ICIIBMS 2015, T2A-20953, Okinawa, Japan

Maximizing the average reward in episodic reinforcement learning tasks

Chris Reinke, Eiji Uchibe, Kenji Doya
Okinawa Institute of Science and Technology
Neural Computation Unit
Onna-son, Japan
chris.reinke@oist.jp

***Abstract*—We propose an ensemble method consisting of several Q-learning modules to optimize the average reward in episodic Markov decision processes (MDPs). It can be proven that the method learns and optimizes the average reward in MDPs where non-zero rewards are only given by transitions into goal states and the decision for a trajectory to a goal state is only possible in the start state. We introduced a sampling method for MDPs to show that the average reward can also be optimized to a high degree in MDPs which do not fulfill these conditions.**

***Keywords*—reinforcement learning; average reward; ensemble learning**

I. Introduction

A major aspect of intelligent agents is their ability to learn optimal decisions for a given task. One important goal is to learn the decisions that maximize the average outcome or reward over time, because it results in the highest amount of reward with respect to the invested time. This is critical for many natural problem settings such as in situations where the agent has to collect energy to perform other tasks. In such settings the agent needs to obtain as much energy with respect to the invested time so that more time and energy can be spend on the other tasks. Our research explored a new method to optimize the average reward in such episodic reinforcement learning tasks.

In reinforcement learning tasks are formulated as Markov decision processes (MDPs) [1] where the agent goes through a set of states S depending on its actions A. State transitions depend on the transfer function $T(s_t,a_t,s_{t+1})$ which defines the probability of resulting in state s_{t+1} if the agent performs action a_t in state s_t at time point t. At each state transition the agent receives a reward $r_t = R(s_t,a_t)$ resulting in a chain of rewards: $\tau = r_0 \rightarrow r_1 \rightarrow \ldots \rightarrow r_n$. MDPs are usually episodic or non-episodic. Episodic MDPs consist of several episodes each starting in an initial state and ending in a goal state. Many natural task settings are episodic such as foraging tasks in which the agent starts an episode in its home area and learns the way to an energy source, its goal state. In non-episodic MDPs goal states do not exist and the task has no end ($t \rightarrow \infty$) which is often used for control problems such as the balancing of an inverted pendulum.

The formulation of the goal is another important aspect of MDPs. One of the most common goals is to maximize the expected discounted reward (1) with a discount factor $\gamma \in [0,1]$.

$$E\left[\sum_{t=0}^{\infty}\gamma^t r_t\right] \tag{1}$$

Value based temporal difference (TD) methods such as Q-learning or SARSA solve this goal in episodic and non-episodic tasks by learning for every state-action pair the expected discounted reward $Q^\gamma(s,a) = E[r_0 + \gamma_1 r_1 + \gamma_2 r_2 + \ldots + \gamma_n r_n]$. They have the advantage of being simple, well defined and extensively studied [1]. A second important goal is the maximization of the expected average reward (2).

$$\lim_{h\to\infty}\frac{1}{h+1}E\left[\sum_{t=0}^{h}r_t\right] \tag{2}$$

It has been mainly studied in non-episodic tasks with specialized algorithms such as R-learning [2]. These algorithms are less studied than conventional TD methods and have often more learning parameters.

We explored the maximization of the average reward in episodic tasks with help of an ensemble method based on conventional TD algorithms due to their advantage of being simpler compared to more specialized algorithms such as R-learning.

II. Approach

The proposed method consists of an ensemble of Q-learning modules each learning with a different discount factor γ ranging from 0 to 1. We therefore refer to it as a *γ-Ensemble*. It was first introduced to explain human discounting behavior [3]. The value of a state-action pair for the γ-Ensemble is the integral over its modules (3).

$$Q^e(s,a) = \int_0^1 Q^\gamma(s,a)\,d\gamma \tag{3}$$

978-1-4799-8562-3/15/$31.00 ©2015 IEEE

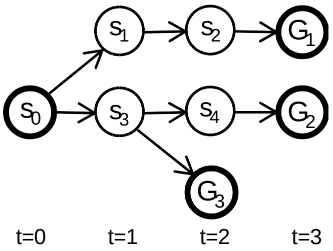


Fig. 1. Example of a sampled MDP.

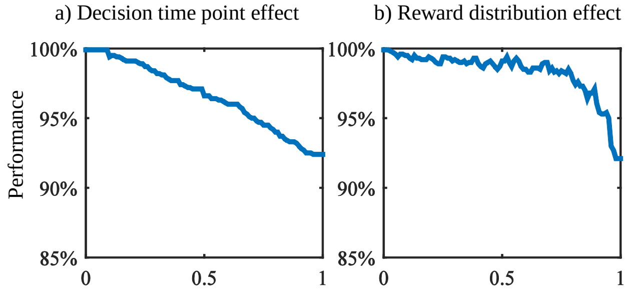
Each Q-module learns the expected discounted reward following the actions given by the γ-Ensemble (4).

$$Q^{\gamma}(s_t,a_t) = r_t + \gamma\, Q^{\gamma}\left(s_{t+1}, \arg\max_{a_{t+1}} Q^{e}(s_{t+1},a_{t+1}) \right) \tag{4}$$

The integral over the γ-Ensemble results in an hyperbolic discounting, i.e. a future reward r_t at time point t is discounted by the factor $1/(t+1)$ [3]. In contrast, classical TD methods such as Q-learning use an exponential discounting: γ^t.

The advantage of the hyperbolic discounting of the γ-Ensemble is that in tasks where non-zero rewards are only given when a goal state is reached ($r_0 = 0, r_1 = 0, \ldots, r_n = R$) its value is equivalent to the average reward: $Q^{e}(s_0,a_0) = R/(n+1)$. Therefore it can solve the average reward maximization problem. Unfortunately it computes the average reward of an episode only correctly for the start state of an MDP at $t = 0$. In states after the start state ($t > 0$) the value of the γ-Ensemble represents the average reward from the current state s_t to the goal state and not the average reward of the episode starting from the start state s_0. This results in an inconsistency effect in which an agent might not choose the trajectory to the goal state with the maximum average reward for an episode if the decision between trajectories to different goals is performed after the start state ($t > 0$). The state s_3 in Fig. 1 is such a decision point. Our goal was to analyze how strong the ability of the γ-Ensemble to maximize the average reward depends on the time point t at which decisions between trajectories have to be made. Furthermore we also wanted to analyze how much the average maximization depends on the condition that non-zero rewards are only given when goal states are reached.

We used a Monte Carlo sampling method to analyze the behavior of the γ-Ensemble. It samples episodic MDPs such as illustrated in Fig. 1 and computes if a γ-Ensemble would be able to optimize their average reward if it learned all Q-values correctly. MDPs are sampled by sampling first the number of goal states in an MDP. Then for each goal state the length of its trajectory, i.e. the number of states to reach it, and the sum of the rewards on the trajectory to it are sampled. At last the decision points between trajectories to the goal states are sampled. The procedure iterates over the trajectories. For each trajectory i one of the previous trajectories $j < i$ is randomly selected to which i gets connected. The time point t at which the trajectory i gets connected to j is then sampled from a distribution over the trajectory length of j.


a) Decision time point effect

b) Reward distribution effect

Fig. 2. Ability in percent to maximize the average reward in sampled MDPs (n=1000) for different MDP distributions (x-axis). MDPs have a uniform distribution over number of goal states U(2,5), length of trajectories U(1,50) and reward sum per trajectory U(1,50). (a) Effect of distributions over different decision time points between trajectories. 0 means that only at the start state a decision between goal trajectories exists and 1 means that decision points are uniformly distributed over all time points. (b) Effect of different reward distributions over trajectories. 0 means that reward is only given in goal states and 1 means the reward is uniformly distributed over the trajectory.

III. Results

We tested the γ-Ensemble on different distributions over MDPs. The results show that the γ-Ensemble is able to maximize the average reward for many MDPs even if the necessary conditions to ensure optimality are not fulfilled. In the case of MDPs where decision points between trajectories are uniformly distributed over all possible time points the average reward can still be optimized in more than 92% of the sampled MDPs (Fig. 2, a). The same can be observed in MDPs where rewards are distributed over the whole trajectories and not only given in goal states (Fig. 2, b). For MDPs where both, decision points and rewards, are uniformly distributed the ability drops to 70%.

IV. Conclusion

We proposed the γ-Ensemble as a possible solution to maximize the average reward in episodic reinforcement learning tasks. It has the ability to solve the average reward in MDPs where non-zero rewards are only given when a goal state is reached and the decision between trajectories to a goal state are performed in the start state. Nonetheless even if these conditions are violated it has the ability to converge to the optimal solution in most MDPs. Therefore it can be a viable alternative to other average reward algorithms which have often more parameters. Future research will compare the learning performance of the γ-Ensemble to other learning algorithms and will try to overcome the inconsistency problem associated with the hyperbolic discounting.

References

[1] R.S. Sutton and A.G. Barto, Reinforcement learning: An introduction. Cambridge Univ Press, 1998.

[2] A. Schwartz, “A reinforcement learning method for maximizing undiscounted rewards,” in Proceedings of the tenth international conference on machine learning, vol. 298, 1993, pp. 298-305.

[3] Z. Kurth-Nelson and A.D. Redish, “Temporal-difference reinforcement learning with distributed representations,” PloS One, 4(10):e7362, 2009

978-1-4799-8562-3/15/$31.00 ©2015 IEEE