On Q-learning Convergence Beyond Markov Decision Processes

Sultan Javed Majeed and Marcus Hutter

Research School of Computer Science, Australian National University

International Joint Conference on Artificial Intelligence
July 16th, 2018
Artificial General Intelligence (AGI) agents are versatile.

An AGI agent needs to perform “well” in a wide range of environments.

One of the weakest forms of performing “well” is to converge on the optimal policy asymptotically.

The General Reinforcement Learning (GRL) framework can (possibly) realise an AGI agent.

Arguably, GRL admits the largest possible class of environments. (details in the next slide)
A Typical GRL Setup

- The agent and the environment interact in cycles.
- This interaction generates a history $h$.
- The agent takes an action $a$, then the environment provides an observation-reward tuple $(o', r')$.
- The history extends for the next cycle as $h' = hao'r'$.
- There are no restrictions on the environment dynamics $P(o'r'|ha)$.
- Every history is unique and appears at most once.
- Hence, in general, this History-based Decision Process (HDP) is not learnable.
A model $\phi$ which sends histories to a finite set of states.

The modeling results in a marginalized process

$$P_\phi(s' r'| h a) = \sum_{o'}: \phi(hao'r') = s' P(o'r'| h a).$$

**Definition: A Markov Decision Process (MDP) Model**

A model $\phi$ is an MDP if there exists a $p$ such that

$$p(s' r'| sa) = P_\phi(s' r'| h a) \forall a, h : \phi(h) = s.$$  

In words: next state-reward probability only depends on $h$ through $\phi(h)$.

An MDP model has state-based/stationary Markovian dynamics, (optimal) Q-function, and optimal policies.

Q-learning, an off-policy algorithm, converges in MDPs.
Going Beyond MDP Models

- An MDP model is restrictive, e.g. it *cannot* model non-stationarity.

- Often, an **aggregated MDP** is not an MDP anymore.

- However, it provides a **necessary condition**\(^1\) for Q-learning convergence by preserving\(^1\) the optimal Q-function.

- Which is a **strong condition**\(^2\) for convergence of Q-learning.

- The (optimal) Q-function preservation is not only necessary but the **sufficient condition**\(^1\) for Q-learning to converge.

---

\(^1\)A preserved quantity is modeled perfectly by the model.
\(^2\)One of our main results, more details later.
(Compartively) Less Restrictive Subclass of HDP

**Definition: A Q-uniform Decision Process (QDP) Model**

A model $\phi$ is a QDP if there exists a $q$ such that
$$q(s, a) = Q^*(h, a) \quad \forall a, h : \phi(h) = s.$$  

- QDP only preserves the **optimal** Q-function.
- The QDP class is **strictly larger** than MDP.
- It still admits **stationary optimal policies**.
- Whereas, the Partially Observable MDP (POMDP) class **does not** have stationary optimal policies.
Why do we need Q-learning for AGI?

- Q-learning, in the **tabular case**, converges in MDPs.
- As far as convergence is the only performance criteria, Q-learning can serve as a learning and/or planning module for an AGI for finite-MDPs.

**Definition: Q-learning (Sketch)**

The Q-learning algorithm applies the following **Q-iteration** for each time-step $t$,

$$q_{t+1}(s,a) = (1 - \alpha_t(s,a)) q_t(s,a) + \alpha_t(s,a) \left( r' + \max_b q_t(s',b) \right)$$

With a set of **appropriate** learning rates ($\alpha_t$), the Q-iteration asymptotically converges to the optimal.
Does Q-learning also converge in QDPs?
Yes, it does. Because,

- The operators are contractions, and
- they still have the same fix point.

**Theorem: Q-learning Convergence in QDPs**

Q-learning converges in QDPs, if the rewards are bounded and the set of learning rates satisfies the *appropriate conditions*\(^3\).

- Hence, Q-learning can also be used as a learning and/or planning module for an AGI for QDPs.
- The convergence also implies the existence of a stationary optimal policy.
- The preservation of optimal Q-function is not only necessary but a sufficient condition for Q-learning convergence.

\[3 \sum_{t=0}^{\infty} \alpha_t(s,a) = \infty, \sum_{t=0}^{\infty} \alpha_t^2(s,a) < \infty\]
Q-learning on a non-stationary (toy) domain

- The agent has to input the **right key**.
- The key acceptance probability is **non-stationary**.
- More **wrong inputs** in the past, lower the acceptance probability.
- But, the optimal Q-function is **not** a function of history.
Where to go from here?

- The exact $Q^*$-uniformity (i.e. preservation of the optimal Q-function) is brittle, an extension to the approximate $Q^*$-uniformity case is a natural next step.

- Can Q-learning also converge with high probability if the $Q^*$-uniformity condition is only met in expectation with small variance?

- Construct a natural sub-class of QDP environments beyond MDPs.

- Develop a QDP learning (i.e. $\phi$ learning) algorithm using Q-learning as a module.
Q-learning not only converges in MDPs but also beyond MDPs in QDPs, which include non-stationary domains.