Projection-Optimal Monotonic Value Function Factorization in Multi-Agent Reinforcement Learning

Extended Abstract

Yongsheng Mei
The George Washington University
Washington, DC, USA
ysmei@gwu.edu

Hanhan Zhou
The George Washington University
Washington, DC, USA
hanhan@gwu.edu

Tian Lan
The George Washington University
Washington, DC, USA
tlan@gwu.edu

ABSTRACT

Value function factorization has emerged as the prevalent method for cooperative multi-agent reinforcement learning under the centralized training and decentralized execution paradigm. Many of these algorithms ensure the coherence between joint and local action selections for decentralized decision-making by factorizing the optimal joint action-value function using a monotonic mixing function of agent utilities. Despite this, utilizing monotonic mixing functions also induces representational limitations, and finding the optimal projection of an unconstrained mixing function onto monotonic function classes remains an open problem. In this paper, we propose QPro, which casts this optimal projection problem for value function factorization as regret minimization over projection weights of different transitions. This optimization problem can be relaxed and solved using the Lagrangian multiplier method to obtain the optimal projection weights in a closed form, where we narrow the gap between optimal and restricted monotonic mixing functions by minimizing the policy regret of expected returns, thereby enhancing the monotonic value function factorization. Our experiments demonstrate the effectiveness of our method, indicating improved performance in environments with non-monotonic value functions.

KEYWORDS

Multi-agent Reinforcement Learning; Value Function Factorization; Optimization

ACM Reference Format:
Yongsheng Mei, Hanhan Zhou, and Tian Lan. 2024. Projection-Optimal Monotonic Value Function Factorization in Multi-Agent Reinforcement Learning: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

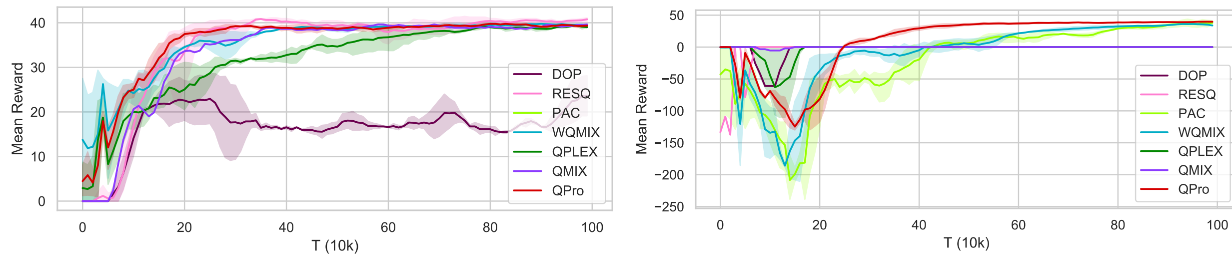
In this paper, we propose QPro, formulating the optimal projection problem for value function factorization as regret minimization over the projection weights of different state-action values. Our method involves constructing an optimal policy based on the optimal joint action-value function and a restricted policy using its projection onto monotonic mixing functions. We then define policy regret as the difference between the expected discounted reward of the optimal policy and that of the restricted policy. By minimizing this policy regret through an upper bound, we can minimize the gap between the optimal and restricted policies, leading to an optimal monotonic factorization with minimum regret. Our proposed regret minimization problem can be solved using the Lagrangian method considering an upper bound. We derive the optimal projection weights in closed form by examining a weighted Bellman equation involving monotonic mixing functions and per-agent critics and leveraging the implicit function theorem [3] and Karush-Kuhn-Tucker conditions [1]. Our results shed light on the key principles that contribute to optimal monotonic value function factorization. The optimal projection weights consist of four components: Bellman error, value underestimation, the gradient of the monotonic mixing function, and the on-policiness of available transitions. We note that the first two components are consistent with the weighting heuristics proposed in WQMIX [8] and provide a quantitative justification for this method. Furthermore, our analysis shows that an optimal value function factorization should also consider the gradient of the monotonic mixing function and the positive impact of more current transitions.

2 BACKGROUND

Partially Observable Markov Decision Process. In decentralized partially observable Markov decision process (Dec-POMDP) [7], the task is a tuple $G = \langle S, U, P, R, Z, O, n, \gamma \rangle$, where $s \in S$ describes the global state of the environment. Every time, each agent $a \in A \equiv \{1, \dots, n\}$ selects an action $u_a \in U$, and all selected actions are combined to form a joint action $\boldsymbol{u} \in \mathbf{U}$, which causes a transition in the environment based on the state transition function $P(s'|s, \boldsymbol{u}) : S \times \mathbf{U} \times S \rightarrow [0, 1]$. All agents share the same reward function $r(s, \boldsymbol{u}) : S \times \mathbf{U} \rightarrow \mathbb{R}$ with a discount factor $\gamma \in [0, 1)$. In the partially observable environment, the agents' individual observations $z \in Z$ are generated by the observation function $O(s, u) : S \times A \rightarrow Z$. Each agent has an action-observation history $\tau_a \in T \equiv (Z \times U)^*$, and the policy $\pi^a(u_a|\tau_a) : T \times U \rightarrow [0, 1]$ is conditioned on the history. The joint policy $\boldsymbol{\pi}$ has a joint action-value function: $Q^{\boldsymbol{\pi}}(s_t, \boldsymbol{u}_t) = \mathbb{E}_{s_{t+1:\infty}, \boldsymbol{u}_{t+1:\infty}}[R_t|s_t, \boldsymbol{u}_t]$, where t is the timestep and $R_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i}$ is the discounted return.

Regret of Expected Returns. Regret has been widely adopted in many existing works [2, 4, 6]. In the MARL context, the objective is to find a joint policy $\boldsymbol{\pi}$ that can maximize the expected return: $\eta(\boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\pi}}[\sum_{i=0}^{\infty} \gamma^i r_{t+i}]$. For a fixed policy, the Markov decision process becomes a Markov reward process, where the discounted

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).



(a) No punishment

(b) Punishment $= -2$

Figure 1: Average reward per episode on the Predator-Prey tasks for QPro and other baseline algorithms.

state distribution is defined as $d^{\pi}(s)$. Similarly, the discounted state-action distribution is defined as $d^{\boldsymbol{\pi}}(s, \boldsymbol{u}) = d^{\boldsymbol{\pi}}(s)\boldsymbol{\pi}(\boldsymbol{u}|s)$. Thus, we write the expected return $\eta(\boldsymbol{\pi})$ as $\frac{1}{1-\gamma}\mathbb{E}_{d^{\boldsymbol{\pi}}(s,\boldsymbol{u})}[r(s, \boldsymbol{u})]$. We assume there exists an optimal joint policy $\boldsymbol{\pi}^*$ such that $\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi}} \eta(\boldsymbol{\pi})$. The regret of the joint policy $\boldsymbol{\pi}$ is defined as $\eta(\boldsymbol{\pi}^*) - \eta(\boldsymbol{\pi})$. The policy regret measures the expected loss when following the current policy $\boldsymbol{\pi}$ instead of optimal policy $\boldsymbol{\pi}^*$. Since $\eta(\boldsymbol{\pi}^*)$ is a constant, minimizing the regret is consistent with maximizing of expected return $\eta(\boldsymbol{\pi})$. By minimizing the regret, the current policy $\boldsymbol{\pi}_k$ following a monotonic value factorization will approach the optimum $\boldsymbol{\pi}^*$ following an unrestricted value function.

3 OPTIMAL PROJECTION ONTO MONOTONIC VALUE FUNCTIONS

3.1 Problem Formulation

Let Q^* be the unrestricted joint action value and $f_s(Q^1, \ldots, Q^n)$ be its estimation obtained through a monotonic mixing function $f_s(\cdot)$ of per-agent utilities Q^a for $a = 1, \ldots, n$. To formulate the regret with respect to this projection, we consider a Boltzmann policy $\boldsymbol{\pi}_k$ following the agent's individual utilities Q_k^a at step k obtained from such monotonic value factorization, as well as a similar policy $\boldsymbol{\pi}^*$ following the unrestricted value function Q^* defined over joint actions. Our objective is to minimize the regret $\eta(\boldsymbol{\pi}^*) - \eta(\boldsymbol{\pi})$ over non-negative projection weights under relevant constraints, i.e.,

$$
\begin{aligned}
\min_{w_k} \quad & \eta(\boldsymbol{\pi}^*) - \eta(\boldsymbol{\pi}_k) \\
\text{s.t.} \quad & (Q_k^1, \ldots, Q_k^n) = \\
& \quad \underset{(Q^1,\ldots,Q^n)\in\mathcal{Q}}{\arg\min} \; \mathbb{E}_{\mu}[w_k(s, \boldsymbol{u})(f_s(Q^1, \ldots, Q^n) - \mathcal{B}^* Q_{k-1}^*)^2], \\
& \boldsymbol{\pi}_k = \{\pi_k^a\}_{a=1}^n, \quad \pi_k^a = \frac{\exp(Q_k^a(\tau_a, u_a))}{\sum_{\tau_a, u_a'} \exp(Q_k^a(\tau_a, u_a'))}, \\
& \mathbb{E}_{\mu}[w_k(s, \boldsymbol{u})] = 1, \quad w_k(s, \boldsymbol{u}) \geq 0
\end{aligned}
$$

3.2 Optimal Projection Weights

We can solve the proposed regret minimization problem and obtain optimal projection weights in closed form in Theorem 1. The proof is provided in the full version of our paper [5].

Theorem 1 (Optimal weighting scheme). *The optimal weight $w_k(s, \boldsymbol{u})$ to a relaxation of the regret minimization problem with discrete action space is given by:*

$$w_k(s, \boldsymbol{u}) = \frac{1}{Z^*}(E_k(s, \boldsymbol{u}) + \epsilon_k(s, \boldsymbol{u})),$$

where when $Q_k \leq \mathcal{B}^* Q_{k-1}^*$, we have:

$$E_k(s, \boldsymbol{u}) = \frac{d^{\boldsymbol{\pi}_k}(s, \boldsymbol{u})}{\mu(s, \boldsymbol{u})}(\mathcal{B}^* Q_{k-1}^* - Q_k) \exp(Q_{k-1}^* - Q_k) \left(\sum_{j=1}^n \frac{1 - \pi^j}{f'_{s,Q^j}} - 1 \right),$$

and when $Q_k > \mathcal{B}^* Q_{k-1}^*$, we have $E_k(s, \boldsymbol{u}) = 0$, where Z^* is the normalization factor, and $\epsilon_k(s, \boldsymbol{u})$ is a negligible term.

The theoretical results shed light on the key factors determining an optimal projection onto monotonic mixing functions. Specifically, when the Bellman error $\mathcal{B}^* Q_{k-1}^* - Q_k$ of a particular transition is high indicating a wide gap, we consider assigning a larger weight to it. Similarly, value underestimation $\exp(Q_{k-1}^* - Q_k)$ works as a correction term for incoming transitions, which compensates the underestimated Q_k with larger importance while penalizing over-estimated Q_k with a smaller weighting modifier. Additionally, our analysis identifies two new terms: the gradient of the monotonic mixing function $\sum_{j=1}^n (1 - \pi^j) f'^{-1}_{s,Q^j} - 1$ and measurement of on-policy transitions $d^{\boldsymbol{\pi}_k}(s, \boldsymbol{u})\mu(s, \boldsymbol{u})^{-1}$, which are crucial in obtaining an optimal projection onto monotonic value function factorization.

4 EXPERIMENTS

Predator-Prey. We present results in Predator-Prey environment as the demonstration. Figure 1 shows the performance of seven algorithms with two punishments, where all results demonstrate the superiority of QPro over others. Besides, regarding efficiency, we can spot that QPro has the fastest convergence speed in seeking the best policy. In Figure 1b, QPro significantly outperforms other algorithms in a hard setting requiring a higher level of coordination among agents as learning the best policy with improved joint action representation is required in this setting. Most algorithms, such as QMIX [9], ResQ [10], and DOP [11], end up learning a sub-optimal policy where agents learn to work together with limited coordination. Although QPro and WQMIX [8] acquired good results eventually, compared to the latter, QPro achieves better performance and converges to the optimal policy profoundly faster.

ACKNOWLEDGMENTS

This research is based on work supported by the Office of Naval Research under grants N00014-23-1-2850 and N00014-20-1-2146.

REFERENCES

[1] Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. 2021. KKT Conditions, First-Order and Second-Order Optimization, and Distributed Optimization: Tutorial and Survey. *arXiv preprint arXiv:2110.01858* (2021).

[2] Peter Jin, Kurt Keutzer, and Sergey Levine. 2018. Regret minimization for partially observable deep reinforcement learning. In *International conference on machine learning*. PMLR, 2342–2351.

[3] Steven George Krantz and Harold R Parks. 2002. *The implicit function theorem: history, theory, and applications.* Springer Science & Business Media.

[4] Xuhui Liu, Zhenghai Xue, Jingcheng Pang, Shengyi Jiang, Feng Xu, and Yang Yu. 2021. Regret Minimization Experience Replay in Off-Policy Reinforcement Learning. *Advances in Neural Information Processing Systems* 34 (2021), 17604–17615.

[5] Yongsheng Mei, Hanhan Zhou, and Tian Lan. 2023. Projection-Optimal Monotonic Value Function Factorization in Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:2302.05593.*

[6] Yongsheng Mei, Hanhan Zhou, Tian Lan, Guru Venkataramani, and Peng Wei. 2023. MAC-PO: Multi-Agent Experience Replay via Collective Priority Optimization. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 466–475.

[7] Frans A Oliehoek and Christopher Amato. 2016. *A concise introduction to decentralized POMDPs.* Springer.

[8] Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. 2020. Weighted QMIX: Expanding Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. arXiv:2006.10800 [cs.LG]

[9] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*. PMLR, 4295–4304.

[10] Siqi Shen, Mengwei Qiu, Jun Liu, Weiquan Liu, Yongquan Fu, Xinwang Liu, and Cheng Wang. 2022. ResQ: A Residual Q Function-based Approach for Multi-Agent Reinforcement Learning Value Factorization. *Advances in Neural Information Processing Systems* 35 (2022), 5471–5483.

[11] Yihan Wang, Beining Han, Tonghan Wang, Heng Dong, and Chongjie Zhang. 2020. Dop: Off-policy multi-agent decomposed policy gradients. In *International Conference on Learning Representations*.