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

1 Control and Power Research Group (CAP), Imperial College London, UK
2 Imperial Computer Vision & Learning Lab (ICVL), Imperial College London, UK
3 Laiye Network Technology Co.Ltd., China

Jianhong Wang[12], Yuan Zhang[3], Tae-Kyun Kim[2], Yunjie Gu[1]

open source code: https://github.com/hsvgbkhgbv/SQDDPG

Syllabus

- Motivation of Cooperative Game
- Global Reward Approach vs. Local Reward Approach
- Cooperative Game Theory
- Shapley Q-value
- Shapley Q-value Deep Deterministic Policy Gradient
- Experimental Results
- Conclusion

Motivation of Cooperative（Global Reward）Game

1. As opposed to competing with others in a competitive game, agents in a cooperative game aim to cooperate to solve a joint task or maximize the global reward [1].
2. Many scenarios can be modelled as cooperative games, e.g., Electric Power System, Transportation System, and Communication Networks.

Motivation of Cooperative Game (Cont.)

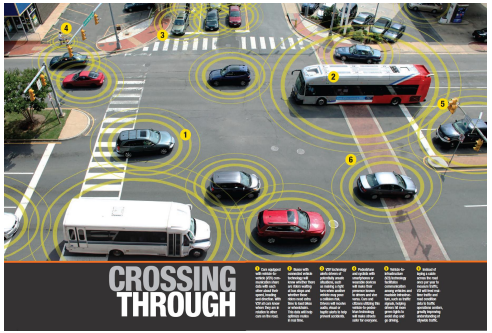


Figure 1: Transportation System

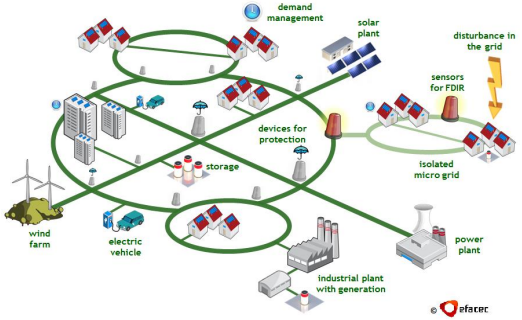


Figure 2: Electric Power System

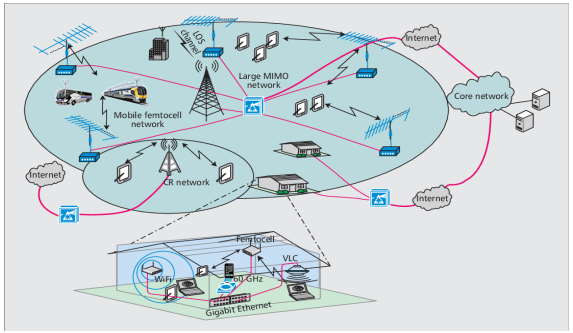


Figure 3: Communication System

Global Reward Approach vs. Local Reward Approach

- It gives each agent the global reward directly.
- The game converges to Nash Equilibrium.

- It gives each agent an accurate reward for its contribution.
- The accurate credit assignment can accelerate exploring the optimal policy.

Cooperative Game Theory – Shapley Value

- Definition: It is a theoretical framework mainly focusing on constructing groups and finding the proper credit assignment to each agent.
- Shapley Value [2] is a conventional method to distribute the credits in a Convex Game.
  - It can guarantee the existence of a solution in the Core in a Convex Game.
  - It considers all possible permutations of sequences to form a grand coalition.
  - It calculates the marginal contribution [3] of each agent when the agent joins in an arbitrary intermediate coalition.

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(N - |S| - 1)!}{N!} (v(S \cup \{i\}) - v(S))$$

The probability of the occurrence of some possible intermediate coalition that the agent can join in.

The marginal value of the agent when it intends to join in some intermediate coalition.

Shapley Q-value

- Extend the vanilla Shapley Value (i.e. stateless) by state and action.

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

The Q-value of some coalition when the agent joins in.

The Q-value of some coalition without the agent.

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

The probability of the occurrence of some possible intermediate coalition that the agent can join in.

The marginal value of the agent when it intends to join in some intermediate coalition.

Shapley Q-value (Cont.)

- Since the drastic complexity and instability to compute the Shapley Q-value analytically, we approximate it.

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

Probabilistic Form of Shapley Q-value $$Q^{\Phi_i}(s, a_i) = \mathbb{E}_{\mathcal{C}\sim Pr(\mathcal{C}|\mathcal{N}\setminus\{i\})}[\ \Phi_i(\mathcal{C})\ ]$$

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

Shapley Q-value Deep Deterministic Policy Gradient (SQDDPG)

Deterministic Policy Gradient

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

The objective function to optimize Shapley Q-values.

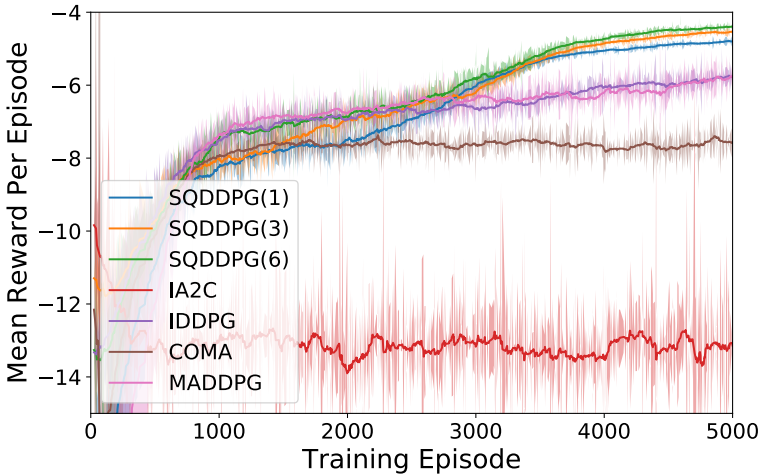
$$\min_{\omega_1,\omega_2,\ldots,\omega_{|\mathcal{N}|}} \mathbb{E}_{s^t,\mathbf{a}^t_{\mathcal{N}},r^t(\mathcal{N}),s^{t+1}}\Big[\ \frac{1}{2}\Big(\ r^t(\mathcal{N}) + \gamma \sum_{i\in\mathcal{N}} Q^{\Phi_i}(s^{t+1}, a_i^{t+1};\omega_i)|_{a_i^{t+1}=\mu_{\theta_i}(s^{t+1})} - \sum_{i\in\mathcal{N}}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

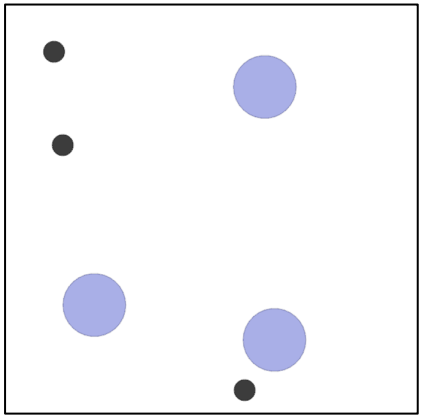
Experimental Results

- Evaluation on three environments: Cooperative Navigation, Prey-and-Predator [4], Traffic Junction [5].
- Baselines: IA2C [6], COMA [3], IDDPG [7], MADDPG [4].

Experimental Results (Cooperative Navigation)

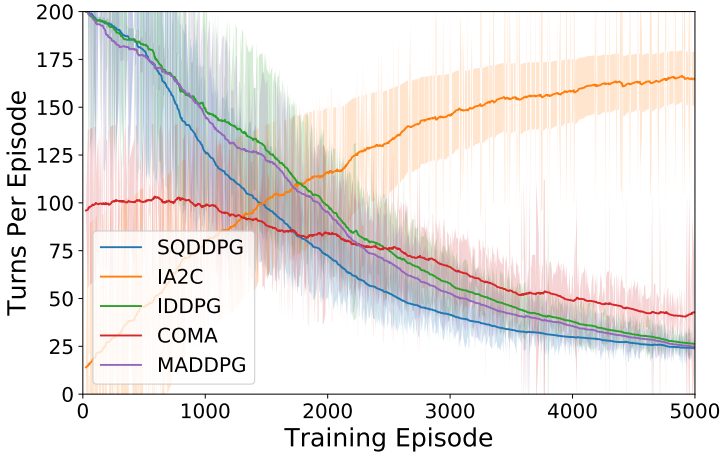
- Task: 3 agents moves to 3 targets respectively without any pre-allocations.
- Control variables: force direction.

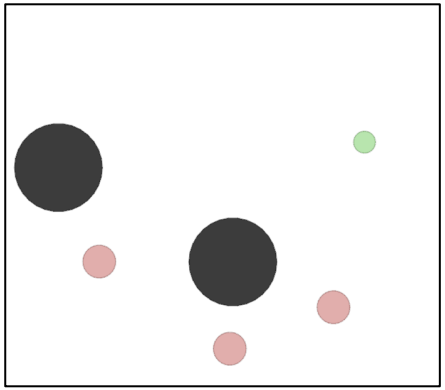




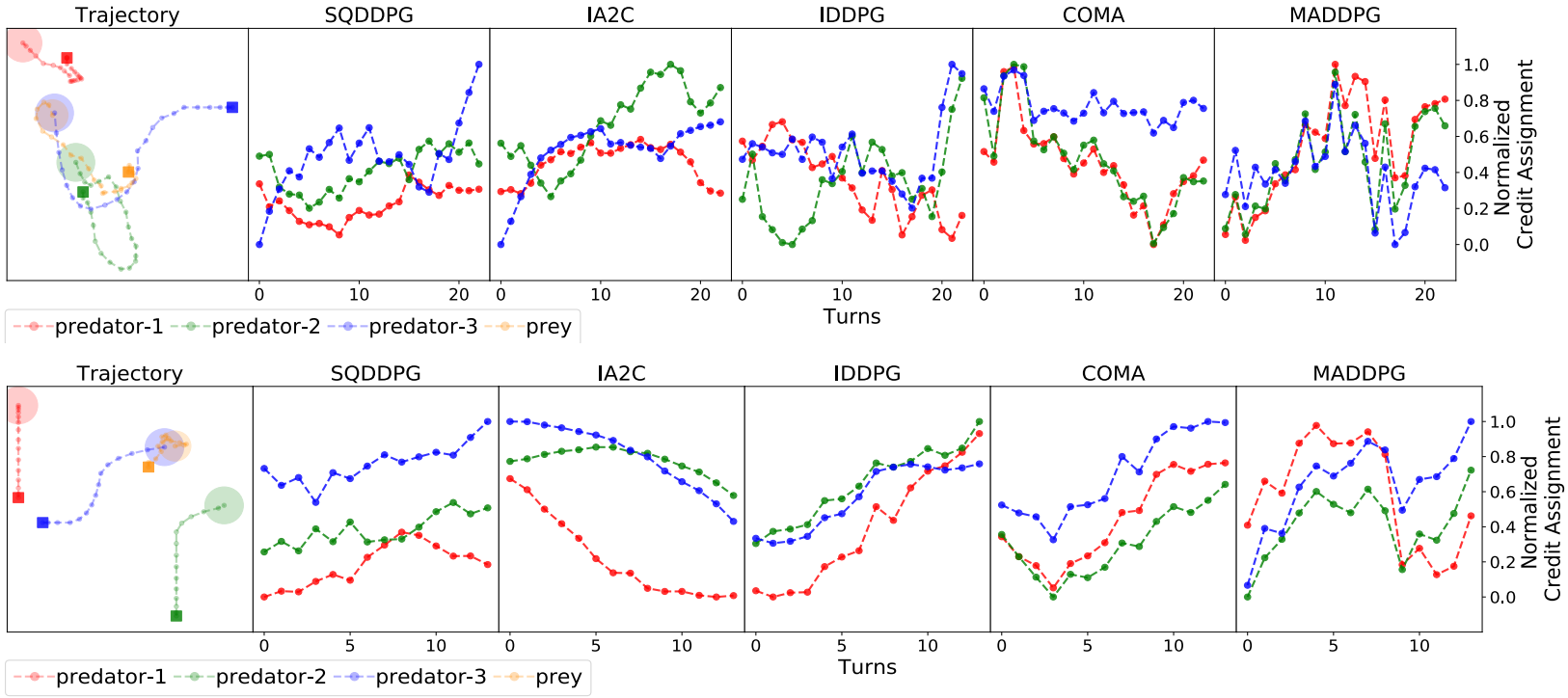
Experimental Results (Prey-and-Predator)

- Task: 3 predators attempt to cooperate to capture the prey.
- Control variables: force direction.





Experimental Results (Prey-and-Predator Cont.)



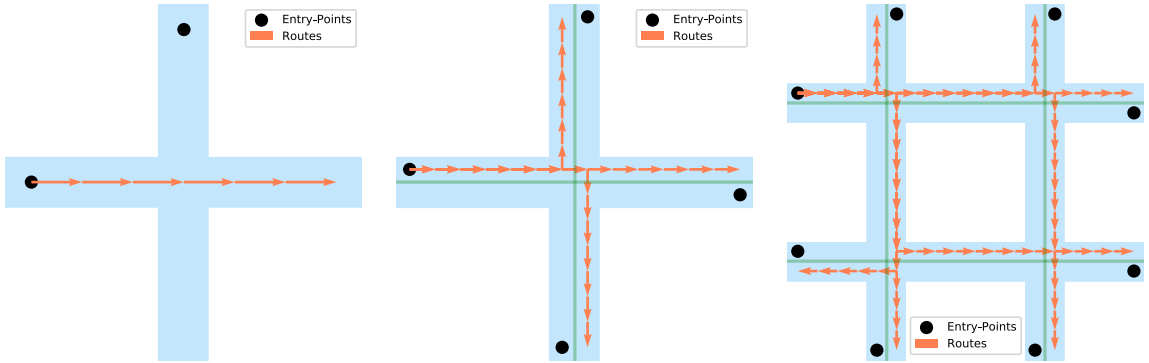
Experimental Results (Prey-and-Predator Cont.)

	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**

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.

Experimental Results (Traffic Junction)

- Task: Several vehicles move along the road of a traffic junction in case that the collision happens. This needs the corporation among vehicles.
- Control variables: Gas and Brake.



Experimental Results (Traffic Junction Cont.)

- Task Level Difficulty:
  - Easy: 7x7 Grid, 1 junction with 2 one-way roads, N_{max}=5, p_{arrive}=0.3
  - Medium: 14x14 Grid, 1 junction with 2 two-way roads, N_{max}=10, p_{arrive}=0.2
  - Hard: 18x18 Grid, 4 junctions with 4 two-way roads, N_{max}=20, p_{arrive}=0.05

Experimental Results (Traffic Junction Cont.)

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%**

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.

Conclusion

- We extend Shapley Value (i.e. a credit assignment method in Convex Game) to Shapley Q-value by considering state and action.
- We utilize Shapley Q-value to assign the credit to each agent in Global Reward Game.
- We propose an algorithm called SQDDPG according to Shapley Q-value.
- SQDDPG achieves better performances in 3 experiments compared with other baselines

Reference List

1. Chalkiadakis, Georgios, Edith Elkind, and Michael Wooldridge. "Computational aspects of cooperative game theory." *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5.6 (2011): 1-168.
2. Shapley, Lloyd S. "A value for n-person games." *Contributions to the Theory of Games* 2.28 (1953): 307-317.
3. Foerster, Jakob N., et al. "Counterfactual multi-agent policy gradients." *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.
4. Lowe, Ryan, et al. "Multi-agent actor-critic for mixed cooperative-competitive environments." *Advances in Neural Information Processing Systems*. 2017.
5. Sukhbaatar, Sainbayar, and Rob Fergus. "Learning multiagent communication with backpropagation." *Advances in Neural Information Processing Systems*. 2016.
6. Konda, Vijay R., and John N. Tsitsiklis. "Actor-critic algorithms." *Advances in neural information processing systems*. 2000.
7. Lillicrap, Timothy P., et al. "Continuous control with deep reinforcement learning." *arXiv preprint arXiv: 1509.02971* (2015).

The End!