Multi-Agent Reinforcement Learning with General Utilities via Decentralized Shadow Reward Actor-Critic

JUNYU ZHANG, Princeton University, USA
AMRIT SINGH BEDI, U.S. Army Research Laboratory, USA
MENGDI WANG, Princeton University and Deepmind, USA
ALEC KOPPEL, U.S. Army Research Laboratory, USA

CCS Concepts: • **Computing methodologies** → **Machine Learning**; *Artificial Intelligence*; • **Mathematics of computing** → Probability and Statistics; *Information Theory*.

Additional Key Words and Phrases: reinforcement learning, multi-agent systems, risk-sensitivity, cooperation, exploration

ACM Reference Format:
Junyu Zhang, Amrit Singh Bedi, Mengdi Wang, and Alec Koppel. 2021. Multi-Agent Reinforcement Learning with General Utilities via Decentralized Shadow Reward Actor-Critic. In *RLNQ 2021: Reinforcement Learning in Networks and Queues, a workshop in conjunction with ACM SIGMETRICS 2021, June 14, 2021, Virtual-Only Conference.* ACM, New York, NY, USA, 3 pages. https://doi.org/DOITBD

EXTENDED ABSTRACT

Reinforcement learning (RL) is a framework for directly estimating the parameters of a controller through repeated interaction with the environment, and has gained attention for its ability to alleviate the need for a physically exact model across a number of domains, such as robotic manipulation [12], web services [31], and various games [25]. In RL, an agent in a given state takes an action, and transits to another according to a Markov transition density, whereby a reward informing the merit of the action is revealed by the environment. Mathematically, this setting may be encapsulated by a Markov Decision Process (MDP) [23], in which the one seeks to select the action sequence to maximize the long-term accumulation of rewards.

In many domains, multiple agents interact in order to obtain favorable outcomes, as in finance [17], social networks [9], and games [27]. In multi-agent RL (MARL) and more generally, stochastic games, a key question is the payoff structure [2]. We focus on common payoffs among agents, i.e., the utility of the team is the sum of local utilities [4], which contrasts with competitive settings where one agent's gain is another's loss, or combinations thereof [18]. Whereas typically cooperative MARL defines the global utility as the average over agents' local reward accumulations, here we define a *new mechanism for cooperation* that permits agents to incorporate risk-sensitivity [3, 22], prior experience [1], or exploration [8]. The usual common-payoff setting focuses on global cumulative return of rewards, which is a linear function of the the state-action occupancy measure. By contrast, the aforementioned decision-making goals define *nonlinear* functions of the state-action occupancy measure [10]. Such functions, we call *general utilities*, have recently yielded impressive performance in practice via prioritizing exploration [7], risk-sensitivity [21], and prior

Manuscript submitted to ACM

Objective	Approach	Convergence
Cumulative Return	Value-Based [6, 11, 15, 24, 28]	✓
	Policy-Based [5, 30]	✓
Risk	[21]	✗
Exploration	[7, 20]	✗
Priors	[14, 16]	✗

This Work ✓

Table 1. Cumulative Returns, Risk-Sensitivity, Exploration, and the incorporation of Priors are common goals in MARL, and subsumed by the general utilities considered here. We focus on the setting when agents are **cooperative** and transition according to a common **global** dynamics model [4]. We defer a discussion of centralized training decentralized execution (CTDE), partial observability, and different transition models to appendices, with the understanding that our focus is on **decentralized training** under **full observability**. The respective technical settings of [7, 14, 16, 20, 21] are different; their inclusion here is to underscore their use of goals beyond cumulative return, which is given a conceptual underpinning for the first time in this work.

experience [16]. To date, however, there exists few formal guarantees for algorithms designed to optimize general utilities in multi-agent settings, to the best of our knowledge.

This gap motivates us to put forth the first decentralized MARL scheme for general utilities, and establish its consistency and sample complexity. Our approach hinges upon first noting that the embarking point for most RL methodologies is the Policy Gradient Theorem [26] or Bellman's equation, both of which break down for general utilities. One potential path forward is a recent generalization of the PG Theorem for general utilities [29], which expresses the gradient as product of the partial derivative of the utility with respect to the occupancy measure, and the occupancy measure with respect to the policy. However, in the team setting, this later factor is a *global nonlinear function* of agents' policies, and hence does not permit decentralization. Thus, we define an agent's local occupancy measure as the joint occupancy measure of all agents' polices with all others' marginalized out, and its local general utility as any (not-necessarily concave) function of its marginal occupancy measure. The team objective, then, is the global aggregation of all local utilities.

From this definition, we derive a new variant of the Policy Gradient where each agent estimate its policy gradient based on local information and message passing with neighbors. Specifically, we derive a model-free algorithm, **D**ecentralized **S**hadow Reward **A**ctor-**C**ritic (DSAC), that generalizes multi-agent actor-critic (see [13]) beyond cumulative return [30]. Each agent's procedure follows four stages: (i) a marginalized occupancy measure estimation step used to evaluate the instantaneous gradient of the local utility with respect to the occupancy measure, which we dub the "shadow reward"; (ii) accumulate "shadow rewards" along a trajectory to estimate "shadow" critic parameters (critic); (iii) average critic parameters with those of its neighbors (information mixing); and (iv) a stochastic policy gradient ascent step along trajectories (actor).

Contributions. Overall, our contributions are:

- present the first MARL formulation for broader goals than the cumulative return and specialization among agents' roles;
- derive a variant of multi-agent actor-critic to solve this problem that employs an occupancy measure estimation step to construct the gradient of the general utility with respect to the occupancy measure, which serves as a "shadow reward" for the critic step;
- for ϵ-stationarity with high probability, we respectively establish that DSAC requires $\mathcal{O}(1/\epsilon^{2.5})$ and $\mathcal{O}(1/\epsilon^{2})$ steps if agents exchange information onceor multiple times per policy update. Under proper assumptions, we further establish the convergence to the globally optimal policy under diminishing step-sizes.
- provide experimental evaluation for exploration maximization and safe navigation in cooperative settings [19].

REFERENCES

[1] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robotics and autonomous systems* 57, 5 (2009), 469–483.

[2] Tamer Başar and Geert Jan Olsder. 1998. *Dynamic noncooperative game theory*. SIAM.

[3] Vivek S Borkar and Sean P Meyn. 2002. Risk-sensitive optimal control for Markov decision processes with monotone cost. *Mathematics of Operations Research* 27, 1 (2002), 192–209.

[4] Lucian Busoniu, Robert Babuska, and Bart De Schutter. 2008. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, 2 (2008), 156–172.

[5] Tianyi Chen, Kaiqing Zhang, Georgios B Giannakis, and Tamer Başar. 2018. Communication-efficient distributed reinforcement learning. *arXiv preprint arXiv:1812.03239* (2018).

[6] Thinh Doan, Siva Maguluri, and Justin Romberg. 2019. Finite-Time Analysis of Distributed TD (0) with Linear Function Approximation on Multi-Agent Reinforcement Learning. In *International Conference on Machine Learning*. 1626–1635.

[7] Tarun Gupta, Anuj Mahajan, Bei Peng, Wendelin Böhmer, and Shimon Whiteson. 2020. UneVEn: Universal Value Exploration for Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:2010.02974* (2020).

[8] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. 2019. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*. PMLR, 2681–2691.

[9] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*. PMLR, 3040–3049.

[10] L. C. M. Kallenberg. 1994. Survey of linear programming for standard and nonstandard Markovian control problems. Part I: Theory. *Zeitschrift für Operations Research* 40, 1 (1994), 1–42.

[11] Soummya Kar, José MF Moura, and H Vincent Poor. 2013. QD-Learning: A Collaborative Distributed Strategy for Multi-Agent Reinforcement Learning Through Consensus+ Innovations. *IEEE Transactions on Signal Processing* 61, 7 (2013), 1848–1862.

[12] Jens Kober, J Andrew Bagnell, and Jan Peters. 2013. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research* 32, 11 (2013), 1238–1274.

[13] Vijaymohan R Konda and Vivek S Borkar. 1999. Actor-Critic–Type Learning Algorithms for Markov Decision Processes. *SIAM Journal on Control and Optimization* 38, 1 (1999), 94–123.

[14] Hoang M Le, Yisong Yue, Peter Carr, and Patrick Lucey. 2017. Coordinated Multi-Agent Imitation Learning. *Proceedings of Machine Learning Research* 70 (2017), 1995–2003.

[15] Donghwan Lee, Hyungjin Yoon, V Cichella, and N Hovakimyan. 2018. Stochastic primal-dual algorithm for distributed gradient temporal difference learning. *arXiv preprint arXiv:1805.07918* (2018).

[16] Hyun-Rok Lee and Taesik Lee. 2019. Improved cooperative multi-agent reinforcement learning algorithm augmented by mixing demonstrations from centralized policy. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 1089–1098.

[17] Jae Won Lee, Byoung-Tak Zhang, et al. 2002. Stock Trading System Using Reinforcement Learning with Cooperative Agents. In *Proceedings of the Nineteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 451–458.

[18] Michael L Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*. Elsevier, 157–163.

[19] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *Neural Information Processing Systems (NIPS)* (2017).

[20] Anuj Mahajan, Tabish Rashid, Mikayel Samvelyan, and Shimon Whiteson. 2019. Maven: Multi-agent variational exploration. In *Advances in Neural Information Processing Systems*. 7613–7624.

[21] Mysterious Mystery. 2021. RMIX: Risk-Sensitive Multi-Agent Reinforcement Learning. *Under Review at International Conference on Learning Representations* (2021).

[22] LA Prashanth and Mohammad Ghavamzadeh. 2016. Variance-constrained actor-critic algorithms for discounted and average reward MDPs. *Machine Learning* 105, 3 (2016), 367–417.

[23] Martin L Puterman. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

[24] Chao Qu, Shie Mannor, Huan Xu, Yuan Qi, Le Song, and Junwu Xiong. 2019. Value propagation for decentralized networked deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*. 1184–1193.

[25] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.

[26] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*. 1057–1063.

[27] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[28] Hoi-To Wai, Zhuoran Yang, Zhaoran Wang, and Mingyi Hong. 2018. Multi-agent reinforcement learning via double averaging primal-dual optimization. In *Advances in Neural Information Processing Systems*. 9649–9660.

[29] Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. 2020. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems* 33 (2020).

[30] Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. 2018. Fully Decentralized Multi-Agent Reinforcement Learning with Networked Agents. In *International Conference on Machine Learning*. 5872–5881.

[31] Xiangyu Zhao, Long Xia, Liang Zhang, Zhuoye Ding, Dawei Yin, and Jiliang Tang. 2018. Deep reinforcement learning for page-wise recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 95–103.