathbb{E}_{p_{h_t}}\left[r' + \gamma \max_{\tilde{a}} q_t(s',\tilde{a})|\mathcal{F}_t\right]^2 \\
&\overset{(a)}{\le} \frac{1}{4}\left(\max_{s'r'}\left(r' + \gamma \max_{\tilde{a}} q_t(s',\tilde{a})\right)\right. \\
&\quad \left. - \min_{s'r'}\left(r' + \gamma \max_{\tilde{a}} q_t(s',\tilde{a})\right)\right)^2 \\
&\overset{(b)}{\le} \frac{1}{4}\left(\frac{r_{\max} - r_{\min}}{1-\gamma} + \gamma ||q_t||_\infty\right)^2 \\
&\overset{(c)}{\le} A + B||q_t||^2_\infty \qquad (10)
\end{aligned}$$

(a) follows from Popoviciu's inequality, $\mathrm{Var}(X) \le {}^1\!/_4(\max X - \min X)^2$, (b) is due to the bounded rewards assumption; and (c) results from some algebra with constants $A := {}^\Delta\!/_4\left({}^{2\gamma}\!/_{1-\gamma} + \Delta\right)$ and $B := {}^{\gamma^2}\!/_4$, where $\Delta := (r_{\max} - r_{\min})/(1-\gamma)$. We denote a *sup-norm* by $||\cdot||_\infty$.

***F_h* is a contraction.** For a fixed history h, we show that an operator F_h is a contraction mapping.

$$\begin{aligned}
&||F_h q - F_h q'||_\infty \\
= &\max_{s,a}\left|\mathbb{E}_{p_{h_t}}\left[r' + \gamma \max_{\tilde{a}} q(s',\tilde{a})|\mathcal{F}_t\right]\right. \\
&\left. - \mathbb{E}_{p_{h_t}}\left[r' + \gamma \max_{\tilde{a}} q'(s',\tilde{a})|\mathcal{F}_t\right]\right| \\
\overset{(a)}{=} &\max_{s,a}\left|\mathbb{E}_{p_{h_t}}\left[r' + \gamma \max_{\tilde{a}} q(s',\tilde{a})|sa\right]\right. \\
&\left. - \mathbb{E}_{p_{h_t}}\left[r' + \gamma \max_{\tilde{a}} q'(s',\tilde{a})|sa\right]\right| \\
\le &\max_{s,a}\max_{s'}\gamma\left|\max_{\tilde{a}} q(s',\tilde{a}) - \max_{\tilde{a}} q'(s',\tilde{a})\right| \\
\le &\gamma \max_{s,a}|q(s,a) - q'(s,a)| = \gamma||q - q'||_\infty \quad (11)
\end{aligned}$$

(a) for a fixed history, the expectation only depends on the (s,a)-pair rather than the complete history $\mathcal{F}_t$.

Same fixed point. We show that for any history h, the contraction operator F_h has a fixed point q^*. Let h be mapped to state s:

$$\begin{aligned}
q^*(s,a) &\overset{(a)}{=} Q^*(h,a) \\
&\equiv \sum_{o'r'} P(o'r'|ha)\left(r' + \gamma \max_{\tilde{a}} Q^*(h',\tilde{a})\right) \\
&\overset{(b)}{=} \sum_{s'r'} p_h(s'r'|sa)\left(r' + \gamma \max_{\tilde{a}} q^*(s',\tilde{a})\right) \\
&\equiv F_h q^*(s,a) \qquad (12)
\end{aligned}$$

(a) is the QDP assumption; and (b) follows from (2) and again using the QDP assumption. We also show that for any history h the operator F_h has a same contraction factor γ.

$$||F_h q - q^*||_\infty \overset{(a)}{=} ||F_h q - F_h q^*||_\infty \overset{(b)}{\le} \gamma||q - q^*||_\infty \quad (13)$$

(a) follows from (12); and (b) is due to (11). Therefore, for any history h the operator F_h has the same fixed point q^* with the same contraction factor γ.

We have all the conditions to invoke a convergence result from Bertsekas and Tsitsiklis [1995]. We adopt[5] and state Proposition 4.5 from Bertsekas and Tsitsiklis [1995] without reproducing the complete proof.

Proposition 8 (Prop. 4.5 [Bertsekas and Tsitsiklis, 1995]) *Let $(q_t)_{t\in\mathbb{N}}$ be the sequence generated by the iteration (4). We assume the following.*

(a) The learning rates $\alpha_t(s,a)$ are nonnegative and satisfy,

$$\sum_{t=0}^{\infty}\alpha_t(s,a)=\infty, \qquad \sum_{t=0}^{\infty}\alpha_t^2(s,a)<\infty.$$

(b) The noise term $w_t(s,a)$ satisfies,

$$\begin{aligned}\mathbb{E}_{p_t}\left[w_t(s,a)|\mathcal{F}_t\right] &= 0,\\ \mathbb{E}_{p_t}\left[w_t^2(s,a)|\mathcal{F}_t\right] &\leq A+B||q_t||_\infty^2, \quad \forall s,a,t\end{aligned}$$

where, A and B are constants.

(c) There exists a vector q^*, and a scalar $\gamma\in[0,1)$, such that,

$$||F_tq_t-q^*||_\infty\leq\gamma||q_t-q^*||_\infty, \quad \forall t.$$

Then, q_t converges to q^* with probability 1.

Proof Sketch. We have a sequence of maps $(F_t)_{t\in\mathbb{N}}$. At any time-step t, the map F_t is a contraction and every map moves the iterates toward the same fixed point q^*. Since, the contraction factor is the same, the rate of convergence is not affected by the order of the maps. Every map contracts the iterates by a factor γ with respect to the fixed point that asymptotically converges to q^*. ■

Proposition 8 uses a sequence of maps $(F_t)_{t\in\mathbb{N}}$ with a same fixed point. In our case, we have this sequence based on histories, i.e.$(F_{h_t})_{t\in\mathbb{N}}$. Similarly, for w_t and p_t we have corresponding history-based instances. Therefore, Proposition 8 with $p_t=p_{h_t}$, $F_t=F_{h_t}$ and $w_t=w_{h_t}$ provides the main result. This can be done, since α_t, q_t, w_t and F_t are allowed to be random variables. ■

Obviously, state-uniformity is a necessary condition, since otherwise $Q^*(h,a)$ can not even be represented as $q^*(s,a)$. Typically, the state-process is assumed to be an MDP. This makes the state-process p_h independent of history, and leads to a history-independent operator $F:=F_h$ for any history h, which (trivially) all have the same fixed point. We relax this MDP assumption, and only demand state-uniformity of the optimal value-function. The proof shows that this condition is sufficient to provide a unique fix point for the *history-dependent* operators. Therefore, the state-uniformity is not only a necessary but also a sufficient condition for Q-learning convergence.

[5]The original proposition is slightly more general than we need for our proof. It has an extra diminishing noise term which we do not have/require in our formulation.

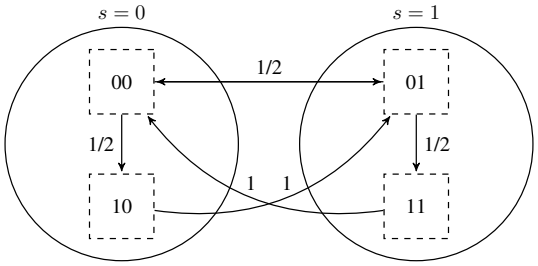

Figure 2: An example MRP aggregated to a non-MDP (non-MRP). The square nodes represent the states of the underlying MRP. The circles are the aggregated states. The solid arrows represent the only available action x and the transition probabilities are shown at the transition edges.

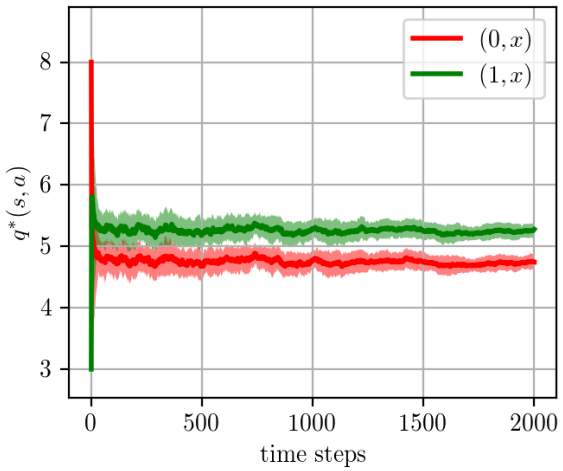

Figure 3: The learning curves of Q-learning are averaged over 40 independent runs with the parameters, $\gamma=0.9$, $q_0(s=0,x)=8$ and $q_0(s=1,x)=3$.

7 Empirical Evaluation

In this section, we empirically evaluate two example non-MDP domains to show the validity of our result.

Non-Markovian Reward Process. Let us consider our first example from [Hutter, 2016] to demonstrate Q-learning convergence to a non-Markovian reward processes (non-MRP). We consider that the underlying HDP is an MDP over the observation space $\mathcal{O}$ (in fact an action-independent MDP, i.e. an MRP) with a transition matrix T and a deterministic reward function R. The state diagram of the process is shown in Figure 2.

$$T=\begin{bmatrix}0 & 1/2 & 1/2 & 0\\ 1/2 & 0 & 0 & 1/2\\ 0 & 1 & 0 & 0\\ 1 & 0 & 0 & 0\end{bmatrix}, \qquad R=\begin{bmatrix}\frac{\gamma/2}{1+\gamma}\\ \frac{1+\gamma/2}{1+\gamma}\\ 0\\ 1\end{bmatrix}. \quad (14)$$

Due to this structure, the HDP is expressible as, $P(o'r'|ha)=T_{oo'}\cdot[\![r'=R(o)]\!]$, such that h has a last observation o, where $[\![\cdot]\!]$ denotes an Iverson bracket. The observation space is $\mathcal{O}=\{00,01,10,11\}$. Let us consider the state

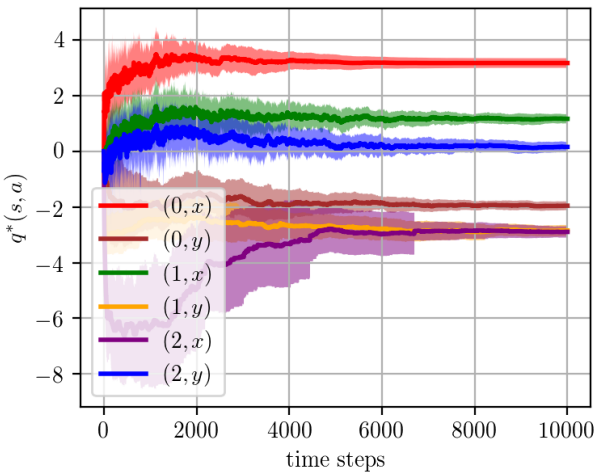

Figure 4: The learning curves of Q-learning are averaged over 50 independent runs with the parameters, $\gamma = 0.9$, $p_{\min} = 0.01$ and $q_0(s,a) = 0$ for all s and a.

space $S = \{0, 1\}$, and the agent experiences the state-process under the following aggregation map:

$$s_t := \phi(h_t) := \begin{cases} 0, & \text{if } o_t = 00 \text{ or } 10, \\ 1, & \text{if } o_t = 01 \text{ or } 11. \end{cases}$$

It is easy to see that the resultant state-process is not an MDP (MRP):

$$\begin{aligned} p_{00}(s' = 0|s = 0) &= T_{00,00} + T_{00,10} = 0 + {}^1\!/_2 = {}^1\!/_2 \\ p_{10}(s' = 0|s = 0) &= T_{10,00} + T_{10,10} = 0 + 0 = 0 \end{aligned}$$

which implies $p_{00} \neq p_{10}$, but this state-process satisfies the optimal Q-value function state-uniformity condition (see below). Hence, the state-process is a QDP $\in$ POMDP$\backslash$MDP: the underlying HDP has a finite set of hidden states, i.e. the states of the underlying MDP. Since it is an action-independent process, the action-value function is the same for any action $a \in \mathcal{A}$. We denote the only available action with x:

$$\begin{aligned} q^*(s = 0, x) &:= Q^*(00, x) = Q^*(10, x) = \frac{\gamma}{1 - \gamma^2} \\ q^*(s = 1, x) &:= Q^*(01, x) = Q^*(11, x) = \frac{1}{1 - \gamma^2}. \end{aligned}$$

We apply Q-learning to the induced state-process and the learning curves plot is shown in Figure 3. The plot shows that Q-learning is able to converge to the optimal action-value function of the process, despite the fact that the state-process is a non-MDP (non-MRP).

Non-Markovian Decision Process. The previous example demonstrated that Q-learning is able to learn a non-MRP $\in$ QDP. Now we provide an example QDP which is a two-action non-MDP $\in$ HDP$\backslash$POMDP: the state space of the underlying HDP is infinite. The agent is facing a *non-stationary* state-process with state space $\mathcal{S} = \{0, 1, 2\}$ and action space $\mathcal{A} = \{x, y\}$. The agent has to input a right *key*-action k_s at a state $s = \phi(h)$ but the environment accepts the key action with a certain history-dependent probability $p_v(h)$ by providing an observation from $\mathcal{O} = \{v, i\}$, where v and i indicate acceptance or rejection of an input, respectively.

$$p_v(h) := \max\{p_{\min}, \%(v, h)\} \tag{15}$$

where, $\%(v, h)$ is the percentage of accepted keys in h and $p_{\min}$ is a minimum acceptance probability. Without loss of generality, we use the key sequence $k_0 := x, k_1 := x, k_2 := y$. The history to state mapping is defined as:

$$\phi(h) = \begin{cases} 0 & \text{if, } h = \begin{cases} \epsilon \\ \ldots v \end{cases} \\ 1 & \text{if, } h = \begin{cases} xi \\ yi \\ \ldots vxi \\ \ldots vyi \end{cases} \\ 2 & \text{if, } h = \ldots \dot{h} \text{ such that } |\dot{h}| \geq 2 \text{ and } v \notin \dot{h} \end{cases} \tag{16}$$

It is apparent from the mapping function that *state-0* is the start state, and it is also the case when a key is accepted in the last time-step, *state-1* is defined when a key input is rejected once, and *state-2* is reached when the key input has been recently rejected at least twice in a row.

The transition probabilities are formally given as follows (see Figure 5 for a graphical representation):

$$p_h(s'|s, a) = \begin{cases} p_v(h) & \text{if, } s' = 0, s = 0|1|2, a = k_s \\ 1 - p_v(h) & \text{if, } \begin{cases} s' = 1, s = 0, a = k_s \\ s' = 2, s = 1, a = k_s \\ s' = 2, s = 2, a = k_s \end{cases} \\ 1 & \text{if, } \begin{cases} s' = 1, s = 0, a \neq k_s \\ s' = 2, s = 1, a \neq k_s \\ s' = 2, s = 2, a \neq k_s \end{cases} \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

The reward is also a function of the complete history.

$$r'(ha) = \begin{cases} 3 - \gamma - 2\gamma p_v(h) & \text{if, } \phi(h) = 0, a = k_0, \\ 1 - 3\gamma p_v(h) & \text{if, } \phi(h) = 1, a = k_1, \\ -3\gamma p_v(h) & \text{if, } \phi(h) = 2, a = k_2, \\ -3 & \text{if, } \phi(h) = s, a \neq k_s \end{cases} \tag{18}$$

It is easy to see that Q-values are only a function of state-action pairs as follows for $\phi(h) = s$:

$$q^*(s, a) = Q^*(h, a) = \begin{cases} 3 & \text{if, } s = 0, a = x, \\ -2 & \text{if, } s = 0, a = y, \\ 1 & \text{if, } s = 1, a = x, \\ -3 & \text{if, } s = 1, a = y, \\ -3 & \text{if, } s = 2, a = x, \\ 0 & \text{if, } s = 2, a = y. \end{cases} \tag{19}$$

Despite the history-based dynamics, Figure 4 shows that Q-learning is able to learn the Q-values of the *non-stationary* state-process due to the fact that it is a QDP.

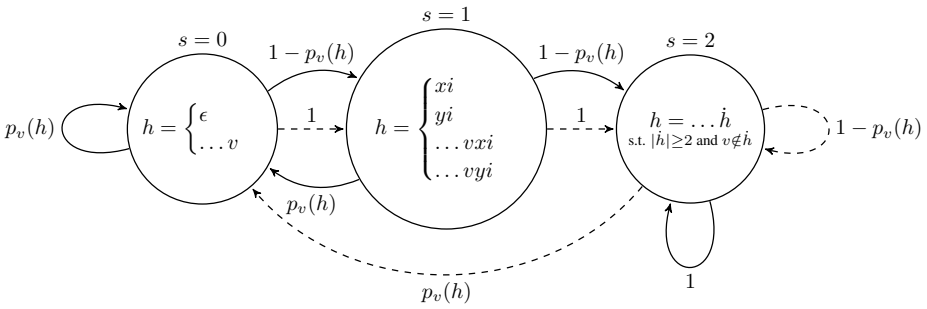

Figure 5: A complete history-dependent process is aggregated to a 3-state non-MDP. The circles are states. Inside the states are the corresponding history patterns mapped to the state. For clarity, the rewards are not shown in the history patterns. The action x is denoted by the solid arrows while the action y is denoted by the dashed arrows. The transition probabilities are indicated at the transition edges.

8 Conclusion

In this paper, we proved that Q-learning convergence can be extended to a much larger class of decision problems than finte-state MDP. We call this class of environments QDP. In QDP, the optimal action-value function of the state-process is still only a function of states, but the dynamics can be a function of the complete history (in effect, a function of time). That enables QDP to allow non-stationary domains in contrast to finite-state MDPs that can only model stationary domains. We also showed that this state-uniformity condition is not only a necessary but also a sufficient condition for Q-learning convergence. An empirical evaluation of a few non-MDP domains is also provided. In the state-aggregation context, we only extended the convergence for an *exact* aggregation case. A natural next step is to investigate if this proof can be extended to the *approximate* aggregation case. Approximate aggregation is a special case of function approximation, but there are known counter-examples of Q-learning diverging with function approximation [Baird, 1995]. Therefore, it is intriguing to know if there exist some non-trivial conditions that provide convergence guarantee for Q-learning in the approximate aggregation case and yet avoid these counter-examples.

References

[Baird, 1995] Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier, 1995.

[Bertsekas and Tsitsiklis, 1995] Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming: an overview*, volume 1. Belmont: Athena Scientific, 1995.

[Cassandra *et al.*, 1994] Anthony R Cassandra, Leslie Pack Kaelbling, and Michael L Littman. Acting optimaly in partially observable stochastic domains. *AAAI*, pages 1023–1023, 1994.

[Cassandra, 1994] Anthony R Cassandra. Optimal Policies for Partially Observable Markov Decision Processes. Technical Report August, Report CS-94-14, Brown Univ, 1994.

[Hutter, 2005] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.

[Hutter, 2016] Marcus Hutter. Extreme state aggregation beyond Markov decision processes. *Theoretical Computer Science*, 650:73–91, 2016.

[Leike, 2016] Jan Leike. *Nonparametric General Reinforcement Learning*. PhD thesis, Australian National University, 2016.

[Li *et al.*, 2006] Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for mdps. In *ISAIM*, 2006.

[Lin and Mitchell, 1992] L.J. Lin and T.M. Mitchell. Memory approaches to reinforcement learning in non-Markovian domains. *Artificial Intelligence*, 8(7597):28, 1992.

[Littman, 1994] Michael L Littman. Memoryless policies: Theoretical limitations and practical results. In *From Animals to Animats 3: Proceedings of the third international conference on simulation of adaptive behavior*, volume 3, page 238. MIT Press, 1994.

[McCallum, 1995] R. Andrew McCallum. Instance-based utile distinctions for reinforcement learning with hidden state. *ICML*, pages 387–395, 1995.

[Murphy, 2000] Kevin P Murphy. A survey of pomdp solution techniques. *environment*, 2:X3, 2000.

[Pendrith and McGarity, 1998] Mark D Pendrith and Michael J McGarity. An analysis of direct reinforcement learning in non-Markovian domains. *ICML*, pages 421–429, 1998.

[Perkins and Pendrith, 2002] Theodore J Perkins and Mark D Pendrith. On the existence of fixed points for Q-learning and Sarsa in partially observable domains. *ICML*, pages 490–497, 2002.

[Singh *et al.*, 1994] S Singh, T Jaakkola, and M Jordan. Learning Without State-Estimation in Partially Observable Markov Decision Processes. *ICML*, 31(0):37, 1994.

[Sutton and Barto, 1984] R S Sutton and A G Barto. Temporal credit assignment in reinforcement learning. *Computer Science*, page 210, 1984.

[Sutton, 1988] Richard S Sutton. Learning to Predict by the Method of Temporal Differences. *Machine Learning*, 3(1):9–44, 1988.

[Thrun *et al.*, 2005] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. Partially Observable Markov Decision Processes. *Probabilistic Robotics*, page 2005, 2005.

[Tsitsiklis, 1994] John N. Tsitsiklis. Asynchronous Stochastic Approximation and Q-Learning. *Machine Learning*, 16(3):185–202, 1994.

[Watkins and Dayan, 1992] Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

[Whitehead and Lin, 1995] Steven D. Whitehead and Long Ji Lin. Reinforcement learning of non-Markov decision processes. *Artificial Intelligence*, 73(1-2):271–306, 1995.