Decentralized Multi-Agent Reinforcement Learning in Average-Reward Dynamic DCOPs (Theoretical Proofs)

Duc Thien Nguyen[†], William Yeoh[‡], Hoong Chuin Lau[†], Shlomo Zilberstein[⋆], and Chongjie Zhang[*]

[†]School of Information Systems, Singapore Management University, Singapore 178902
{dtnguyen.2011,hclau}@smu.edu.sg

[‡]Department of Computer Science, New Mexico State University, Las Cruces, NM 88003, USA
wyeoh@cs.nmsu.edu

[⋆]School of Computer Science, University of Massachusetts, Amherst, MA 01003, USA
shlomo@cs.umass.edu

[*]CSAIL, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
chongjie@csail.mit.edu

Abstract. In this document, we show the proofs for the theoretical results described in the paper titled "Decentralized Multi-Agent Reinforcement Learning in Average-Reward Dynamic DCOPs" submitted to AAAI 2014.

In this paper, we consider MDPs where a joint state can transition to any other joint state with non-zero probability, that is, the MDP is *unichain*. We are going to show the decomposability of the value function of a unichain MDP, thereby leading us to the property that the Distributed RVI Q-learning algorithm converges to an optimal solution.

It is known that there always exists an optimal solution for a given unichain MDP and this solution can be characterized by the $V^*(\mathbf{s})$ value:

Theorem 1. ***(Puterman, 2005)*** *There exists an optimal Q-value $Q^*(\mathbf{s},\mathbf{d})$ for each joint state $\mathbf{s}$ and joint action $\mathbf{a}$ in an average-reward unichain MDP with bounded reward function satisfying:*

$$Q^*(\mathbf{s},\mathbf{d}) + \rho^* = \mathbf{F}(\mathbf{s},\mathbf{d}) + \sum_{\mathbf{s}'} P(\mathbf{s}',\mathbf{s},\mathbf{d}) \max_{\mathbf{d}'\in\mathbf{D}} Q^*(\mathbf{s}',\mathbf{d}') \tag{1}$$

To simplify the analysis, we assume that the sets of global joint states $\mathbf{S}$ and global joint values $\mathbf{D}$ are finite, and that the Markov chains for all the agents, induced by any policy, are aperiodic. A Markov chain is aperiodic when it converges to its stationary distribution in the limit (Puterman, 2005).

Additionally, there exists a unique V-value $V^*(\mathbf{s}) = \max_{\mathbf{d}\in\mathbf{D}} Q^*(\mathbf{s},\mathbf{d})$ for each joint state $\mathbf{s}$ such that

$$V^*(\mathbf{s}) + \rho^* = \max_{\mathbf{d}\in\mathbf{D}}[\mathbf{F}(\mathbf{s},\mathbf{d}) + \sum_{\mathbf{s}'} P(\mathbf{s}',\mathbf{s},\mathbf{d}) V^*(\mathbf{s}')] \tag{2}$$

with $V^*(\mathbf{s}^0) = 0$ *for any initial state* $\mathbf{s}^0$.

To help us to prove the decomposability of the value function, we first show the decomposibility of average reward ρ^* of a given optimal policy:

Lemma 1. *For a given unichain MDP, the optimal average reward* $\rho^* = \sum_{i=1}^{m} \rho_i^*$ *can be decomposed into a sum of local average rewards* ρ_i^* *for each reward function* $f_i \in \mathcal{F}$.

PROOF SKETCH OF LEMMA 1: For a given unichain MDP, there always exists a stationary distribution $P^\pi(\mathbf{s})$ of the global joint state $\mathbf{s} \in \mathbf{S}$ in the limit, where π is the converged global joint policy. Hence, we have the existence of

$$\rho_i^\pi = \sum_{\mathbf{s} \in \mathbf{S}} P^\pi(\mathbf{s}) f_i(\mathbf{s}_i, \mathbf{d}_i \mid \mathbf{s}_i \in \mathbf{s}, \mathbf{d}_i = \pi(\mathbf{s}))$$

for each reward function $f_i \in \mathcal{F}$. ■

From the decomposability of average reward given by Lemma 1 and the characteristic of V^* value given in Theorem 1, we now prove the decomposability of V^* as follows:

Definition 1. $\bar{P}_i(\mathbf{s}', \mathbf{s}, \mathbf{d}_i)$ *is the probability of transitioning to joint state* $\mathbf{s}'$ *from joint state* $\mathbf{s}$ *given joint value* $\mathbf{d}_i$ *and other values following policy* Φ *with* $V_j^*(\mathbf{s}_j^0) = 0$ *for each reward function* $f_j \in \mathcal{F}$.

Theorem 2. *There exists* $V_i^*(\mathbf{s}) = Q_i^*\big(\mathbf{s}, \mathbf{d}_i \mid \mathbf{d}_i \in \operatorname{argmax}_{\mathbf{d} \in \mathbf{D}} Q^*(\mathbf{s}, \mathbf{d})\big)$ *and* ρ_i *for each reward function* $f_i \in \mathcal{F}$ *under an optimal policy* $\Phi(\mathbf{s}) = \mathbf{d}_i \in \operatorname{argmax}_{\mathbf{d} \in \mathbf{D}} Q^*(\mathbf{s}, \mathbf{d})$ *such that*

$$\begin{aligned} V_i^*(\mathbf{s}) + \rho_i^* &= f_i(\mathbf{s}_i, \mathbf{d}_i \mid \mathbf{s}_i \in \mathbf{s}, \mathbf{d}_i \in \operatorname*{argmax}_{\mathbf{d} \in \mathbf{D}} Q^*(\mathbf{s}, \mathbf{d})) \\ &+ \sum_{\mathbf{s}'} \bar{P}_i(\mathbf{s}', \mathbf{s}, \mathbf{d}_i \mid \mathbf{d}_i \in \operatorname*{argmax}_{\mathbf{d} \in \mathbf{D}} Q^*(\mathbf{s}, \mathbf{d}))\, V_i^*(\mathbf{s}') \end{aligned} \tag{3}$$

and $V^*(\mathbf{s}) = \sum_i V_i^*(\mathbf{s})$.

PROOF SKETCH OF THEOREM 2: We do not show how to decompose $Q^*(\mathbf{s}, \mathbf{d})$ into $Q_i^*(\mathbf{s}, \mathbf{d}_i)$ but only show that there exists such a decomposition. The proof is based on the uniqueness of an optimal solution for any unichain MDP, which is given by Theorem 1.

Step 1: We first propose a modified MD-DCOP and decompose it into a set of subproblems, where each subproblem has a corresponding reward function f_i.

Step 2: Suppose we know the optimal policy of the original problem, which always exists due to Theorem 1. Then, for each subproblem in the modified problem, if we were to fix the other variables (that are not in the subproblem) according to the optimal policy of the original problem, we can then compute the decomposed optimal Q-values Q_i^*. Additionally, Theorem 1 guarantees the existence of these decomposed optimal Q-values.

Step 3: Next, we show that the global optimal Q-values (sum of the decomposed optimal Q-values) of the modified MD-DCOP is the same as the global optimal Q-values of the original MD-DCOP.

Step 4: Finally, we show how to decompose the global optimally Q-values, which concludes the proof.

Step 1: Consider a modified MD-DCOP where the transition probabilities are the same as the original MD-DCOP, but the reward functions for each joint state $\mathbf{s}$ and joint value $\mathbf{d}_i$ are defined as follows:

$$\bar{f}_i(\mathbf{s}, \mathbf{d}_i) = \begin{cases} f_i(\mathbf{s}_i, \mathbf{d}_i \mid \mathbf{s}_i \in \mathbf{s}) & \text{if } \mathbf{d}_i \in \text{argmax}_{\mathbf{d} \in \mathbf{D}}\, Q^*(\mathbf{s}, \mathbf{d}) \\ -C & \text{otherwise} \end{cases} \tag{4}$$

where C is a very large constant.

Step 2: We now show the existence of the decomposed Q-values $\bar{Q}_i^*(\mathbf{s}, \mathbf{d}_i)$ for each reward function f_i. First, set the policy of every other variable that is not in the subproblem defined by reward function f_i to their respective optimal policy in the original MD-DCOP. Also set the transition probabilities $\bar{P}_i(\mathbf{s}', \mathbf{s}, \mathbf{d}_i)$ according to the premise of Theorem 2 and set the reward functions $\bar{f}_i(\mathbf{s}, \mathbf{d}_i)$ according to Equation 4.

According to Theorem 1, there exists a decomposed Q-value $\bar{Q}_i^*$ for this subproblem such that

$$\begin{aligned} \bar{Q}_i^*(\mathbf{s}, \mathbf{d}_i) + \rho_i^* &= \bar{f}_i(\mathbf{s}, \mathbf{d}_i) \\ &+ \sum_{\mathbf{s}'} \bar{P}_i(\mathbf{s}', \mathbf{s}, \mathbf{d}_i)\, \bar{Q}_i^*\big(\mathbf{s}', \mathbf{d}'_i \mid \mathbf{d}'_i \in \underset{\mathbf{d}' \in \mathbf{D}}{\text{argmax}}\, Q^*(\mathbf{s}', \mathbf{d}')\big) \end{aligned} \tag{5}$$

where ρ_i^* corresponds to the local average reward of the subproblem, as shown in Lemma 1.

Step 3: Then, for the globally optimal joint value $\mathbf{d}^* = \text{argmax}_{\mathbf{d} \in \mathbf{D}}\, Q^*(\mathbf{s}, \mathbf{d})$, let $\mathbf{d}_i^*$ to denote the local joint value in $\mathbf{d}^*$, $\bar{Q}^*(\mathbf{s}, \mathbf{d}^*) = \sum_i \bar{Q}_i^*(\mathbf{s}, \mathbf{d}_i^*)$, and $\bar{\mathbf{F}}(\mathbf{s}, \mathbf{d}^*) = \sum_i \bar{f}_i(\mathbf{s}, \mathbf{d}_i^*)$. Summing over all subproblems, we get

$$\begin{aligned} &\bar{Q}^*(\mathbf{s}, \mathbf{d}^*) + \rho^* \\ &\quad = \sum_i \Big[\bar{Q}_i^*(\mathbf{s}, \mathbf{d}_i^*) + \rho_i^*\Big] \\ &\quad = \sum_i \Big[\bar{f}_i(\mathbf{s}, \mathbf{d}_i^*) \\ &\qquad + \sum_{\mathbf{s}'} \bar{P}_i(\mathbf{s}', \mathbf{s}, \mathbf{d}_i^*)\, \bar{Q}_i^*\big(\mathbf{s}', \mathbf{d}'_i \mid \mathbf{d}'_i \in \underset{\mathbf{d}' \in \mathbf{D}}{\text{argmax}}\, Q^*(\mathbf{s}', \mathbf{d}')\big)\Big] \\ &\quad = \sum_i \bar{f}_i(\mathbf{s}, \mathbf{d}_i^*) \\ &\qquad + \sum_{i, \mathbf{s}'} \Big[\bar{P}_i(\mathbf{s}', \mathbf{s}, \mathbf{d}_i^*)\, \bar{Q}_i^*\big(\mathbf{s}', \mathbf{d}'_i \mid \mathbf{d}'_i \in \underset{\mathbf{d}' \in \mathbf{D}}{\text{argmax}}\, Q^*(\mathbf{s}', \mathbf{d}')\big)\Big] \\ &\quad = \bar{\mathbf{F}}(\mathbf{s}, \mathbf{d}^*) \\ &\qquad + \sum_{\mathbf{s}'} \Big[P(\mathbf{s}', \mathbf{s}, \mathbf{d}^*) \sum_i \bar{Q}_i^*\big(\mathbf{s}', \mathbf{d}'_i \mid \mathbf{d}'_i \in \underset{\mathbf{d}' \in \mathbf{D}}{\text{argmax}}\, Q^*(\mathbf{s}', \mathbf{d}')\big)\Big] \end{aligned}$$

$$= \bar{\mathbf{F}}(\mathbf{s}, \mathbf{d}^*) + \sum_{\mathbf{s}'} \Big[P(\mathbf{s}', \mathbf{s}, \mathbf{d}^*) \max_{\mathbf{d}' \in \mathbf{D}} \bar{Q}^*(\mathbf{s}', \mathbf{d}') \Big] \tag{6}$$

This equation is in the form of Equation 1, which characterizes $\bar{Q}^*(\mathbf{s}, \mathbf{d}^*)$ as a solution to the modified problem. Additionally, one can also show that the Q-value $Q^*(\mathbf{s}, \mathbf{d}^*)$ of the original problem is also a solution to the modified problem using the same process as in Equation 6. Since $V^*(\mathbf{s})$ is unique according to Theorem 1, it must be the case that $V^*(\mathbf{s}) = Q^*(\mathbf{s}, \mathbf{d}^*) = \bar{Q}^*(\mathbf{s}, \mathbf{d}^*)$.

Step 4: Now, let's define the decomposed Q- and V-values for $\mathbf{d}_i^*$ as follows:

$$\begin{aligned} Q_i^*(\mathbf{s}, \mathbf{d}_i^*) + \rho_i^* &= f_i(\mathbf{s}_i, \mathbf{d}_i^*) \\ &+ \sum_{\mathbf{s}'} \bar{P}_i(\mathbf{s}', \mathbf{s}, \mathbf{d}_i^*) \, \bar{Q}_i^*(\mathbf{s}', \mathbf{d}_i' \mid \mathbf{d}_i' \in \operatorname*{argmax}_{\mathbf{d}' \in \mathbf{D}} Q^*(\mathbf{s}', \mathbf{d}')) \qquad (7) \\ V_i^*(\mathbf{s}) &= Q_i^*(\mathbf{s}, \mathbf{d}_i^*) \qquad (8) \end{aligned}$$

We now show that $Q_i^*(\mathbf{s}, \mathbf{d}_i^*) = \bar{Q}_i^*(\mathbf{s}, \mathbf{d}_i^*)$:

$$\begin{aligned} Q_i^*(\mathbf{s}, \mathbf{d}_i^*) &= f_i(\mathbf{s}_i, \mathbf{d}_i^*) - \rho_i^* \\ &\quad + \sum_{\mathbf{s}'} \bar{P}_i(\mathbf{s}', \mathbf{s}, \mathbf{d}_i^*) \, \bar{Q}_i^*(\mathbf{s}', \mathbf{d}_i' \mid \mathbf{d}_i' \in \operatorname*{argmax}_{\mathbf{d}' \in \mathbf{D}} Q^*(\mathbf{s}', \mathbf{d}')) \\ &= \bar{f}_i(\mathbf{s}_i, \mathbf{d}_i^*) - \rho_i^* \\ &\quad + \sum_{\mathbf{s}'} \bar{P}_i(\mathbf{s}', \mathbf{s}, \mathbf{d}_i^*) \, \bar{Q}_i^*(\mathbf{s}', \mathbf{d}_i' \mid \mathbf{d}_i' \in \operatorname*{argmax}_{\mathbf{d}' \in \mathbf{D}} Q^*(\mathbf{s}', \mathbf{d}')) \\ &= \bar{Q}_i^*(\mathbf{s}, \mathbf{d}_i^*) \qquad (9) \end{aligned}$$

Therefore,

$$\begin{aligned} & V_i^*(\mathbf{s}) + \rho_i^* \\ &\quad = Q_i^*(\mathbf{s}, \mathbf{d}_i^*) + \rho_i^* \\ &\quad = f_i(\mathbf{s}_i, \mathbf{d}_i^*) + \sum_{\mathbf{s}'} \bar{P}_i(\mathbf{s}', \mathbf{s}, \mathbf{d}_i^*) \, \bar{Q}_i^*(\mathbf{s}', \mathbf{d}_i' \mid \mathbf{d}_i' \in \operatorname*{argmax}_{\mathbf{d}' \in \mathbf{D}} Q^*(\mathbf{s}', \mathbf{d}')) \\ &\quad = f_i(\mathbf{s}_i, \mathbf{d}_i^*) + \sum_{\mathbf{s}'} \bar{P}_i(\mathbf{s}', \mathbf{s}, \mathbf{d}_i^*) \, Q_i^*(\mathbf{s}', \mathbf{d}_i' \mid \mathbf{d}_i' \in \operatorname*{argmax}_{\mathbf{d}' \in \mathbf{D}} Q^*(\mathbf{s}', \mathbf{d}')) \\ &\quad = f_i(\mathbf{s}_i, \mathbf{d}_i^*) + \sum_{\mathbf{s}'} \bar{P}_i(\mathbf{s}', \mathbf{s}, \mathbf{d}_i^*) \, V_i^*(\mathbf{s}') \qquad (10) \end{aligned}$$

Finally, we now show that the V-value $V^*(\mathbf{s})$ is a sum of its decomposed components $V_i^*(\mathbf{s}, \mathbf{d}_i^*)$:

$$\begin{aligned} V^*(\mathbf{s}) &= \bar{Q}^*(\mathbf{s}, \mathbf{d}^*) \\ &= \sum_i \bar{Q}_i^*(\mathbf{s}, \mathbf{d}_i^*) \\ &= \sum_i Q_i^*(\mathbf{s}, \mathbf{d}_i^*) \\ &= \sum_i V_i^*(\mathbf{s}, \mathbf{d}_i^*) \qquad (11) \end{aligned}$$

which concludes the proof. ■

As a result of the existence of the local value $V_i^*(\mathbf{s})$, we can derive the convergence proof of our distributed RVI Q-learning algorithm:

Theorem 3. *The Distributed RVI Q-learning algorithm converges to an optimal solution.*

PROOF SKETCH OF THEOREM 3: Let $\mathbf{d}$ denote the global joint value taken by all the variables in the current iteration, and $\mathbf{d}_i$ denote the local joint value taken by variables in the scope of reward function f_i. Additionally, let $\mathbf{s}$ denote the current global state, and $\mathbf{s}'$ denote the next global state as a result of taking the joint value $\mathbf{d}$.

Now, let H_i be the mapping defined by

$$\begin{aligned}(H_i Q_i)(\mathbf{s}, \mathbf{d}_i) &= f_i(\mathbf{s}, \mathbf{d}_i) \\ &\quad + \sum_{\mathbf{s}'_i} \bar{P}_i(\mathbf{s}', \mathbf{s}, \mathbf{d}_i) Q(\mathbf{s}', \mathbf{d}'_i \mid \mathbf{d}'_i \in \operatorname*{argmax}_{\mathbf{d}' \in \mathbf{D}} Q(\mathbf{s}', \mathbf{d}')) - \rho_i \\ &= f_i(\mathbf{s}, \mathbf{d}_i) + \sum_{\mathbf{s}'_i} \bar{P}_i(\mathbf{s}', \mathbf{s}, \mathbf{d}_i) V(\mathbf{s}') - \rho_i\end{aligned}$$

with $V(\mathbf{s}') = Q(\mathbf{s}', \mathbf{d}'_i \mid \mathbf{d}'_i \in \operatorname{argmax}_{\mathbf{d}' \in \mathbf{D}} Q(\mathbf{s}', \mathbf{d}'))$. Since H_i is non-expansive, that is,

$$\|H_i Q_i - H_i Q'_i\|_\infty \leq \|Q_i - Q'_i\|_\infty$$

and the corresponding ODE of $H_i Q_i$

$$\dot{Q}_i(t) = H_i(Q_i(t)) - Q(t)$$

has at least one solution according to Theorem 2, Q_i thus converges to the optimal value Q_i^* using the result in (Abounadi, Bertsekas, and Borkar, 2001). ■

Bibliography

Abounadi, J.; Bertsekas, D.; and Borkar, V. 2001. Learning algorithms for markov decision processes with average cost. *SIAM Journal on Control and Optimization* 40(3):681–698.

Puterman, M. 2005. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc.