Shapley Q-Value: A Local Reward Approach to Solve Global Reward Games

Jianhong Wang,[1,2*] Yuan Zhang,[3*] Tae-Kyun Kim,[2] Yunjie Gu[1]
[1]Control and Power Research Group, Imperial College London, UK
[2]Imperial Computer Vision and Learning Lab, Imperial College London, UK
[3]Laiye Network Technology Co.Ltd., China

Abstract

Cooperative game is a critical research area in the multi-agent reinforcement learning (MARL). Global reward game is a subclass of cooperative games, where all agents aim to maximize the global reward. Credit assignment is an important problem studied in the global reward game. Most of previous works stood by the view of non-cooperative-game theoretical framework with the shared reward approach, i.e., each agent being assigned a shared global reward directly. This, however, may give each agent an inaccurate reward on its contribution to the group, which could cause inefficient learning. To deal with this problem, we i) introduce a cooperative-game theoretical framework called extended convex game (ECG) that is a superset of global reward game, and ii) propose a local reward approach called Shapley Q-value. Shapley Q-value is able to distribute the global reward, reflecting each agent's own contribution in contrast to the shared reward approach. Moreover, we derive an MARL algorithm called Shapley Q-value deep deterministic policy gradient (SQDDPG), using Shapley Q-value as the critic for each agent. We evaluate SQDDPG on Cooperative Navigation, Prey-and-Predator and Traffic Junction, compared with the state-of-the-art algorithms, e.g., MADDPG, COMA, Independent DDPG and Independent A2C. In the experiments, SQDDPG shows a significant improvement on the convergence rate. Finally, we plot Shapley Q-value and validate the property of fair credit assignment.

1 Introduction

Cooperative game is a critical research area in multi-agent reinforcement learning (MARL). Many real-life tasks can be modeled as cooperative games, e.g., the coordination of autonomous vehicles (Keviczky et al. 2007), autonomous distributed logistics (Schuldt 2012) and search-and-rescue robots (Koes, Nourbakhsh, and Sycara 2006; Ramchurn et al. 2010). Global reward game (Chang, Ho, and Kaelbling 2004) is a subclass of cooperative games where agents aim to maximize the global reward. In this game, credit assignment is an important problem, which targets on finding a method to distribute the global reward. There are two categories of approaches to solve out this problem, namely shared reward approach (also known as shared reward game or fully cooperative game) (Sukhbaatar, szlam, and Fergus 2016; Omidshafiei et al. 2018; Kim et al. 2019) and local reward approach (Panait and Luke 2005). The shared reward approach directly assigns the global reward to all agents. The local reward approach, on the other hand, distributes the global reward according to each agent's contribution, and turns out to have superior performances in many tasks (Foerster et al. 2018; Nguyen, Kumar, and Lau 2018).

Whatever approach is adopted, a remaining open question is whether there is an underlying theory to explain credit assignment. Conventionally, a global reward game is built upon non-cooperative game theory, which primarily aims to find Nash equilibrium as the stable solution (Osborne and Rubinstein 1994; Basar and Olsder 1999). This formulation can be extended to a dynamic environment with infinite horizons via stochastic game (Shapley 1953a). However, Nash equilibrium focuses on the individual reward and has no explicit incentives for cooperation. As a result, the shared reward function has to be applied to force cooperation, which could be used as a possible explanation to the shared reward approach, but not the local reward approach.

In our work, we introduce and investigate the cooperative game theory (or the coalitional game theory) (Chalkiadakis, Elkind, and Wooldridge 2011) in which local reward approach becomes rationalized. In cooperative game theory, the objective is dividing coalitions and binding agreements among agents who belong to the same coalition. We focus on convex game (CG) which is a typical game in cooperative game theory featuring the existence of a stable coalition structure with an efficient payoff distribution scheme (i.e., a local reward approach) called core. This payoff distribution is equivalent to the credit assignment, thereby the core rationalizes and well explains the local reward approach (Peleg and Sudhölter 2007).

Referring to the previous works (Suijs and Borm 1999; Chalkiadakis and Boutilier 2004), we extend CG to an infinite-horizon scenario, namely extended CG (ECG). In addition, we show that a global reward game is equivalent to an ECG with the grand coalition and an efficient payoff

*The first two authors have equal contributions. Correspondence to jianhong.wang16@imperial.ac.uk.

distribution scheme. Furthermore, we propose Shapley Q-value, extending Shapley value (i.e., an efficient payoff distribution scheme) for the credit assignment in an ECG with the grand coalition. Therefore, it results in a conclusion that Shapley Q-value is able to work in a global reward game. Finally, we derive an algorithm called Shapley Q-value deep deterministic policy gradient (SQDDPG) according to the actor-critic framework (Konda and Tsitsiklis 2000) to learn decentralized policies with centralized critics (i.e., Shapley Q-values). SQDDPG is evaluated on the environments such as Cooperative Navigation, Prey-and-Predator (Lowe et al. 2017), and Traffic Junction (Sukhbaatar, szlam, and Fergus 2016), compared with the state-of-the-art baselines, e.g., MADDPG (Lowe et al. 2017), COMA (Foerster et al. 2018), Independent DDPG (Lillicrap et al. 2015) and Independent A2C (Sutton and Barto 2018).

2 Related Work

Multi-agent Learning Multi-agent learning refers to a category of methods that tackle the games with multiple agents such as cooperative games. Among these methods, we only focus on using reinforcement learning to deal with a cooperative game, which is called multi-agent reinforcement learning (MARL). Incredible progresses have recently been made on MARL. Some researchers (Sukhbaatar, szlam, and Fergus 2016; Kim et al. 2019; Das et al. 2018) focus on distributed executions, which allow communications among agents. Others (Chang, Ho, and Kaelbling 2004; Foerster et al. 2018; Nguyen, Kumar, and Lau 2018; Lowe et al. 2017; Iqbal and Sha 2018) consider decentralized executions, where no communication is permitted during the execution. Nevertheless, all of them study on centralized critics, which means information can be shared on the value function during training. In our work, we pay our attention to the decentralized execution and the centralized critic.

Cooperative Game As opposed to competing with others, agents in a cooperative game aim to cooperate to solve a joint task or maximize the global payoff (also known as the global reward) (Chalkiadakis, Elkind, and Wooldridge 2011). Shapley (1953a) proposed a non-cooperative game theoretical framework called stochastic game, which models the dynamics of multiple agents in zero-sum game with infinite horizons. Hu and Wellman (1998) introduced a general-sum stochastic game theoretical framework, which generalises the zero-sum game. To force cooperation under this framework, potential function (Monderer and Shapley 1996) was applied such that each agent shares the same objective, namely global reward game (Chang, Ho, and Kaelbling 2004). In this paper, we use cooperative game theory whereas the existing cooperative game framework are built under the non-cooperative game theory. Our framework gives a new view on the global reward game and well explains the reason why credit assignment is important. We show that the global reward game is a subclass of our framework if we interpret that the agents in a global reward game forms a grand coalition (i.e., the group including the whole agents). Under our framework, it is more rational to use a local reward approach to distribute the global reward.

Credit Assignment Credit assignment is a significant problem that has been studied in cooperative games for a long period. There are two sorts of credit assignment approaches, i.e., shared reward approach and local reward approach. The shared reward approach directly assigns each agent the global reward (Sukhbaatar, szlam, and Fergus 2016; Kim et al. 2019; Das et al. 2018; Lowe et al. 2017). We show that this is actually equivalent to distributing the global reward equally to individual agents. The global reward game with this credit assignment scheme is also called shared reward game (also known as fully cooperative game) (Panait and Luke 2005). However, Wolpert and Tumer (2002) claimed that the shared reward approach does not give each agent the accurate contribution. Thus, it may not perform well in difficult problems. This motivates the study on the local reward approach, which can distribute the global reward to agents according to their contributions. The existing question is how to quantify the contributions. To investigate the answer to this question, Chang, Ho, and Kaelbling (2004) attempted using Kalman filter to infer the contribution of each agent. Recently, Foerster et al. (2018) and Nguyen, Kumar, and Lau (2018) modelled the marginal contributions inspired by the reward difference (Wolpert and Tumer 2002). Under our proposed framework (i.e., ECG), we propose a new method called Shapley Q-value to learn a local reward. This method is extended from the Shapley value (Shapley 1953b). It is theoretically guaranteed to distribute the global reward fairly. Although Shapley value can be regarded as the expectation of the marginal contributions, it is different from the previous works (Foerster et al. 2018; Nguyen, Kumar, and Lau 2018): it considers all possible orders of agents to form a grand coalition, which has not been mentioned in these works.

3 Preliminaries

3.1 Convex Game

Convex game (CG) is a typical transferable utility game in the cooperative game theory. The definitions below are referred to the textbook (Chalkiadakis, Elkind, and Wooldridge 2011). A CG is formally represented as $\Gamma = \langle \mathcal{N}, v \rangle$, where $\mathcal{N}$ is the set of agents and v is the value function to measure the profits earned by a coalition (i.e., a group). $\mathcal{N}$ itself is called the grand coalition. The value function $v : 2^{\mathcal{N}} \rightarrow \mathbb{R}$ is a mapping from a coalition $\mathcal{C} \subseteq \mathcal{N}$ to a real number $v(\mathcal{C})$. In a CG, its value function satisfies two properties, i.e., 1) $v(\mathcal{C} \cup \mathcal{D}) \geq v(\mathcal{C}) + v(\mathcal{D}), \forall \mathcal{C}, \mathcal{D} \subset \mathcal{N}, \mathcal{C} \cap \mathcal{D} = \phi$; 2) the coalitions are independent. The solution of a CG is a tuple $(\mathcal{CS}, \mathbf{x})$, where $\mathcal{CS} = \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_m\}$ is a coalition structure and $\mathbf{x} = (\mathrm{x}_i)_{i \in \mathcal{N}}$ indicates the payoff (i.e., the local reward) distributed to each agent, which satisfies two conditions, i.e., 1) $\mathrm{x}_i \geq 0, \forall i \in \mathcal{N}$; 2) $\mathbf{x}(\mathcal{C}) \leq v(\mathcal{C}), \forall \mathcal{C} \subseteq \mathcal{CS}$, where $\mathbf{x}(\mathcal{C}) = \sum_{i \in \mathcal{C}} \mathrm{x}_i$. $\mathcal{CS}_{\mathcal{N}}$ denotes the set of all possible coalition structures. The *core* is a stable solution set of a CG, which can be defined mathematically as $\texttt{core}(\Gamma) = \{(\mathcal{C}, \mathbf{x}) | \mathbf{x}(\mathcal{C}) \geq v(\mathcal{C}), \forall \mathcal{C} \subseteq \mathcal{N}\}$. The core of a CG ensures a reasonable payoff distribution and inspires our work on credit assignment in MARL.

3.2 Shapley Value

Shapley value (Shapley 1953b) is one of the most popular methods to solve the payoff distribution problem for a grand coalition (Fatima, Wooldridge, and Jennings 2008; Michalak et al. 2013; Faigle and Kern 1992). Given a cooperative game $\Gamma = (\mathcal{N}, v)$, for any $\mathcal{C} \subseteq \mathcal{N} \backslash \{i\}$ let $\delta_i(\mathcal{C}) = v(\mathcal{C} \cup \{i\}) - v(\mathcal{C})$ be a marginal contribution, then the Shapley value of each agent i can be written as:

$$\mathrm{Sh}_i(\Gamma) = \sum_{\mathcal{C} \subseteq \mathcal{N} \backslash \{i\}} \frac{|\mathcal{C}|!(|\mathcal{N}| - |\mathcal{C}| - 1)!}{|\mathcal{N}|!} \cdot \delta_i(\mathcal{C}). \quad (1)$$

Literally, Shapley value takes the average of marginal contributions of all possible coalitions, so that it satisfies: 1) efficiency: $\mathbf{x}(\mathcal{N}) = v(\mathcal{N})$; 2) fairness: if an agent i has no contribution, then $x_i = 0$; if i-th and j-th agents have the same contribution, then $x_i = x_j$ (Chalkiadakis, Elkind, and Wooldridge 2011). As we can see from Eq.1, if we calculate Shapley value for an agent, we have to consider $2^{|\mathcal{N}|} - 1$ possible coalitions that the agent could join in to form a grand coalition, which causes the computational catastrophe. To mitigate this issue, we propose an approximation in the scenarios with infinite horizons called approximate Shapley Q-value which is introduced in the next section.

3.3 Multi-agent Actor-Critic

Different from the value based method, i.e., Q-learning (Watkins and Dayan 1992), policy gradient (Williams 1992) directly learns the policy by maximizing $J(\theta) = \mathbb{E}_{s \sim \rho^\pi, a \sim \pi_\theta}[r(s,a)]$, where $r(s,a)$ is the reward of an arbitrary state-action pair. Since the gradient of $J(\theta)$ w.r.t. θ cannot be directly calculated, policy gradient theorem (Sutton and Barto 2018) is used to approximate the gradient such that $\nabla_\theta J(\theta) = \mathbb{E}_{s \sim \rho^\pi, a \sim \pi_\theta}[Q^\pi(s,a) \nabla_\theta \log \pi_\theta(a|s)]$. In the actor-critic framework (Konda and Tsitsiklis 2000) (that is derived from the policy gradient theorem), $\pi_\theta(a|s)$ is called actor and $Q^\pi(s,a)$ is called critic. Additionally, $Q^\pi(s,a) = \mathbb{E}_\pi[\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) | s_1 = s, a_1 = a]$. Extending to the multi-agent scenarios, the gradient of each agent i can be represented as $\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s \sim \rho^\pi, \mathbf{a} \sim \pi_\theta}[Q_i^\pi(s, a_i) \nabla_{\theta_i} \log \pi_{\theta_i}^i(a_i|s)]$. $Q_i^\pi(s, a_i)$ can be regarded as the estimation of the contribution of each agent i. If the deterministic policy (Silver et al. 2014) needs to be learned in MARL problems, we can reformulate the approximate gradient of each agent as $\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s \sim \rho^\mu}[\nabla_{\theta_i} \mu_{\theta_i}^i(s) \nabla_{a_i} Q_i^\mu(s, a_i)|_{a_i = \mu_{\theta_i}^i(s)}]$. In this work, we applies this formulation to learn the deterministic policy for each agent.

4 Our Work

In this section, we (i) extend convex game (CG) with the infinite horizons and decisions, namely extended convex game (ECG) and show that a global reward game is equivalent to an ECG with the grand coalition and an efficient distribution scheme, (ii) show that the shared reward approach is an efficient distribution scheme in an ECG with the grand coalition, (iii) propose Shapley Q-value by extending and approximating Shapley value to distribute the credits in a global reward game, because it can accelerate the convergence rate compared with shared reward approach, and (iv) derive an MARL algorithm called Shapley Q-value deep deterministic policy gradient (SQDDPG), using the Shapley Q-value as each agent's critic.

4.1 Extended Convex Game

Referring to the previous works (Suijs and Borm 1999; Chalkiadakis and Boutilier 2004), we extend the CG to the scenarios with infinite horizons and decisions, named as extended CG (ECG). The set of joint actions of agents is defined as $\mathcal{A} = \times_{i \in \mathcal{N}} \mathcal{A}_i$, where $\mathcal{A}_i$ is the feasible action set for each agent i. $\mathcal{S}$ is the set of possible states in the environment. The dynamics of the environment are defined as $Pr(s'|s, \mathbf{a})$, where $s, s' \in \mathcal{S}$ and $\mathbf{a} \in \mathcal{A}$. Inspired by Nash (1953), we construct the ECG by two stages. In the stage 1, an oracle arranges the coalition structure and contracts the cooperation agreements, i.e., the credit assigned to an agent for his optimal long-term contribution if he joins in some coalition. We assume that this oracle can observe the whole environment and be familiar with each agent's feature. In the stage 2, after joining in the allocated coalition, each agent will further make a decision by $\pi_i(a_i|s)$ to maximize the social value of its coalition, so that the optimal social value of each coalition and individual credit assignment can be achieved, where $a_i \in \mathcal{A}_i$ and $s \in \mathcal{S}$. Mathematically, the optimal value of a coalition $\mathcal{C} \in \mathcal{CS}$ can be written as $\max_{\pi_\mathcal{C}} v^{\pi_\mathcal{C}}(\mathcal{C}) = \mathbb{E}_{\pi_\mathcal{C}}[\sum_{t=1}^{\infty} \gamma^{t-1} r_t(\mathcal{C})]$; $\pi_\mathcal{C} = \times_{i \in \mathcal{C}} \pi_i$; $r_t(\mathcal{C})$ is the reward gained by coalition $\mathcal{C}$ at each time step. According to the property (1) of the CG aforementioned, the formula $\max_{\pi_{\mathcal{C} \cup \mathcal{D}}} v^{\pi_{\mathcal{C} \cup \mathcal{D}}}(\mathcal{C} \cup \mathcal{D}) \geq \max_{\pi_\mathcal{C}} v^{\pi_\mathcal{C}}(\mathcal{C}) + \max_{\pi_\mathcal{D}} v^{\pi_\mathcal{D}}(\mathcal{D}), \forall \mathcal{C}, \mathcal{D} \subset \mathcal{N}, \mathcal{C} \cap \mathcal{D} = \phi$ holds. In this paper, we denote the joint policy of the whole agents as $\pi = \times_{i \in \mathcal{N}} \pi_i(a_i|s)$ and assume that each agent can observe the global state.

Lemma 1 (Shapley 1971; Chalkiadakis, Elkind, and Wooldridge 2011)**.** *1) Every convex game has a non-empty core. 2) If a solution $(\mathcal{CS}, \boldsymbol{x})$ is in the core of a characteristic function game $\langle \mathcal{N}, v \rangle$ and the payoff distribution scheme is efficient, then $v(\mathcal{CS}) \geq v(\mathcal{CS}')$ for every coalition structure $\mathcal{CS}' \in \mathcal{CS}_\mathcal{N}$.*

Theorem 1. *With the efficient payoff distribution scheme, for an extended convex game (ECG), one solution in the core must exist with the grand coalition and the objective is $\max_\pi v^\pi(\{\mathcal{N}\})$, which can lead to the maximal social welfare, i.e., $\max_\pi v^\pi(\{\mathcal{N}\}) \geq \max_\pi v^\pi(\mathcal{CS}')$ for every coalition structure $\mathcal{CS}' \in \mathcal{CS}_\mathcal{N}$.*

Proof. This theorem is proved based on Lemma 1. See Appendix for the complete proof. □

Corollary 1. *For an extended convex game (ECG) with the grand coalition, Shapley value must be in the core.*

Proof. Since Shapley value must be in the core for a CG with the grand coalition (Shapley 1971) and ESG still conserves the property of CG, this statement holds. □

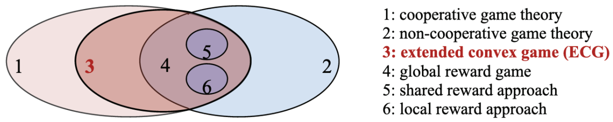

Figure 1: The relationship between the concepts mentioned in this paper.

As seen from Theorem 1, with an appropriate efficient payoff distribution scheme, an ECG with the grand coalition is actually equivalent to a global reward game. Both of them aim to maximize the global value (i.e., the global reward). Here, we assume that the agents in a global reward game are regarded as the grand coalition. Consequently, the local reward approach in ECG is feasible in the global reward game.

4.2 Looking into the Shared Reward Approach by the View of ECG

Shared reward approach assigns each agent the global reward directly in the global reward game (Panait and Luke 2005). Each agent unilaterally maximizes this global reward to seek its optimal policy such that

$$\max_{\pi_i} v^{\pi}(\{\mathcal{N}\}) = \max_{\pi_i} \mathbb{E}_{\pi}[\, \sum_{t=1}^{\infty} r_t(\mathcal{N}) \,], \tag{2}$$

where $r_t(\mathcal{N})$ is the global reward and $\pi = \times_{i \in \mathcal{N}} \pi_i$. If $v^{\pi}(\{\mathcal{N}\})$ is multiplied by a normalizing factor, i.e., $1/|\mathcal{N}|$, then the objective of the new optimization problem for each agent i is equivalent to Eq.2. We can express it mathematically as

$$\begin{aligned} \max_{\pi_i} v^{\pi}(\{\mathcal{N}\}) &= \max_{\pi_i} \sum_{i \in \mathcal{N}} \mathrm{x}_i = \max_{\pi_i} \mathrm{x}_i \\ &= \max_{\pi_i} \frac{1}{|\mathcal{N}|} \cdot v^{\pi}(\{\mathcal{N}\}). \end{aligned} \tag{3}$$

Then, the credit assigned to each agent in the shared reward approach is actually $x_i = v^{\pi}(\{\mathcal{N}\})/|\mathcal{N}|$, and the sum of the whole agents' credits is equal to the global reward. It suffices the condition of efficient payoff distribution scheme. Therefore, we show that the shared reward approach is an efficient payoff distribution scheme in an ECG with the grand coalition (i.e., a global reward game). Nevertheless, from the view of ECG, the shared reward approach is not guaranteed to find the optimal solution. By Corollary 1 and Theorem 1, we know that Shapley value has to lie in the core and theoretically guarantees the convergence to the maximal global value. This is one of the reasons why we are interested in this local reward approach. To clarify the concepts we mentioned before, we draw a Venn diagram shown as Fig.1.

4.3 Shapley Q-value

Although the shared reward approach successfully solves out a global reward game in practice, it has been shown that local reward approach gives the faster convergence rate (Balch 1997; 1999). For the two aforementioned reasons, we use the Shapley value (i.e., a local reward approach) for the credit assignment to each agent. Because $v^{\pi_{\mathcal{C}'}}(\mathcal{C}')$ represents the global reward earned by coalition $\mathcal{C}'$ in an ECG, we can model it as a Q-value, where s represents the state and $\mathbf{a}_{\mathcal{C}'} = (a_i)_{i \in \mathcal{C}'}$. According to Eq.1, the Shapley Q-value of each agent can be written as

$$\Phi_i(\mathcal{C}) = Q^{\pi_{\mathcal{C} \cup \{i\}}}(s, \mathbf{a}_{\mathcal{C} \cup \{i\}}) - Q^{\pi_{\mathcal{C}}}(s, \mathbf{a}_{C}), \tag{4}$$

$$Q^{\Phi_i}(s, a_i) = \sum_{\mathcal{C} \subseteq \mathcal{N} \setminus \{i\}} \frac{|\mathcal{C}|!(|\mathcal{N}| - |\mathcal{C}| - 1)!}{|\mathcal{N}|!} \cdot \Phi_i(\mathcal{C}). \tag{5}$$

Approximate Marginal Contribution As seen from Eq.4, it is difficult and unstable to learn two Q-value functions (where one is for representing the Q-value of $\mathcal{C} \cup \{i\}$ and the other is for representing the Q-value of $\mathcal{C}$) to estimate the marginal contributions (i.e., making the difference between two Q-values) for different coalitions. To mitigate this problem, we propose a method called approximate marginal contribution (AMC) to directly estimate the marginal contribution of each coalition (i.e., $\Phi_i(\mathcal{C})$).

In cooperative game theory, each agent is assumed to join in the grand coalition sequentially. $|\mathcal{C}|!(|\mathcal{N}| - |\mathcal{C}| - 1)!/|\mathcal{N}|!$ in Eq.1, denoted as $Pr(\mathcal{C}|\mathcal{N} \setminus \{i\})$, is interpreted as that an agent randomly joins in an existing coalition $\mathcal{C}$ (which could be empty) to form the complete grand coalition with the subsequent agents (Chalkiadakis, Elkind, and Wooldridge 2011). According to this interpretation, we model a function to approximate the marginal contribution directly such that

$$\hat{\Phi}_i(s, \mathbf{a}_{\mathcal{C} \cup \{i\}}) : \mathcal{S} \times \mathcal{A}_{\mathcal{C} \cup \{i\}} \mapsto \mathbb{R}, \tag{6}$$

where $\mathcal{S}$ is the state space; $\mathcal{C}$ is the ordered coalition that agent i would like to join in; $\mathcal{A}_{\mathcal{C} \cup \{i\}} = (\mathcal{A}_j)_{j \in \mathcal{C} \cup \{i\}}$, and the actions are ordered. For example, if the order of a coalition is $(0, 2)$ (where 0 and 2 are two agents), then $\mathbf{a}_{\mathcal{C} \cup \{1\}} = (a_0, a_2, a_1)$. By such a formulation, we believe that the property of the marginal contribution (i.e., mapping from every possible combination of coalition $\mathcal{C}$ and agent i to a numerical value) can be maintained. Hence, AMC is reasonable to replace the exact marginal contribution. In practice, we represent $\mathbf{a}_{\mathcal{C} \cup \{i\}}$ by the concatenation of each agent's action vector. To keep the input size of $\hat{\Phi}_i(s, \mathbf{a}_{\mathcal{C} \cup \{i\}})$ constant in different cases, we fix the actions as the concatenation of all agents' actions and mask the actions of irrelevant agents (i.e., the agents who do not exist in the coalition) with zeros.

Approximate Shapley Q-value Followed by the interpretation above, Shapley Q-value can be rewritten as

$$Q^{\Phi_i}(s, a_i) = \mathbb{E}_{\mathcal{C} \sim Pr(\mathcal{C}|\mathcal{N} \setminus \{i\})}[\, \Phi_i(\mathcal{C}) \,]. \tag{7}$$

To enable Eq.7 to be tractable in realization, we can sample $Q^{\Phi_i}(s, a_i)$ here. Also, substituting AMC in Eq.6 for the marginal contribution in Eq.4, we can approximate Shapley Q-value such that

$$Q^{\Phi_i}(s, a_i) \approx \frac{1}{M} \sum_{k=1}^{M} \hat{\Phi}_i(s, \mathbf{a}_{\mathcal{C}_k \cup \{i\}}), \;\; \forall \mathcal{C}_k \sim Pr(\mathcal{C}|\mathcal{N} \setminus \{i\}). \tag{8}$$

4.4 Shapley Q-value Deep Deterministic Policy Gradient

In an ECG with the grand coalition, each agent only needs to maximize its own Shapley value (i.e., $Q^{\Phi_i}(s, a_i)$) so that $\max_{\pi} v^{\pi}(\mathcal{N})$ can be achieved such that

$$\begin{aligned}\max_{\pi} v^{\pi}(\mathcal{N}) = \max_{\pi} \sum_{i \in \mathcal{N}} \mathrm{x}_i &= \sum_{i \in \mathcal{N}} \max_{\pi_i} \mathrm{x}_i \\ &= \sum_{i \in \mathcal{N}} \max_{\pi_i} Q^{\Phi_i}(s, a_i).\end{aligned} \tag{9}$$

Therefore, if we can show that $\max_{\pi_i} Q^{\Phi_i}(s, a_i)$ for each agent i is approached, then we show that the maximal global value $\max_{\pi} v^{\pi}(\mathcal{N})$ is sufficient to be achieved. Now, the problem transfers to solving $\max_{\pi_i} Q^{\Phi_i}(s, a_i)$ for each agent i. Aforementioned, a global reward game is identical to a potential game [1]. Additionally, Monderer and Shapley (1996) showed that in a potential game there exists a pure Nash equilibrium (i.e., a deterministic optimal policy solution). For these reasons, we apply deterministic policy gradient (DPG) (Silver et al. 2014) to search out a deterministic optimal policy. If we substitute Shapley Q-value for $Q_i^{\pi}(s, a_i)$ in DPG, we can directly write the policy gradient of each agent i such that

$$\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s \sim \rho^{\mu}}[\nabla_{\theta_i} \mu_{\theta_i}(s) \nabla_{a_i} Q^{\Phi_i}(s, a_i)|_{a_i = \mu_{\theta_i}(s)}], \tag{10}$$

where $Q^{\Phi_i}(s, a_i)$ is the Shapley Q-value for agent i and μ_{θ_i} is agent i's deterministic policy, parameterized by θ_i. A global reward is received each time step in a global reward game. Since the Shapley Q-value of each agent is correlated to a local reward, we cannot update each $\hat{Q}^{\Phi_i}(s, a_i)$ directly by the global reward. However, benefited by the property of efficiency (see Section 3.2), we can solve it out according to the minimization problem such that

$$\begin{aligned}\min_{\omega_1, \omega_2, \ldots, \omega_{|\mathcal{N}|}} \mathbb{E}_{s^t, \mathbf{a}_{\mathcal{N}}^t, r^t(\mathcal{N}), s^{t+1}} \Big[\ \frac{1}{2}\Big(\ & r^t(\mathcal{N}) \\ & + \gamma \sum_{i \in \mathcal{N}} Q^{\Phi_i}(s^{t+1}, a_i^{t+1}; \omega_i)|_{a_i^{t+1} = \mu_{\theta_i}(s^{t+1})} \\ & - \sum_{i \in \mathcal{N}} Q^{\Phi_i}(s^t, a_i^t; \omega_i) \ \Big)^2 \ \Big],\end{aligned} \tag{11}$$

where $r(\mathcal{N})$ is the global reward received from the environment each time step and $Q^{\Phi_i}(s, a_i; \omega_i)$ for each agent i is parameterized by ω_i. Constrained by this objective function, Shapley Q-value suffices the property of efficiency. Accordingly, the condition of efficient payoff distribution scheme stated in Theorem 1 is promised. Because Shapley Q-value takes all of feasible agents' actions and states as input, our algorithm actually uses the centralized critic. Nevertheless, the policies are decentralized in execution.

Silver et al. (2014) showed that DPG has the familiar machinery of policy gradient. Besides, Sutton and Barto (2018) emphasized that with a small learning rate, policy gradient algorithm can converge to a local optimum. Consequently, we can conclude that with a small learning rate, each agent can find a local maximizer and the global value $v^{\pi}(\mathcal{N})$ converges to a local maximum. The convexity of $v^{\pi}(\mathcal{N})$ is impossible to be guaranteed in applications, so the global maximum stated in Theorem 1 may not be always fulfilled.

Implementation In implementation, we replace the exact Shapley Q-value by the approximate Shapley Q-value in Eq.8. Accordingly, the Shapley Q-value here is actually a linear combination of the approximate marginal contributions (AMCs) (i.e., $\hat{\Phi}_i(s, \mathbf{a}_{\mathcal{C} \cup \{i\}}; \omega_i)$). Thanks to the powerful off-policy training strategy and function approximation on AMCs by the deep neural networks, we use the deep deterministic policy gradient (DDPG) method (Lillicrap et al. 2015) for learning policies. Additionally, we apply the reparameterization technique called Gumbel-Softmax trick (Jang, Gu, and Poole 2017) to deal with the discrete action space. Since our algorithm aims to find the optimal policies by Shapley Q-value and DDPG, we call it Shapley Q-value deep deterministic policy gradient (SQDDPG). The pseudo code for the SQDDPG is given in Appendix.

5 Experiments

We evaluate SQDDPG on Cooperative Navigation, Prey-and-Predator (Lowe et al. 2017) and Traffic Junction (Sukhbaatar, szlam, and Fergus 2016) [2]. In the experiments, we compare SQDDPG with two Independent algorithms (with decentralised critics), e.g., Independent DDPG (IDDPG) (Lillicrap et al. 2015) and Independent A2C (IA2C) (Sutton and Barto 2018), and two state-of-the-art methods with centralised critics, e.g., MADDPG (Lowe et al. 2017) and COMA (Foerster et al. 2018). In the experiments, to keep the fairness of comparison, the policy and critic networks for all MARL algorithms are parameterized by MLPs. All models are trained by the Adam optimizer (Kingma and Ba 2014). The details of experimental setups are given in Appendix.

5.1 Cooperative Navigation

Environment Settings In this task, there are 3 agents and 3 targets. Each agent aims to move to a target, with no prior allocations of the targets to each agent. The state of each agent in this environment includes its current position and velocity, the displacement to the three targets, and the displacement to other agents. The action space of each agent is `move_up`, `move_down`, `move_right`, `move_left` and `stay`. The global reward of this environment is the negative sum of the distance between each target and the nearest agent to it. Besides, if a collision happens, then the global reward will be reduced by 1.

Results As seen from Fig.2, SQDDPGs with variant sample sizes (i.e., M in Eq.8) outperform the baselines on the convergence rate. We believe that if more training episodes are permitted, the algorithms except for IA2C can achieve

[1] A potential game is a game where there exists a potential function (Monderer and Shapley 1996).

[2] The code of experiments is available on: https://github.com/hsvgbkhgbv/SQDDPG

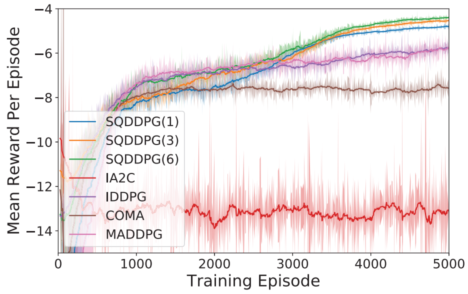

Figure 2: Mean reward per episode during training in Cooperative Navigation. SQDDPG(n) indicates SQDDPG with the sample size of n. In the rest of experiments, since only SQDDPG with the sample size (i.e., M in Eq.8) of 1 is run, we just use SQDDPG to represent SQDDPG(1).

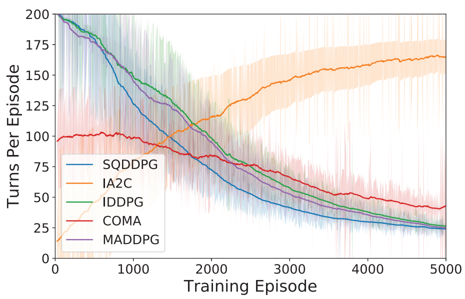

Figure 3: Turns to capture the prey per episode during training in Prey-and-Predator.

the similar performance as that of SQDDPG. Therefore, our result supports the previous argument that the local reward approach converges faster than the global reward approach (Balch 1997; 1999). As the sample size grows, the approximate Shapley Q-value estimation in Eq.7 could be more accurate and easier to converge to the optimal value. This explains the reason why the convergence rate of SQDDPG becomes faster when the sample size increases. Since we show that SQDDPG with the sample size of 1 can finally obtain nearly the same performance as that of other variants, we just run it in the rest of experiments to reduce the computational complexity.

5.2 Prey-and-Predator

Environment Settings In this task, we can only control three predators and the prey is a random agent. The aim of each predator is coordinating to capture the prey by as less steps as possible. The state of each predator contains its current position and velocity, the respective displacement to the prey and other predators, and the velocity of the prey. The action space is the same as that defined in Cooperative Navigation. The global reward is the negative minimal distance between any predator and the prey. If the prey is caught by any predator, the global reward will be added by 10 and the game terminates.

Results As Fig.3 shows, SQDDPG converges fastest with around 25 turns to capture the prey finally, followed by MADDPG and IDDPG. To study and understand the credit assignment, we visualize the Q-values of each MARL algorithm for one randomly selected trajectory of states and actions from an expert policy. For the convenient visualization, we normalize the Q-values by min-max normalization (Patro and Sahu 2015) for each MARL algorithm. We can see from Fig.4 that the credit assignment of SQDDPG is more explainable than that of the baselines. Specifically, it is intuitive that the credit assigned to each agent by SQDDPG is inversely proportional to its distance to the prey. However, other MARL algorithms do not explicitly show such a property. To validate this hypothesis, we also evaluate it quantitatively by Pearson correlation coefficient (PCC) on 1000 randomly selected transition samples for the correlation between the credit assignment and the reciprocal of each predator's distance to the prey. The value of PCC is greater, and the inverse proportion is stronger. As Tab.1 shows, SQDDPG expresses the inverse proportion significantly, with the PCC of 0.3210 and the two-tailed p-value of 1.9542e-19. If a predator is closer to the prey, it is more likely to catch the prey and the contribution of that predator should be more significant. Consequently, we demonstrate that the credit assignment of Shapley Q-value is fair.

5.3 Traffic Junction

Environment Settings In this task, cars move along the predefined routes which intersect on one or more traffic junctions. At each time step, new cars are added to the environment with the probability p_{arrive}, and the total number of cars is limited below $N_{\max}$. After a car finishes its mission, it will be removed from the environment and possibly sampled back to a new route. Each car has a limited vision of 1, which means it can only observe the circumstance within the 3x3 region surrounding it. No communication between cars is permitted in our experiment, in contrast to the others' experiments on the same task (Sukhbaatar, szlam, and Fergus 2016; Das et al. 2018). The action space of each car is `gas` and `brake`, and the global reward function is $\sum_{i=1}^{N}$ - $0.01 t_i$, where t_i is the time steps that car i is continuously active on the road in one mission and N is the total number of cars. Additionally, if a collision occurs, the global reward will be reduced by 10. We evaluate the performance by the success rate (i.e., the episode that no collisions happen).

Results We compare our method with the baselines on the easy, medium and hard version of Traffic Junction. The easy version is constituted of one traffic junction of two one-way roads on a 7×7 grid with $N_{\max} = 5$ and $p_{\text{arrive}} = 0.3$. The medium version is constituted of one traffic junction of two-way roads on a 14×14 grid with $N_{\max} = 10$ and $p_{\text{arrive}} = 0.2$. The hard version is constituted of four connected traffic junctions of two-way roads on a 18×18 grid with $N_{\max} = 20$ and $p_{\text{arrive}} = 0.05$. From Tab.2, we can see that on the easy version, except for IA2C, other algorithms can get a success rate over 93%, since this scenario is too easy. On the medium and hard version, SQDDPG outperforms the baselines with the success rate of 88.98% on the medium version and 87.04% on the hard version. Moreover,

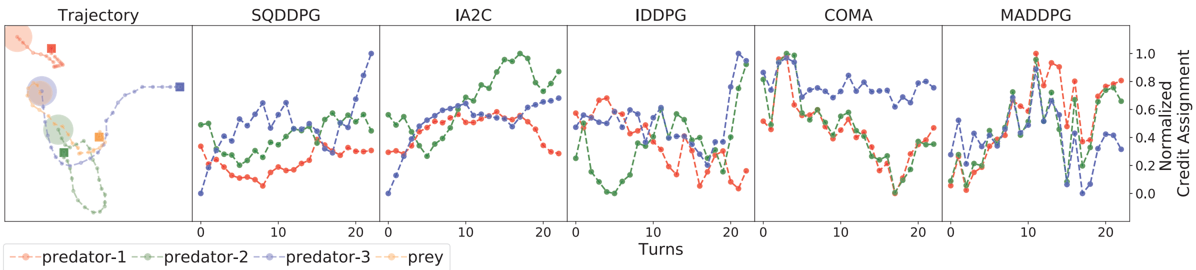

Figure 4: Credit assignment to each predator for a fixed trajectory. The leftmost figure records a trajectory sampled by an expert policy. The square represents the initial position whereas the circle indicates the final position of each agent. The dots on the trajectory indicates each agent's temporary positions. The other figures show the normalized credit assignments generated by different MARL algorithms according to this trajectory.

	IA2C	**IDDPG**	**COMA**	**MADDPG**	**SQDDPG**
coefficient	0.0508	0.0061	0.1274	0.0094	**0.3210**
two-tailed p-value	1.6419e-1	8.6659e-1	4.6896e-4	7.9623e-1	**1.9542e-19**

Table 1: The Pearson correlation coefficient between the credit assignment to each predator and the reciprocal of its distance to the prey. This test is conducted by 1000 randomly selected episode samples.

Difficulty	**IA2C**	**IDDPG**	**COMA**	**MADDPG**	**SQDDPG**
Easy	65.01%	93.08%	93.01%	**93.72%**	93.26%
Medium	67.51%	84.16%	82.48%	87.92%	**88.98%**
Hard	60.89%	64.99%	85.33%	84.21%	**87.04%**

Table 2: The success rate on Traffic Junction, tested with 20, 40, and 60 steps per episode in easy, medium and hard versions respectively. The results are obtained by running each algorithm after training for 1000 episodes.

the performance of SQDDPG significantly exceeds that of no-communication algorithms reported as 84.9% and 74.1% in (Das et al. 2018). We demonstrate that SQDDPG can solve out the large-scale problems.

5.4 Discussion

In the experimental results, it is surprising that IDDPG achieves a good performance. The possible reason could be that a potential game (i.e., a global reward game) can be solved by the fictitious play (Monderer and Shapley 1996) and DDPG is analogous to it, finding an optimal deterministic policy by fitting the other agents' behaviours. However, the convergence rate is not guaranteed when the number of agents becomes large, such as the result shown in the hard version of Traffic Junction. To deal with both of competitive and cooperative games, MADDPG assigns each agent a centralised critic to estimate the global reward. Theoretically the credits assigned to agents are identical, though in experiment it does not always exhibit this property. The possible reason could be the bias existing in the Q-values. We can see from Tab.1 that both of COMA and SQDDPG exhibit the feature of credit assignment. Nonetheless, SQDDPG performs more significantly on the fairness. This well validates our motivation of importing the local reward approach (e.g., Shapley value) to solve out the credit assignment problems.

6 Conclusion

We introduce the cooperative game theory to extend the existing global reward game to a broader framework called extended convex game (ECG). Under this framework, we propose an algorithm namely Shapley Q-value deep deterministic policy gradient (SQDDPG), leveraging a local reward approach called Shapley Q-value, which is theoretically guaranteed to find out the optimal solution in an ECG with the grand coalition (i.e., a global reward game). We evaluate SQDDPG on three global reward games (i.e., Cooperative Navigation, Prey-and-Predator and Traffic Junction) and show the promising performance compared with the baselines (e.g., MADDPG, COMA, IDDPG and IA2C). During the experiments, SQDDPG exhibits the fair credit assignments and fast convergence rate. In the future work, we plan to dynamically group the agents at each time step with some theoretical guarantees and jump out of the restriction of global reward game.

Acknowledgements

This work is sponsored by EPSRC-UKRI Innovation Fellowship EP/S000909/1. Jianhong Wang especially thanks Ms Yunlu Li for helpful and patient explanations on mathematics.

References

Balch, T. 1997. Learning roles: Behavioral diversity in robot teams. *College of Computing Technical Report GIT-CC-97-12, Georgia Institute of Technology, Atlanta, Georgia* 73.

Balch, T. 1999. Reward and diversity in multirobot foraging. *In IJCAI-99 Workshop on Agents Learning About, From and With other Agents*.

Basar, T., and Olsder, G. J. 1999. *Dynamic noncooperative game theory*, volume 23. Siam.

Chalkiadakis, G., and Boutilier, C. 2004. Bayesian reinforcement learning for coalition formation under uncertainty. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 3*, 1090–1097. IEEE Computer Society.

Chalkiadakis, G.; Elkind, E.; and Wooldridge, M. 2011. Computational aspects of cooperative game theory. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5(6):1–168.

Chang, Y.-H.; Ho, T.; and Kaelbling, L. P. 2004. All learning is local: Multi-agent learning in global reward games. In *Advances in neural information processing systems*, 807–814.

Das, A.; Gervet, T.; Romoff, J.; Batra, D.; Parikh, D.; Rabbat, M.; and Pineau, J. 2018. Tarmac: Targeted multi-agent communication. *arXiv preprint arXiv:1810.11187*.

Faigle, U., and Kern, W. 1992. The shapley value for cooperative games under precedence constraints. *International Journal of Game Theory* 21(3):249–266.

Fatima, S. S.; Wooldridge, M.; and Jennings, N. R. 2008. A linear approximation method for the shapley value. *Artificial Intelligence* 172(14):1673–1699.

Foerster, J. N.; Farquhar, G.; Afouras, T.; Nardelli, N.; and Whiteson, S. 2018. Counterfactual multi-agent policy gradients. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Hu, J., and Wellman, M. P. 1998. Multiagent reinforcement learning: theoretical framework and an algorithm. In *ICML*, volume 98, 242–250. Citeseer.

Iqbal, S., and Sha, F. 2018. Actor-attention-critic for multi-agent reinforcement learning. *arXiv preprint arXiv:1810.02912*.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Keviczky, T.; Borrelli, F.; Fregene, K.; Godbole, D.; and Balas, G. J. 2007. Decentralized receding horizon control and coordination of autonomous vehicle formations. *IEEE Transactions on control systems technology* 16(1):19–33.

Kim, D.; Moon, S.; Hostallero, D.; Kang, W. J.; Lee, T.; Son, K.; and Yi, Y. 2019. Learning to schedule communication in multi-agent reinforcement learning. In *International Conference on Learning Representations*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koes, M.; Nourbakhsh, I.; and Sycara, K. 2006. Constraint optimization coordination architecture for search and rescue robotics. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, 3977–3982. IEEE.

Konda, V. R., and Tsitsiklis, J. N. 2000. Actor-critic algorithms. In *Advances in neural information processing systems*, 1008–1014.

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, O. P.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, 6379–6390.

Michalak, T. P.; Aadithya, K. V.; Szczepanski, P. L.; Ravindran, B.; and Jennings, N. R. 2013. Efficient computation of the shapley value for game-theoretic network centrality. *Journal of Artificial Intelligence Research* 46:607–650.

Monderer, D., and Shapley, L. S. 1996. Potential games. *Games and economic behavior* 14(1):124–143.

Nash, J. 1953. Two-person cooperative games. *Econometrica: Journal of the Econometric Society* 128–140.

Nguyen, D. T.; Kumar, A.; and Lau, H. C. 2018. Credit assignment for collective multiagent rl with global rewards. In *Advances in Neural Information Processing Systems*, 8102–8113.

Omidshafiei, S.; Kim, D.-K.; Liu, M.; Tesauro, G.; Riemer, M.; Amato, C.; Campbell, M.; and How, J. P. 2018. Learning to teach in cooperative multiagent reinforcement learning. *arXiv preprint arXiv:1805.07830*.

Osborne, M. J., and Rubinstein, A. 1994. *A course in game theory*. MIT press.

Panait, L., and Luke, S. 2005. Cooperative multi-agent learning: The state of the art. *Autonomous agents and multi-agent systems* 11(3):387–434.

Patro, S., and Sahu, K. K. 2015. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*.

Peleg, B., and Sudhölter, P. 2007. *Introduction to the theory of cooperative games*, volume 34. Springer Science & Business Media.

Ramchurn, S. D.; Farinelli, A.; Macarthur, K. S.; and Jennings, N. R. 2010. Decentralized coordination in robocup rescue. *The Computer Journal* 53(9):1447–1461.

Schuldt, A. 2012. Multiagent coordination enabling autonomous logistics. *KI-Künstliche Intelligenz* 26(1):91–94.

Shapley, L. S. 1953a. Stochastic games. *Proceedings of the national academy of sciences* 39(10):1095–1100.

Shapley, L. S. 1953b. A value for n-person games. *Contributions to the Theory of Games* 2(28):307–317.

Shapley, L. S. 1971. Cores of convex games. *International journal of game theory* 1(1):11–26.

Silver, D.; Lever, G.; Heess, N.; Degris, T.; Wierstra, D.; and Riedmiller, M. 2014. Deterministic policy gradient algorithms. In *ICML*.

Suijs, J., and Borm, P. 1999. Stochastic cooperative games: superadditivity, convexity, and certainty equivalents. *Games and Economic Behavior* 27(2):331–345.

Sukhbaatar, S.; szlam, a.; and Fergus, R. 2016. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc. 2244–2252.

Sutton, R. S., and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Watkins, C. J., and Dayan, P. 1992. Q-learning. *Machine learning* 8(3-4):279–292.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8(3-4):229–256.

Wolpert, D. H., and Tumer, K. 2002. Optimal payoff functions for members of collectives. In *Modeling complexity in economic and social systems*. World Scientific. 355–369.