# A Distributional Perspective on Value Function Factorization Methods for Multi-Agent Reinforcement Learning

Extended Abstract

Wei-Fang Sun
Department of Computer Science
National Tsing Hua University
j3soon@gapp.nthu.edu.tw

Cheng-Kuang Lee
NVIDIA AI Technology Center
NVIDIA Corporation
ckl@nvidia.com

Chun-Yi Lee
Department of Computer Science
National Tsing Hua University
cylee@cs.nthu.edu.tw

## ABSTRACT

Distributional reinforcement learning (RL) provides beneficial impacts for the single-agent domain. However, distributional RL methods are not directly compatible with value function factorization methods for multi-agent reinforcement learning. This work provides a distributional perspective on value function factorization, offering a solution for bridging the gap between distributional RL and value function factorization methods.

## KEYWORDS

Reinforcement Learning; Multi-Agent RL; Distributional RL

**ACM Reference Format:**
Wei-Fang Sun, Cheng-Kuang Lee, and Chun-Yi Lee. 2021. A Distributional Perspective on Value Function Factorization Methods for Multi-Agent Reinforcement Learning: Extended Abstract. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 3 pages.

## 1 INTRODUCTION

Value function factorization methods (e.g. VDN [1], QMIX [2, 3]) for multi-agent reinforcement learning (MARL) offer promising performance in complex multi-agent scenarios [4]. In MARL settings, the environments are highly stochastic due to each agent's partial observability and the continuously changing policies of the other agents. To deal with the above issues, distributional reinforcement learning (RL) is a potential solution that has been empirically proven successful in a wide range of single-agent domains [5–9]. However, distributional RL has not been applied to value function factorization methods in MARL domains to decompose the approximation of joint return distributions into individual utility value functions (or simply '*utilities*' hereafter). In this paper, we bridge the gap between distributional RL and value function factorization by decomposing the approximation of (1) the mean and (2) the shape of joint return distributions. Based on such decomposition, we further propose two practical implementations to generalize QMIX to its distributional variants. The first implementation models the probability mass function (PMF) of individual value function distributions as categorical distributions, and combines them through 1D convolution. The second implementation models the quantile functions of individual value function distributions, and combines them through quantile mixture. To validate the two implementations, we demonstrate their ability to factorize stochastic rewards and present the visualizations of their approximation results.

Wei-Fang Sun contributed to the work during his NVIDIA internship.

*Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.*


## 2 BACKGROUND

### 2.1 Cooperative MARL and CTDE

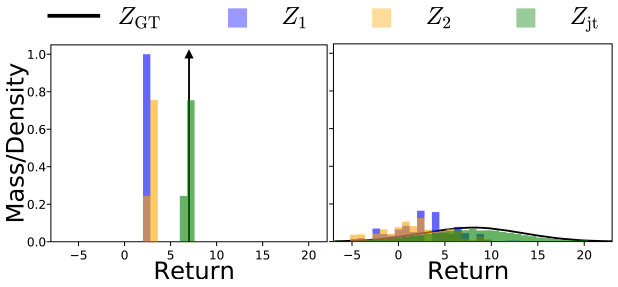
A fully cooperative MARL environment can be modeled as a DEC-POMDP [10], which is described as a tuple $\langle \mathbb{S}, \mathbb{K}, \mathbb{O}, \mathbb{U}, P, O, R, \gamma \rangle$, where $\mathbb{S}$ is the state space, $\mathbb{K}$ is the set of all agents, $\mathbb{O}$ is the set of joint observations, $\mathbb{U}$ is the set of joint actions, $P$ is the state transition function. Each agent perceives a partial observation according to an observation function $O$. All agents share the same joint reward function $R$, and use $\gamma$ as the discount factor. Under such an MARL formulation, we focus on centralized training with decentralized execution (CTDE) methods, where the agents are trained in a centralized fashion and executed in a decentralized manner. In other words, the agents can freely share information during the training phase, while each agent's policy must condition on its own observation during execution.

### 2.2 Value-based Learning Methods for MARL

IQL [11] is the simplest value-based learning method for MARL, where each agent attempts to maximize the total rewards individually. Such a method causes nonstationarity due to the changing policies of the other agents and may not converge. Thus, value function factorization methods (e.g., VDN and QMIX) are introduced to enable centralized training of factorizable tasks [12] based on the IGM condition [12], where optimal individual actions result in the optimal joint actions. In this work, we focus on extending QMIX.

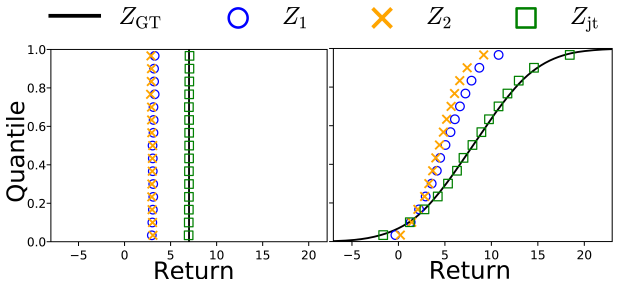
### 2.3 Distributional Reinforcement Learning

A number of distributional RL methods are proposed in the single-agent RL (SARL) domain to generalize expected RL methods. In [5], the distributional Bellman operator is proved to be a contraction in $p$-Wasserstein distance. Based on the proof, the authors further introduced C51 [5] to approximate the PMF of return distributions by categorical distributions. In [6], quantile regression (QR) is proposed to approximate the quantile function of return distributions, while ensuring the contraction theoretically. IQN [7] further improves the data efficiency of QR-DQN [6] by forming an implicit distribution through randomly selecting quantile samples. Distributional RL methods have been proved empirically to outperform expected RL methods in various environments [5–9], and enable additional improvements [13–15] that require the information of full return distributions.




(a) Factorized PMFs of $\mathcal{N}(7, 0)$. (b) Factorized PMFs of $\mathcal{N}(8, 29)$.

**Figure 1: (a) and (b) illustrate the value function factorization of different states. The black line/curve shows the true return PMFs. The blue bars denote agent 1's learned utilities, while the orange bars depict agent 2's learned utilities. The green bars indicate the estimation of the joint return.**




(a) Factorized CDFs of $\mathcal{N}(7, 0)$. (b) Factorized CDFs of $\mathcal{N}(8, 29)$.

**Figure 2: (a) and (b) illustrate the value function factorization of different states. The black line/curve shows the true return CDFs. The blue bars denote agent 1's learned utilities, while the orange bars depict agent 2's learned utilities. The green bars indicate the estimation of the joint return.**

## 3 DECOMPOSING THE MEAN AND SHAPE

The procedure of applying distributional RL to value function factorization methods must satisfy the IGM condition for the correctness of CTDE. Given an arbitrary factorization function (e.g., monotonic network in QMIX), the IGM condition may be violated when modeling the stochasticity in individual utilities. In order to enable the use of distributional RL methods while maintaining the IGM condition, we propose the decomposition of the mean and the shape of joint return distributions. Given a stochastic joint return $Z_{jt}$, it can be decomposed as follows:

$$\begin{aligned} Z_{jt} &= \mathbb{E}[Z_{jt}] + (Z_{jt} - \mathbb{E}[Z_{jt}]) \\ &= Z_{mean} + Z_{shape}\,, \end{aligned}$$

where $\mathrm{Var}(Z_{mean}) = 0$ and $\mathbb{E}[Z_{shape}] = 0$. We decompose a joint return $Z_{jt}$ into its deterministic part $Z_{mean}$ (i.e., the expected value) and stochastic part $Z_{shape}$ (i.e., the higher moments), which are approximated by two different functions $\Psi$ and $\Phi$, respectively. To satisfy IGM, the factorization function $\Psi$ is required to precisely factorize the joint expectation $\mathbb{E}[Z_{jt}]$ into individual expectations $[\mathbb{E}[Z_k]]_{k\in\mathbb{K}}$. On the other hand, the shape function $\Phi$ is allowed to roughly factorize the shape of $Z_{jt}$ into shapes of $[Z_k]_{k\in\mathbb{K}}$, since the main objective of modeling the return distribution is to assist nonlinear approximation of the joint expectation $\mathbb{E}[Z_{jt}]$, rather than accurately model the shape of $Z_{jt}$ [16]. Under this formulation, $\Psi$ can be selected from existing value function factorization methods (e.g., VDN or QMIX), and $\Phi$ can be defined as 1D convolution or quantile mixture. Based on the chosen $\Psi$ and $\Phi$, individual utilities can be approximated by distributional RL methods (e.g., C51 and IQN), while maintaining the IGM condition for CTDE.

## 4 EXPERIMENTAL RESULTS

In this section, we present two possible shape decomposition methods based on QMIX, and show the visualized results. We choose two different implementations of approximating the shape. The first implementation uses C51 to approximate the probability mass functions (PMFs) of individual utilities, and combine their shapes through 1D convolution. The second implementation uses IQN to approximate the quantile functions of individual utilities, and combine their shapes through quantile mixture. We focus on two terminal states in our environment, with rewards sampled from normal distributions: $\mathcal{N}(7, 0)$ and $\mathcal{N}(8, 29)$. Fig. 1 illustrates the factorized PMFs of the two states under the first implementation, while Fig. 2 illustrates the factorized cumulative distribution functions (CDFs) of the two states under the second implementation.

In Figs. 1-2, we can observe that both implementations can successfully factorize the joint return distributions into individual utilities. The deterministic rewards in the first state can be modeled accurately in Fig. 2 (a) since IQN models the distribution implicitly, while there are some approximation errors observed in Fig. 1 (a) due to the limited categories of C51 when modeling the PMF. The stochastic rewards in the second state can be modeled accurately in both Fig. 1 (b) and Fig. 2 (b).

## 5 DISCUSSION

In more complex scenarios, joint return distributions may be multimodal distributions instead of normal distributions. In such cases, the shape of the distributions may not be modeled as precisely as in Figs. 1-2 due to the use of 1D convolution and quantile mixture. Nevertheless, the expectation can still be modeled accurately due to the decomposition of mean and shape. As for the representable tasks, the distributional variant share the same set of representable tasks as its base value function factorization method (in our case, QMIX). Therefore, value function factorization methods proposed in the future can be similarly extended to its distributional variant under our formulation.

## 6 CONCLUSION

In this paper, we provided a distributional perspective on value function factorization methods, and decompose the approximation of the joint return's mean and shape to ensure the IGM condition holds. Then, we propose two practical implementations based on QMIX to demonstrate the ability to factorize return distributions. The decomposition concept can be extended to more value function factorization methods, and can be extended to factorize all factorizable tasks. Please refer to the full paper [17] for the detailed descriptions and the formal proofs of our method.

## REFERENCES


[1] P. Sunehag *et al.* Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proc. Int. Conf. on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 2085–2087, May 2018.

[2] T. Rashid *et al.* QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 4295–4304, Jul. 2018.

[3] T. Rashid *et al.* Monotonic value function factorisation for deep multi-agent reinforcement learning. *J. Machine Learning Research (JMLR)*, 21(178):1–51, Jan. 2020.

[4] M. Samvelyan *et al.* The starcraft multi-agent challenge. In *Proc. Int. Conf. on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 2186–2188, May 2019.

[5] M. G. Bellemare, W. Dabney, and R. Munos. A distributional perspective on reinforcement learning. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 449–458, Jul. 2017.

[6] W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos. Distributional reinforcement learning with quantile regression. In *Proc. AAAI Conf. on Artificial Intelligence (AAAI)*, pages 2892–2901, Feb. 2018.

[7] W. Dabney, G. Ostrovski, D. Silver, and R. Munos. Implicit quantile networks for distributional reinforcement learning. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 1096–1105, Jul. 2018.

[8] M. Rowland *et al.* Statistics and samples in distributional reinforcement learning. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 5528–5536, Jul. 2019.

[9] D. Yang *et al.* Fully parameterized quantile function for distributional reinforcement learning. In *Proc. Conf. Advances in Neural Information Processing Systems (NeurIPS)*, pages 6190–6199, Dec. 2019.

[10] F. A. Oliehoek and C. Amato. *A Concise Introduction to Decentralized POMDPs*. *Springer*, 2016.

[11] M. Tan. Multi-agent reinforcement learning: Independent versus cooperative agents. In *Proc. Int. Conf. on Machine Learning (ICML)*, page 330–337, Jun. 1993.

[12] K. Son, D. Kim, W. J. Kang, D. E. Hostallero, and Y. Yi. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 5887–5896, Jul. 2019.

[13] N. Nikolov, J. Kirschner, F. Berkenkamp, and A. Krause. Information-directed exploration for deep reinforcement learning. In *Proc. Int. Conf. on Learning Representations (ICLR)*, May 2019.

[14] S. Zhang and H. Yao. Quota: The quantile option architecture for reinforcement learning. In *Proc. AAAI Conf. on Artificial Intelligence (AAAI)*, pages 5797–5804, Feb. 2019.

[15] B. Mavrin, H. Yao, L. Kong, K. Wu, and Y. Yu. Distributional reinforcement learning for efficient exploration. In *Proc. Int. Conf. on Machine Learning (ICML)*, pages 4424–4434, Jul. 2019.

[16] C. Lyle, M. G. Bellemare, and P. S. Castro. A comparative analysis of expected and distributional reinforcement learning. In *Proc. AAAI Conf. on Artificial Intelligence (AAAI)*, pages 4504–4511, Feb. 2019.

[17] W.-F. Sun, C.-K. Lee, and C.-Y. Lee. DFAC framework: Factorizing the value function via quantile mixture for multi-agent distributional q-learning. *arXiv preprint arXiv:2102.07936*, Feb. 2021.