A Distributed Q-Learning Approach for Variable Attention to Multiple Critics

Maryam Tavakol[1], Majid Nili Ahmadabadi[1,2], Maryam Mirian[1], and Masoud Asadpour[1]

[1] Cognitive Robotics Lab, College of Engineering, Faculty of Electrical and Computer Engineering, University of Tehran
[2] School of Cognitive Sciences, Institute for Research in Fundamental Researches
{maryam.tavakol,mnili,mmirian,asadpour}@ut.ac.ir

Abstract. One of the substantial concerns of researchers in machine learning area is designing an artificial agent with an autonomous behaviour in a complex environment. In this paper, we considered a learning problem with multiple critics. The importance of each critic for the agent is different, and attention of agent to them is variable during its life. Inspired from neurological studies, we proposed a distributed learning approach for this problem that is flexible against the variable attention. In this approach, there is a distinct learner for each critic that an algorithm is introduced for aggregating of their knowledge based on combination of model-free and model-based learning methods. We showed that this aggregation method could provide the optimal policy for this problem.

Keywords: Multiple critics, distributed Q-learning, model-based parameters, aggregation method.

1 Introduction

Designing an intelligent system with autonomous decision making ability is one of the principal concerns in machine learning area. In this paper we focus on a learning problem similar to decision making process in human with receiving feedbacks from different sources of reward. A system with multiple users or multiple goals is an example of the application field of this type of problem. We can model this problem as a Reinforcement Learning (RL) process with multiple critics. In RL methods, the agent evaluates its behaviour based on the received reinforcement signals from the environment [1]. In this problem, a set of weights is defined for the importance of critics and the agent attends to each of them according to the corresponding weights. The main problem is that critic's importance can be changed during the agent's life. This is necessary for the learning model to cope with this variable attention without requiring the learning process being done again. While the decision making process of human is flexible to these changes, existing RL methods are not applicable for this purpose.

There is a wealth of research in the domain of human and animal decision making. The neurological studies have revealed the existence of a key RL signal, the

T. Huang et al. (Eds.): ICONIP 2012, Part III, LNCS 7665, pp. 244–251, 2012.

temporal difference prediction error, in the brain [2], and some of the RL methods have been applied to these behavioural data [3]. On the other hand, a wide range of neural data suggests that there is more than one system for behavioural choice in the brain. In fact, beside the model-free learning for habitual control, there is a model-based system for goal-directed decision making in the brain [4][5]. So, we can modify RL methods in order to apply them in the problem of variable attention of agent to more than one critic. We consider a distributed approach for resolving the multiple-critic learning problem. A sort of modular approaches for the learning problem have been introduced. Some of them have been designed for a task with multiple goals [6][7][8], while the others have been used for the task decomposition and behaviour coordination of independent subtasks [9][10]. All of these methods assume that the subtasks are completely independent and they do not need an aggregation of the different decisions. In our approach, there is a distinct learner for each reward source and we need a function to aggregate their decisions in each state. In [11], it is shown that in a modular RL, it is impossible to construct an arbitration function that satisfies a few basic and desirable properties for the social choice. Thus, we proposed an algorithm that uses the parameters of the model-based system to address this problem. By combining the model-free and model-based learning methods, we use the advantages of both of them together. Fig. 1 shows how our approach places in learning structure. In the next section we introduce the details of our proposed approach and the correction algorithm, for aggregation of decisions. In section 3, mathematical terms are used to show that how the mentioned algorithm works optimally. Finally we conclude our approach in section 4.

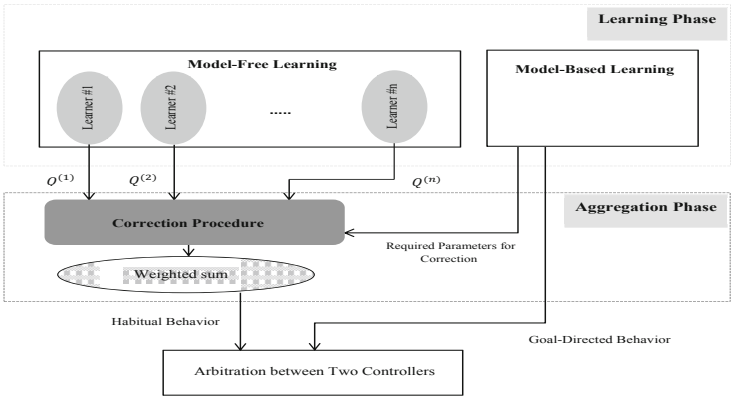

Fig. 1. The position of our proposed approach in the learning process

2 Proposed Approach

In this multiple-critic problem, the received reward from each critic is scaled based on the weight of the corresponding critic. These weights might be changed during the learning life of the agent as the attention of agent to critics changes.

We assume that all the received rewards are summed to form a total reward for learning the action values. However the problem is that if the attention of agent to critics changes, the learned values based on the weighted sum of the rewards will not be valid any more, and the learning has to start from the very beginning. To resolve this problem, a distributed architecture has been introduced that there is a distinct learner for each reward source to learn the Q-value of each state-action pair, independently. In order to apply a learning method to each learner, we preferred an off-policy method over on-policy ones. The Q-values in SARSA, as an on-policy method, rely on behaviour policy of the agent and the current weighted sum of the rewards. The Q-Learning seems a proper method for our purpose. The Q-tables are obtained independent of the behaviour policy and the weights of critics.

2.1 Problem Formulation

In this problem we use the traditional RL framework in an environment with Markov property. In this framework, S is the set of all the possible states and A is all the possible actions. The agent in the state $s \in S$ receives n feedbacks from the n different reward sources, say $r^{(1)}, r^{(2)}, ..., r^{(n)}$ by taking an action $a \in A$. For i^{th} critic, there is $R^{(i)}(s, a)$ as a distribution function for generating the reward sequence. The set of $W = \{w_1, w_2, ..., w_n\}$ indicates the weight of the critics, where $\sum_{i=1}^{n} w_i = 1$. In addition, $P^a_{ss'}$ determines the transition probability between two states. Moreover, $\gamma \in [0, 1]$ shows the discount factor for the delayed reward. When the learning is accomplished, the Q-value for the i^{th} learner will be $Q^{(i)}(s, a)$ for each state-action pair. While $Q_{opt}(s, a)$ will be an optimal Q-table obtained from the centralized learning system. By the centralized system we mean a learning process that tries to maximize $E\{\sum_{i=1}^{n} w_i r^{(i)}\}$. Finally, the Q-tables of the learners are combined based on current weight of the critics, and form Q_{dist} as the Q-table of distributed learning system, $Q_{dist}(s, a) = f(w_1, ..., w_n, Q^{(1)}(s, a), ..., Q^{(n)}(s, a))$. We proposed the approach for a problem with one terminal state with different paths. The agent should find the optimal path with the maximum weighted average reward.

2.2 Aggregation Method

The distributed system needs an aggregation function for combining the Q-tables to find the optimal policy. We chose the weighted sum of the Q-tables as Equation 1 Since we used this function in the centralized system for total reward computing and it is simple enough to work with a linear function.

$$Q_{dist}(s, a) = \sum_{i=1}^{n} w_i Q^{(i)}(s, a) \tag{1}$$

This aggregation function does not produce the optimal policy. So, we introduce a correction approach to make (1) to be optimal. The problem of the off-policy character of Q-learning is that the one-step value updates for each learner are

computed under the assumption that all future actions will be chosen optimally. While if there is an inconsistent state, the state that the optimal actions of the learners are different, in the path of a state-action to the terminal state, this assumption will not hold. We use the model of the environment to correct their values. As mentioned earlier, there is model-based RL that uses the experiences indirectly to build a model of the environment beside the habitual control. The combination of model-free and model-based decision making trades off between flexibility and computational complexity in one view, and two sources of uncertainty: ignorance and computational noise, in another view [12][13]. As Fig. 1 shows, the model of environment is learned in the model-based system and learned parameters are used in the aggregation phase to modify the Q-tables.

2.3 Correction Algorithm

Consider an agent in a stochastic environment with n learned Q-tables and it is going to make an optimal decision in state s. For each action a, the $Q^{(j)}(s,a)$ may need change for j^{th} learner, if two conditions will be hold: First, there is a non-zero probability in the learned model for reaching to an inconsistent state s' after taking a, say $Pr(s,a,s') > 0$. Second, a_1 will be the optimal aggregated action in the s', but $a_1 \neq \arg\max_{a'} Q^{(j)}(s',a') = a_2$. We assumed that s' has been corrected already. The correction procedure is applied to this Q-value based on the following equation.

$$Q^{(j)}_{new}(s,a) = Q^{(j)}(s,a) + Pr(s,a,s') * \gamma^k(-Q^{(j)}(s',a_2) + Q^{(j)}(s',a_1)), \qquad (2)$$

where k is the number of steps between s and s' in the optimal path. Three sets of parameters are required in the correction procedure that should be determined in the learning phase: The set of all the inconsistent states, IS, the transition probability from each state-action to each $s' \in IS$ and all the intermediate states, $Pr(s,a,s')$, and the number of steps between them, $step(s,a,s')$. Algorithm 1 shows the correction procedure for each state s.

In this algorithm $flag$ specifies the corrected states and θ is a threshold parameter that determines how much it is necessary to do correction procedure. In fact, when the transition probability to an inconsistent state is low, we can ignore this small change in order to have an affordable computation. In addition, it is possible for a state to change from consistent to inconsistent. We ignore this change for the first time but the affected states should be corrected in the next passes. Eventually, these modified Q-tables will be valid until the weights of critics change.

3 Analytical Justification

In this section, we are going to show the optimality of the distributed Q-learning approach in two theorems. The estimation of Q-values in the Q-learning for this problem is obtained based on (3) and (4) for the centralized system and i^{th} learner in the distributed system, respectively [14].

Algorithm 1. Correct(s)

Require: IS, Pr, step
  for all $a \in A(s)$ and $s' \in IS$ **do**
    if $Pr(s, a, s') > \theta$ **then**
      $Correct(s')$ if it is not corrected yet
      $a' := \arg\max_{a"} Q_{dist}(s', a")$
      for all $j \in \{1, ..., n\}$ where $a' \neq a_j : \arg\max_{a"} Q^{(j)}(s', a")$ **do**
        $Q^{(j)}_{new}(s, a) = Q^{(j)}(s, a) + Pr(s, a, s') * \gamma^{step(s,a,s')}(-Q^{(j)}(s', a_j) + Q^{(j)}(s', a'))$
        if $inconsistent(s) = true$ **then**
          add s to IS
          for all $s" \in S$ and $a" \in A(s")$ where $Pr(s", a", s) > \theta$ **do**
            $flag(s") = 0$
            $Pr(s", a", s') = Pr(s", a", s') - Pr(s", a", s)$
          end for
        end if
      end for
    end if
  end for
  $flag(s) = true$

$$Q_{opt}(s, a) = \sum_{i=1}^{n} w_i R^{(i)}(s, a) + \gamma \sum_{s'} P^a_{ss'} \max_{a'} Q_{opt}(s', a') \tag{3}$$

$$Q^{(i)}(s, a) = R^{(i)}(s, a) + \gamma \sum_{s'} P^a_{ss'} \max_{a'} Q^{(i)}(s', a') \tag{4}$$

It should be noted that the proofs for the uncertain conditions are straight forward and they are omitted from our proofs because of the space limitation. We illustrate that the following theorems are hold.

Theorem 1. *In the distributed system for each state, s, and action, a, if there is no inconsistent state on their path to the terminal state, (1) and (3) will be equivalent.*

Proof. Consider a consistent state, s, with a as an optimal action of all the learners, say for $i = 1, .., n : a = \arg\max_{a'} Q^{(i)}(s, a)$. Hence, a will be the optimal action in the aggregated form as follows:

$$\forall a' \in A(s) : Q^{(i)}(s, a') < Q^{(i)}(s, a) \quad \Rightarrow \sum_{i=1}^{n} w_i Q^{(i)}(s, a') < \sum_{i=1}^{n} w_i Q^{(i)}(s, a)$$

The mathematical induction is used to show the equivalence of two formulas, backward from the terminal state.

1. $N = 1$: Consider s_1 as a state that exactly takes place before the terminal state and a_1 is the optimal action of the learners. For $i = 1, ..., n$:

$$Q^{(i)}(s_1, a_1) = R^{(i)}(s_1, a_1) + \gamma \times 0 \quad \Rightarrow Q_{dist}(s_1, a_1) = \sum_{i=1}^{n} w_i R^{(i)}(s_1, a_1)$$

On the other hand, the Q-value of the centralized system will be as follows where both of them are equal.

$$Q_{opt}(s_1, a_1) = \sum_{i=1}^{n} w_i R^{(i)}(s_1, a_1) + \gamma \times 0$$

2. $N = k$: Assume for the consistent state s_k in k steps before the terminal state, (5) will be hold. Hence, a_k will be the optimal action of two systems.

$$Q_{dist}(s_k, a_k) = Q_{opt}(s_k, a_k) \tag{5}$$

3. $N = k + 1$: Let a_{k+1} takes the agent from the state s_{k+1} to s_k. The Q-value for this state-action pair is obtained based on the following equations.

$$Q^{(i)}(s_{k+1}, a_{k+1}) = R^{(i)}(s_{k+1}, a_{k+1}) + \gamma \times Q^{(i)}(s_k, a_k)$$

In the above equation, $Q^{(i)}(s_k, a_k) = max_{a'} Q^{(i)}(s_k, a')$ will be hold based on the assumption of induction. So, for the distributed and centralized system we have as follows that the result of two systems are equivalent.

$$\Rightarrow Q_{dist}(s_{k+1}, a_{k+1}) = \sum_{i=1}^{n} w_i R^{(i)}(s_{k+1}, a_{k+1}) + \gamma \underbrace{\sum_{i=1}^{n} w_i Q^{(i)}(s_k, a_k)}_{(1) \Rightarrow Q_{dist}(s_k, a_k)}$$

$$Q_{opt}(s_{k+1}, a_{k+1}) = \sum_{i=1}^{n} w_i R^{(i)}(s_{k+1}, a_{k+1}) + \gamma Q_{opt}(s_k, a_k) \qquad \square$$

Theorem 2. *If there is at least an inconsistent state between s and the terminal state while the agent takes the action a, the correction procedure will make (1) and (3) equivalent.*

Proof. It is assumed that all the intermediate states corrected already. The correction for the current state-action will be taken by the closest inconsistent state, s_{incns}. Consider, in s_{incns} the following is hold where a is the optimal action after correction:

$$a = \arg\max_{a"} Q^{(i)}(s_{incns}, a") \quad i \neq j \ \textit{and} \ a' = \arg\max_{a"} Q^{(j)}(s_{incns}, a") \quad i = j$$

First, we show the proof for the one step before the inconsistent state. Let s_1 be a consistent state and a_1 is the optimal action of each learner to s_{incns}. So, in the distributed case there is

$$Q^{(i)}(s_1, a_1) = R^{(i)}(s_1, a_1) + \gamma Q^{(i)}(s_{incns}, a) \quad i \neq j,$$

$$Q^{(j)}(s_1, a_1) = R^{(j)}(s_1, a_1) + \gamma Q^{(j)}(s_{incns}, a') \quad i = j,$$

$$\Rightarrow Q_{dist}(s_1, a_1) = \sum_{i=1}^{n} w_i R^{(i)}(s_1, a_1) + \gamma \sum_{i=1, i\neq j}^{n} w_i Q^{(i)}(s_{incns}, a) + \gamma w_j Q^{(j)}(s_{incns}, a'),$$

and for the centralized one,

$$Q_{opt}(s_1, a_1) = \sum_{i=1}^{n} w_i R^{(i)}(s_1, a_1) + \gamma \sum_{i=1}^{n} w_i Q^{(i)}(s_{incns}, a).$$

Hence, the correction based on (2) will make two results equivalent as follows:

$$\begin{aligned} Q^{(j)}_{new}(s_1, a_1) &= Q^{(j)}(s_1, a_1) - \gamma Q^{(j)}(s_{incns}, a') + \gamma Q^{(j)}(s_{incns}, a) \\ &= R^{(j)}(s_1, a_1) + \gamma Q^{(j)}(s_{incns}, a). \end{aligned}$$

In this situation, the optimal action may be changed for j^{th} learner. This change makes the state to be inconsistent.

If we propagate this effect to the s_k and a_k, while the intermediate states are consistent, there will be the following equations for two systems. Clearly, using (2) makes both results to be equivalent.

$$\begin{aligned} Q_{dist}(s_k, a_k) = &\sum_{i=1}^{n} w_i [R^{(i)}(s_k, a_k) + ... + \gamma^{k-1} R^{(i)}(s_1, a_1)] \\ &+\gamma^k \sum_{i=1, i\neq j}^{n} w_i Q^{(i)}(s_{incns}, a) + \gamma^k w_j Q^{(j)}(s_{incns}, a') \\ Q_{opt}(s_k, a_k) = &\sum_{i=1}^{n} w_i [R^{(i)}(s_k, a_k) + ... + \gamma^{k-1} R^{(i)}(s_1, a_1)] + \gamma^k \sum_{i=1}^{n} w_i Q^{(i)}(s_{incns}, a) \end{aligned}$$

□

Therefore, the correction procedure modifies the Q-values of the state-action pairs one by one if needed and the weighted sum of them gives the optimal policy. But in our algorithm, the agent does not move backward and just the current state-action will be corrected based on the inconsistent states, recursively. So, if a state changes from consistent to inconsistent after the correction of the earlier states, the agent may miss the optimal action for the first round.

4 Discussion

In this paper, we proposed a new learning method for a real problem with multiple critics. We introduced a distributed approach inspired from decision making process in the human brain that considers a distinct Q-learner for each critic. Using the Q-learning as an off-policy method makes the Q-tables to be learned independent of the current importance of critics and any behaviour policy. So, the agent learns just based on the received rewards and its learning result will remain usable when its attention to the critics changes. Then the Q-tables are

modified for each set of weights based on the learned model without any relearning. Finally, the weighted sum of them determines the greedy policy of the agent. The combination of model-free and model-based learning is helpful. Because model-free control is inflexible to change, while model-based choices are computationally expensive. Hence, when the learned model is inaccurate the former will be used, while the later will be preferred when there are variations in the environment. We will pursue the effect of inaccurate model in our approach in the future works. In the future works, we will study how changing θ affects the resulting policy. In addition, we will investigate the difference of performance of our approach with the optimal system when each new inconsistent state is ignored for the first time.

References

1. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT Press (1998)
2. Bayer, H.M., Glimcher, P.W.: Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47, 129–141 (2005)
3. Niv, Y.: Reinforcement learning in the brain. Journal of Mathematical Psychology 53, 139–154 (2009)
4. Dayan, P., Daw, N.D.: Decision theory, reinforcement learning, and the brain. Cognitive, Affective, & Behavioral Neuroscience 8, 429–453 (2008)
5. Gläscher, J., Daw, N., Dayan, P., O'Doherty, J.P.: States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66, 585–595 (2010)
6. Shelton, C.R.: Balancing multiple sources of reward in reinforcement learning. DTIC Document (2006)
7. Raicevic, P.: Parallel reinforcement learning using multiple reward signals. Neurocomputing 69, 2171–2179 (2006)
8. Sprague, N., Ballard, D.: Multiple-goal reinforcement learning with modular sarsa (0). In: International Joint Conference on Artificial Intelligence, vol. 18, pp. 1445–1447 (2003)
9. Park, K.H., Kim, Y.J., Kim, J.H.: Modular Q-learning based multi-agent cooperation for robot soccer. In: Robotics and Autonomous Systems, vol. 35, pp. 109–122 (2001)
10. Samejima, K., Doya, K., Kawato, M.: Inter-module credit assignment in modular reinforcement learning. Neural Networks 16, 985–994 (2003)
11. Bhat, S., Isbell, C.L., Mateas, M.: On the difficulty of modular reinforcement learning for real-world partial programming. In: Proceedings of the National Conference on Artificial Intelligence, vol. 21, p. 318. AAAI Press (2006)
12. Daw, N.D.: Model-based reinforcement learning as cognitive search: Neurocomputational theories. Evolution, Algorithms and the Brain (2011)
13. Simon, D.A., Daw, N.D.: Environmental statistics and the trade-off between model-based and TD learning in humans. In: Advances in Neural Information Processing Systems, vol. 24, pp. 127–135 (2011)
14. Szepesvári, C.: Algorithms for reinforcement learning. Algorithms for Reinforcement Learning 4, 1–103 (2010)