ResQ: A Residual Q Function-based Approach for Multi-Agent Reinforcement Learning Value Factorization

Anonymous Author(s)
Affiliation
Address
email

Abstract

The factorization of state-action value functions for Multi-Agent Reinforcement Learning (MARL) is important. Existing studies are limited by their representation capability, sample efficiency, and approximation error. To address these challenges, we propose, ResQ, a MARL value function factorization method, which can find the optimal joint policy for any state-action value function through residual functions. ResQ masks some state-action value pairs from a joint state-action value function, which is transformed as the sum of a main function and a residual function. ResQ can be used with mean-value and stochastic-value RL. We theoretically show that ResQ can satisfy both the individual global max (IGM) and the distributional IGM principle without representation limitations. Through experiments on matrix games, the predator-prey, and StarCraft benchmarks, we show that ResQ can obtain better results than multiple expected/stochastic value factorization methods.

1 Introduction

Many real-world tasks involve multiple agents acting together, such as robot control [1]. Such tasks can be modelled as Multi-Agent Reinforcement Learning (MARL) problems where a group of agents must cooperate to achieve a common goal. MARL has attracted great research interest because of its social and economic impact. However, MARL is highly stochastic and hard to learn due to the partial-observability [2] and changing policies of agents. To address these issues, many approaches adopt the Centralized Training with Decentralized Execution paradigm (CTDE) [3].

In the CTDE paradigm, value factorization approaches [4] are widely adopted due to their remarkable performance and sample efficiency. A joint state-action value function is factorized into per-agent utilities Q_i, and each agent i acts greedily according to Q_i. To enable effective CTDE, it is critical to ensure the Individual-Global-Max (IGM) principle [5] that the optimal joint action should be equivalent to the collection of each agent's greedy actions. VDN [6] and QMIX [4] are two popular value factorization methods, which satisfy the IGM principle.

Most value factorization approaches focus on the deterministic state-action value functions. Albeit MARL is highly stochastic, the full distribution of per-agent utilities and the multi-agent system are overlooked and represented only as expected value of the full distribution. Such distributional information could be beneficial for policy learning. DMIX and DDN [7] extend QMIX and VDN with distributional RL. They can factorize stochastic joint state-action value pairs into stochastic per-agent utilities, and satisfy the distributional individual-global-max(DIGM) principle [7]. However, because of the representation limitations of QMIX and VDN, DMIX and DDN cannot factorize distributional value functions whose mean value is non-monotonic.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Agent 2

Agent 1

u_1 \ u_2	A	B	C
A	8	-12	-12
B	-12	0	0
C	-12	0	7.999

Q_{jt}

=

u_1 \ u_2	A	B	C
A	8	-12	-12
B	0	0	0
C	0	0	0

Q_{tot}

+

u_1 \ u_2	A	B	C
A	0	0	0
B	-12	0	0
C	-12	0	7.999

Q_r

Figure 1: Motivating Example of ResQ. The reward (state-action value function) of a one-step game, denoted as $Q_{jt}(u_1, u_2)$. The game consists of two agents; each has three actions (A, B, and C). The optimal policy is to choose action A for both two agents.

To address the representation limitations, QTRAN [5], WQMIX [8], and QPLEX [9] have been proposed. However, QTRAN and QPLEX are sample-inefficient, and WQMIX has high approximation errors to non-optimal actions. These methods may lead to incorrect estimations of optimal values. Moreover, they satisfy only the IGM principle. Achieving the IGM and the DIGM principles with low approximation errors and high sample efficiency remains an open challenge.

We address these challenges through, *ResQ*, a residual Q function-based approach. Given a hard-to-factorize joint state-action value function Q_{jt} (e.g., shown in Fig. 1), the key insight of our work is described as follows. We could obtain an easy-to-factorize state-action value function, the main function Q_{tot}, if we can mask out some state-action value pairs from Q_{jt}. If the main function Q_{tot} shares the same optimal policy with Q_{jt}, then we can obtain the optimal policy of Q_{jt} through the factorization of Q_{tot} by just using some simple mixers (e.g., [6]). After the masking, Q_{jt} is divided into two value functions: a main function Q_{tot} and a residual function Q_r. Q_r are the masked state-action value pairs from Q_{jt}.

Besides mean-value RL (e.g., Q_{jt}), the idea of ResQ can be used with distributional RL [10] as well. ResQ can factorize stochastic joint state-action value functions Z_{jt} into per-agent stochastic utilities Z_i with a mask and a residual function Z_r. We have shown that ResQ satisfies both *the IGM and the DIGM principles* without representation limitations. Moveover, ResQ can track joint state-action value pairs closely; thus, it could have low approximation errors.

For evaluation, we conduct extensive experiments on one-step matrix games, the StarCraft II MARL tasks [11], and predator-prey tasks. The experimental results show that ResQ can obtain better results than multiple competitive value factorization methods, and the ablation study shows that through the use of residual functions for multiple factorization methods, ResQ can improve the performance of these methods.

2 Background

2.1 Dec-POMDPs

We consider cooperative Multi-Agent Reinforcement Learning (MARL) scenarios which can be modelled as Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs) [12] defined as G for n agents. $G = \langle \mathcal{S}, \{\mathcal{U}_i\}_{i=1}^n, P, r, \{\mathcal{O}_i\}_{i=1}^n, \{\sigma_i\}_{i=1}^n, n, \gamma \rangle$. $\mathcal{S}$ denotes the set of states and $\mathcal{U}_i$ the set of actions available to agent i and we consider discrete actions only. A joint action of all agents is defined as $\boldsymbol{u}^t \in \mathcal{U}^N := \mathcal{U}_1 \times \ldots \times \mathcal{U}_n$. At a discrete time step t and state s^t, after the joint action is issued, the next state $s^{t+1} \in \mathcal{S}$ of the environment is drawn from the transition function $s^{t+1} \sim P(\cdot|s^t, \boldsymbol{u}^t)$. All the agents receive a reward r^t after the state transition happens. It is a partially observable environment that agent i observes a part of the environment $o_i^t \in \mathcal{O}_i$ which is drawn from $o_i^t \sim \sigma^i(\cdot|s^t)$. γ is the discount factor. Each agent i maintains an action-observation history $\tau_i = (O_i \times U_i)^{*1}$, and acts according to policy $\pi_i(u_i|\tau_i)$. We denote $\tau \in \mathcal{T}^N := \tau_1 \times \ldots \times \tau_n$ as the joint-observation history. The learning objective is to find the optimal policy $\pi =< \pi_1, ..., \pi_n >$.

2.2 Value Function Factorization

The Individual-Global-Max (**IGM**) principle proposed in [5] is important to realize the factorization of MARL state-action value function. It is defined as follows.

Definition 1 (IGM)**.** *For a joint state-action value function $Q_{\text{jt}} : \mathcal{T}^N \times \mathcal{U}^N \mapsto \mathbb{R}$, where $\tau \in \mathcal{T}^N$ is a joint action-observation history, if there exist individual state-action functions $[Q_i : \mathcal{T}_i \times \mathcal{U}_i \mapsto \mathbb{R}]_{i=1}^N$,*

[1] $*$ represents 0 to T, where T denotes the time step. $*$ means that the history could be short or long. For example, τ_1 could $\in O_1^1 \times U_1^1 \times O_1^2 \times U_1^2$ and could $\in O_1^1 \times U_1^1$, where the superscript represent time, and the subscript is the index of a agent.

such that the following conditions are satisfied

$$\arg\max_{\mathbf{u}} Q_{\mathrm{jt}}(\boldsymbol{\tau}, \mathbf{u}) = (\arg\max_{u_1} Q_1(\tau_1, u_1),\ \ldots,\ \arg\max_{u_n} Q_n(\tau_n, u_n)), \tag{1}$$

then, $[Q_i]$ *satisfy IGM for* Q_{jt} *under* τ*. We can state that* $Q_{\mathrm{jt}}(\boldsymbol{\tau}, \mathbf{u})$ *is factorized by* $[Q_i(\tau_i, u_i)]_{i=1}^N$*. In this work, we assume the argmax operator is unique, the action with smallest index is selected to break ties if a tie exists*

QMIX is a widely used monotonically increasing mixer function which proposes the following sufficient conditions for IGM:

$$\text{(Monotonicity)} \quad \partial Q_{jt}(\boldsymbol{\tau}, \boldsymbol{u}) / \partial Q_i(\tau_i, u_i) \geq 0, \quad \forall i \in \mathcal{N}. \tag{2}$$

There are tasks whose joint state-action value function cannot be expressed well by monotonic increasing conditions, as shown in Fig. 1.

2.3 Distributional RL

Distributional RL models the stochastic return of state-action pair through $Z(\boldsymbol{\tau}, \boldsymbol{u})$ explicitly, rather than estimating a single scalar Q-value. The distribution of return can be approximated through a categorical distribution [13] or a quantile function [10]. Implicit Quantile Function (IQN) [10] models the stochastic value function $Z(\tau, u)$ as a quantile function $F^{-1}(\tau, u|w)$, where F^{-1} is the generalized inverse cumulative distribution function (CDF), $w \in [0, 1]$ is a quantile sample. Distributional RL models full return distribution $Z(\boldsymbol{\tau}, \boldsymbol{u})$ instead of $Q(\boldsymbol{\tau}, \boldsymbol{u})$. IQN defined a distributional Bellman operator, and use it to updates its $Z(s, u)$. After applying the distributional Bellman operator on $Z(\boldsymbol{\tau}, \boldsymbol{u})$, its resulting $Z(s', u')$ remains in the same distribution as $Z(s, u)$. The loss $\rho_t^{w_i, w_j}$ for temporal-difference error $\delta_t^{w_i, w_j}$ is defined as Huber quantile regression loss, where w_i and w_j are two quantiles. During execution, the action with the largest expected return $\arg\max_u \mathbb{E}[Z(\tau, u)]$ is chosen.

MARL is highly stochastic, and distributional RL could be used to deal with the stochasticity of MARL. [7] proposes the Distributional Individual-Global-Max (**DIGM**) principle, which is defined as follows.

Definition 2 (DIGM)**.** *Given a set of stochastic individual state-action value function* $[Z_i(\tau_i, u_i)]_{i=1}^N$ *and a stochastic joint state-action value function* $Z_{jt}(\boldsymbol{\tau}, \boldsymbol{u})$*, if the following conditions are satisfied*

$$\arg\max_{\mathbf{u}} \mathbb{E}[Z_{jt}(\boldsymbol{\tau}, \boldsymbol{u})] = (\arg\max_{u_1} \mathbb{E}[Z_1(\tau_1, u_1)],\ \ldots,\ \arg\max_{u_n} \mathbb{E}[Z_n(\tau_n, u_n)]), \tag{3}$$

then, $[Z_i]$ *satisfy DIGM for* Z_{jt} *under* τ*. We can state that* $Z_{jt}(\boldsymbol{\tau}, \mathbf{u})$ *is distributionally factorized by* $[Z_i(\tau_i, u_i)]_{i=1}^N$*.*

3 Related Work

Value factorization approaches are widely adopted in MARL [14]. VDN [6] factorizes the value function Q_{tot} as the sum of per-agents' utility Q_i. QMIX [4] supports monotonic relationships among Q_i and Q_{tot}. During execution, each agent acts greedily according to Q_i. RMIX [15] learns return distribution of each agent, and integrates risk-sensitive RL with QMIX. [7] proposes DMIX and DDN, which extend QMIX and VDN with distributional RL through mean-shape decomposition. They satisfy the DIGM principle but suffer from the representation limitation of QMIX and VDN. They cannot model stochastic value functions whose mean-value are non-monotonic. QAtten [16] and REFIL [17] adopt the attention mechanisms to focus on certain agents/scenarios when factorizing value function. However, they cannot represent non-monotonic value functions well.

To address the representational limitations, WQMIX [8] prioritizes the estimation of the optimal state-action value. It assigns higher weights to the optimal joint actions when minimizing approximation errors. WQMIX can be viewed as a special version of ResQ. It pays attention to the optimal actions via masking out all the sub-optimal actions. The Q_r of WQMIX consists of all the sub-optimal state-action pairs, and the Q_{tot} models only the optimal ones. However, it assigns high learning priorities to Q_{tot}, which puts the learning of sub-optimal state-action pairs and Q_{jt} in trouble. Moreover, it is difficult to assign proper weights to obtain satisfactory performance.

QTRAN [5] and QTRAN++ [18] transform the joint state-action value function Q_{jt} into an easy-to-factorize one through a series of linear constraints. They can be viewed as using Q_{tot} to model the optimal actions, and constraining $Q_r \geq Q_{jt}$ for sub-optimal actions. QTRAN over-estimates sub-optimal actions, which may lead to non-optimal decisions, whereas ResQ can estimate correctly non-optimal state-action values.

QPlex decomposes Q_{jt} as the sum of a value function V_{tot} and a non-positive advantage function A_{tot}. The role of A_{tot} is similar to the main function Q_{tot} used in ResQ. QPLEX obtains agents' policies from the factorization of A_{tot}, which must be ≤ 0, but ResQ does not have such restrictions. Although QPLEX does not suffer from representation limitation issues, it may struggle to learn a good policy in complex tasks [8].

The concurrent work, Residual-Q-Network (RQN) [19], deals with MARL value factorization too. It extends QTran by adding an individual correction factor for each utility function to compute an adjusted utility function (individual Q value). And the sum of the adjust utility function is used as the transformed Q value Q_{tran} which is defined in QTran. RQN can be viewed as a special case of ResQ. It uses the sum of per-agent utility as the main function Q_{tot}, and the sum of individual correction factors as the residual function Q_r. In this work, Q_{tran} over-estimates the value of sub-optimal actions (see 6b of Theorem in the RQN paper).

MAVEN [20] deals with the inefficient exploration problem in MARL. Inefficient exploration problems could interact with the representation limitation problems. As it is stated in the MAVEN paper, this may push the algorithm towards sub-optimal policies. In MAVEN, value-based agents condition their behaviour based on a shared latent variable. And the latent variable is controlled by a hierarchical policy. MAVEN relies on the underlying value factorization method to decompose Q values. And it adopts QMIX as its value factorization method. ResQ can be used in MAVEN as its factorization method. We will compare ResQ with MAVEN in the related work in the new version of this work.

COMNet [21] and DIAL [22] study communication among agents. DCG [23] uses graphs for coordination. The communication method adopted by them could be used with ResQ to improve MARL performance. COPA [24] and NDQ [25] factorize the value function with communication. CDS [26] improves the effective cooperation of agents by decomposing value functions. Different from them, ResQ focus on value factorization only. There exist various MARL actor-critic methods, such as MADDPG [27], MAAC [28], and COMA [29]. ResQ is a value-based method. Other methods exist. For example, MAPPO [30] combines MARL with PPO [31]. UPDeT [32] proposes a MARL policy decoupling method. ResQ is orthogonal to them.

4 Method

Motivating Example Fig. 1 shows an example of a non-monotonic payoff matrix, which is hard-to-factorized. Agent 1 has three different action-value orderings, which depend on the action of Agent 2. Moreover, the second-best state-action value 7.999 is only slightly lower than the optimal value 8, which causes difficulties during learning for methods which overestimate non-optimal actions, such as QTRAN.

Given a hard-to-factorize state-action value function Q_{jt}, the key insight to ResQ is that if we can mask out some state-action value pairs from Q_{jt}, then we could obtain an easy-to-factorize value function Q_{tot}. As shown in Fig. 1, if we mask out those action value pairs colored in red, then the main function Q_{tot} can be factorized easily by mixers such as QAtten or QMIX. ResQ seeks to find a mask w, a main functions Q_{tot}, and a residual function Q_r which satisfy the following equation.

$$Q_{jt}(\boldsymbol{\tau}, \boldsymbol{u}) = w_{tot}(\boldsymbol{\tau}, \boldsymbol{u}) Q_{tot}(\boldsymbol{\tau}, \boldsymbol{u}) + w_r(\boldsymbol{\tau}, \boldsymbol{u}) Q_r(\boldsymbol{\tau}, \boldsymbol{u}) \tag{4}$$

where masks $w_{tot}(\boldsymbol{\tau}, \boldsymbol{u}),\ w_r(\boldsymbol{\tau}, \boldsymbol{u}) \in \{0, 1\}$. The main Q_{tot} shares the same greedy optimal policy as Q_{jt}. $Q_{tot}(\boldsymbol{\tau}, \boldsymbol{u})$ is factorized into per-agent utilities Q_i, and each agent selects its action greedily with respect to Q_i. There are many possibilities to construct the masks w_{tot} and w_r, in this work, we set w_{tot} to be constant 1, and focus on w_r.

We describe how ResQ factorizes an expected-value joint state-action value function Q_{jt} into per-agent utilities in Sec. 4.1, and how ResQ decomposes a stochastic joint state-action value function Z_{jt} into stochastic individual utilities in Sec. 4.2.

4.1 Residual Q

For a given join action-observation history $\boldsymbol{\tau}$, consider a factorizable joint state-action value function $Q_{jt}(\boldsymbol{\tau}, \boldsymbol{u})$ that can be expressed in the form as (5). Let $\bar{u_i} = \arg\max_{u_i} Q_i(\tau_i, u_i)$, $\bar{u} = [\bar{u_i}]_{i=1}^N$, Theorem 1 states the sufficient conditions for state-action value functions $[Q_i(\tau_i, u_i)]_{i=1}^N$ that satisfy the IGM principle for Q_{jt}.

Theorem 1. *A joint state-action function*

$$Q_{jt}(\boldsymbol{\tau}, \boldsymbol{u}) = Q_{tot}(\boldsymbol{\tau}, \boldsymbol{u}) + w_r(\boldsymbol{\tau}, \boldsymbol{u}) Q_r(\boldsymbol{\tau}, \boldsymbol{u}) \tag{5}$$

is factorized by $[Q_i(\tau_i, u_i)]_{i=1}^N$*, if* $Q_r(\boldsymbol{\tau}, \boldsymbol{u}) \leq 0$*,* $Q_{tot}(\boldsymbol{\tau}, \boldsymbol{u})$ *and* $[Q_i(\tau_i, u_i)]_{i=1}^N$ *satisfy the monotonicity conditions (2), and*

$$w_r(\boldsymbol{\tau}, \boldsymbol{u}) = \begin{cases} 0 & \boldsymbol{u} = \bar{u}, \quad (6a) \\ 1 & \boldsymbol{u} \neq \bar{u}, \quad (6b) \end{cases}$$

The proof of Theorem 1 is shown in the appendix. In Theorem 1, we have shown that for any factorizable value function Q_{jt} that can be expressed as (5), we can find a main function Q_{tot} that shares the same optimal policy as Q_{jt}. In Theorem 2, we show that for any joint state-action value function Q, we can find a joint state-action value function Q_{jt} that can be expressed as (5), and it has the same optimal policy as Q.

Theorem 2. *For any joint state-action function* $Q(\boldsymbol{\tau}, \boldsymbol{u})$*, we can find* $Q_{jt}(\boldsymbol{\tau}, \boldsymbol{u}) = Q_{tot}(\boldsymbol{\tau}, \boldsymbol{u}) + w_r(\boldsymbol{\tau}, \boldsymbol{u}) Q_r(\boldsymbol{\tau}, \boldsymbol{u})$ *that*

$$\bar{u} = \arg\max_u Q(\boldsymbol{\tau}, \boldsymbol{u}) = \arg\max_u Q_{jt}(\boldsymbol{\tau}, \boldsymbol{u}) \tag{7}$$

$$Q(\tag{8}$$

$$vetau, \boldsymbol{u}) = Q_{jt}(\boldsymbol{\tau}, \boldsymbol{u}) \quad \forall \boldsymbol{u} \neq \tag{9}$$

$$vebaru \tag{10}$$

$Q_{tot}(\boldsymbol{\tau}, \boldsymbol{u})$ *monotonically increases with* $[Q_i(\boldsymbol{\tau}, \boldsymbol{u})]_{i=1}^N$*,* $w_r(\boldsymbol{\tau}, \boldsymbol{u})$ *satisfies (6), and* $Q_r(\boldsymbol{\tau}, \boldsymbol{u}) \leq 0$*.*

The proof of Theorem 2 is provided in the appendix. From Theorem 2, for any joint state-action function $Q(\boldsymbol{\tau}, \boldsymbol{u})$, we can find $Q_{jt}(\boldsymbol{\tau}, \boldsymbol{u})$ that shares the same optimal policy as Q. Moreover, Q_{jt} can be expressed as (5). Combining Theorem 1 and 2, for any $Q(\boldsymbol{\tau}, \boldsymbol{u})$, we can find $[Q_i(\tau_i, u_i)]_{i=1}^N$ satisfy the IGM principle for $Q(\boldsymbol{\tau}, \boldsymbol{u})$ through a residual function $Q_r(\boldsymbol{\tau}, \boldsymbol{u}) \leq 0$, a main function $Q_{tot}(\boldsymbol{\tau}, \boldsymbol{u})$, and a mask function $w_r(\boldsymbol{\tau}, \boldsymbol{u}) \in \{0, 1\}$. The policies $[Q_i(\tau_i, u_i)]_{i=1}^N$ can be obtained through factorizing $Q_{tot}(\boldsymbol{\tau}, \boldsymbol{u})$. Furthermore, Theorem 2 shows that $Q_{jt}(\boldsymbol{\tau}, \boldsymbol{u})$ can track $Q(\boldsymbol{\tau}, \boldsymbol{u})$ closely $\forall \boldsymbol{u} \neq \bar{u}$. Thus, ResQ could have low approximation errors.

4.2 Distributional Residual Q

One promising way to deal with the MARL stochasticity is to model the return of state-action pairs as distributional values [10]. Researchers [7] have proposed a distributional MARL approach that satisfies the DIGM principle. However, it cannot model stochastic joint value functions whose value-expectation are non-monotonic. ResQ can address this issue through using $Z_{dmix}(\boldsymbol{\tau}, \boldsymbol{u})$ as the main function and Z_r as the residual function, where $Z_{dmix}(\boldsymbol{\tau}, \boldsymbol{u})$ is the factorization function of DMIX or DDN [7]. Z_{dimx} and Z_r are both stochastic value functions. Formula (11) in Theorem 3 describes how Z_{jt} is decomposed.

Theorem 3. *A stochastic joint state-action function*

$$Z_{jt}(\boldsymbol{\tau}, \boldsymbol{u}) = Z_{dmix}(\boldsymbol{\tau}, \boldsymbol{u}) + w_r(\boldsymbol{\tau}, \boldsymbol{u}) Z_r(\boldsymbol{\tau}, \boldsymbol{u}) \tag{11}$$

is factorized by $[Z_i(\tau_i, u_i)]_{i=1}^N$*, if* $Z_r(\boldsymbol{\tau}, \boldsymbol{u}) \leq 0$ *and* $w_r(\boldsymbol{\tau}, \boldsymbol{u}) = 0$ *when* $\boldsymbol{u} = \bar{u}$*, otherwise* 1*.* $\bar{u_i} = \arg\max_{u_i} \mathbb{E}[Z_i(\tau_i, u_i)]$ *,* $\bar{u} = [\bar{u_i}]_{i=1}^N$*,* $Z_{dmix}(\boldsymbol{\tau}, \boldsymbol{u}) = Z_{mean}(\boldsymbol{\tau}, \boldsymbol{u}) + Z_{shape}(\boldsymbol{\tau}, \boldsymbol{u})$*,* $\mathbb{E}[Z_{shape}(\boldsymbol{\tau}, \boldsymbol{u})] = 0$*,* $Q_i = \mathbb{E}[Z_i(\tau_i, u_i)]$*.* $Z_{mean}(\boldsymbol{\tau}, \boldsymbol{u})$ *is a monotonically increasing function with respect to* Q_i *or the sum of* $[Q_i]_1^N$*.*

Besides freeing [7] from representation limitations, ResQ can factorize Z_{jt} into individual stochastic utilities Z_i with a linear main function Z_{tot}, a residual function Z_r, and a mask function w_r. Theorem 4 states the sufficient condition for $[Z_i(\tau_i, u_i)]_{i=1}^N$ satify the DIGM principle.

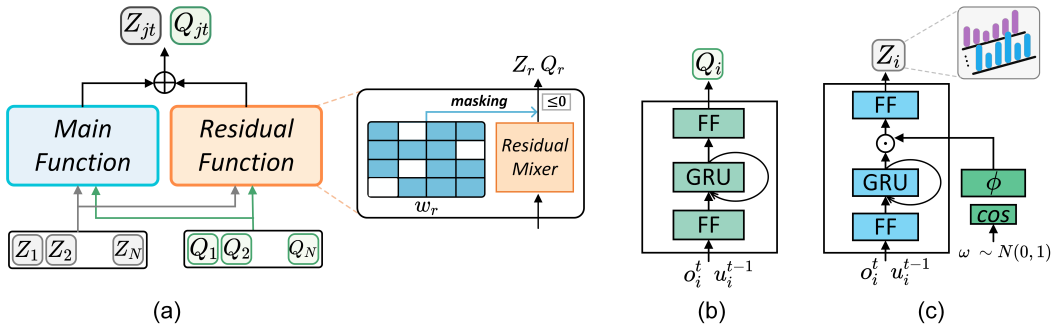

Figure 2: The architecture of ResQ. (a): the main and the residual function, (b): agent utility, (c): stochastic agent utility

Theorem 4. *A stochastic joint state-action function*

$$Z_{jt}(\boldsymbol{\tau}, \boldsymbol{u}) = Z_{tot}(\boldsymbol{\tau}, \boldsymbol{u}) + w_r(\boldsymbol{\tau}, \boldsymbol{u}) Z_r(\boldsymbol{\tau}, \boldsymbol{u}) \tag{12}$$

is factorized by $[Z_i(\tau_i, u_i)]_{i=1}^N$*, if* $Z_r(\boldsymbol{\tau}, \boldsymbol{u}) \leq 0$*,* $Z_{tot}(\boldsymbol{\tau}, \boldsymbol{u}) = \sum_{i=1}^N k_i Z_i(\tau_i, u_i)$ $k_i \geq 0$ *and* $w_r(\boldsymbol{\tau}, \boldsymbol{u}) = 0$ *when* $\boldsymbol{u} = \bar{u}$*, otherwise* 1*.* $\bar{u}_i = \arg\max_{u_i} \mathbb{E}[Z_i(\tau_i, u_i)]$*,* $\bar{u} = [\bar{u}_i]_{i=1}^N$*.*

The proofs of Theorem 3 and 4 are provided in the appendix. Theorem 4 states that ResQ can find individual stochastic utilities satisfying the DIGM principle for Z_{jt}. And it does not limit the representation capacity of stochastic joint state-action function Z_{jt}.

In (12), the main function Z_{tot} is a positive weighted sum of Z_i instead of a monotonic mixing individual utilities as in (5). This is because not all monotonic increasing functions can be used as the main function. Some monotonic increasing function (e.g., $Z_{tot} = \sum_{i=1}^N Z_i^2$) could lead to incorrect estimation of the optimal actions (please refer to the Appendix for an example).

4.3 Neural Networks

For any state-action value function Q_{jt}, if we can construct the main function Q_{tot}, the mask w_r, and the residual function Q_r as (5), then we can recover the optimal policy for Q_{jt}. Similarly, for stochastic function Z_{jt}, we can recover the optimal actions for Z_{jt} through the main function Z_{tot}, the residual function Z_r and the mask as (12). The neural network of ResQ is depicted in Fig. 2.

Main Function For Q_{jt}, $Q_{tot}(\boldsymbol{\tau}, \boldsymbol{u})$ can be any monotonically increasing function. The most well known monotonically increasing mixers of per-agent utility $[Q_i]_{i=1}^N$ are QMIX, VDN, and QAtten. In default, we use QMIX as the main function Q_{tot}. For the stochastic case, the stochastic main function $Z_{tot}(\boldsymbol{\tau}, \boldsymbol{u})$ can be a positive weighted sum of individual utilities $Z_i(\tau_i, u_i)$. $Z_{tot}(\boldsymbol{\tau}, \boldsymbol{u}) = \sum_{i=1}^N k_i Z_i(\tau_i, u_i)$ $k_i \geq 0$. Specifically, we model k_i by the attention mechanism [16].

Residual Function The residual function $Q_r(\boldsymbol{\tau}, \boldsymbol{u})$ is a feed-forward network that takes action-observation history τ, agents' utilities Q_i, and action $[u_i]_{i=1}^N$ as input. Unlike Q_{tot}, Q_r is not constrained to be monotonic or positive weighted-sum of utilities, thus allowing it to model a larger class of functions. For $[Q_i]_{i=1}^N$, Q_r is a simple feed-forward network which takes $[Q_i]_{i=1}^N$ as inputs, and the negative absolute function is the activation function of the last neural network layer to ensure $Q_r \leq 0$. For the stochastic case, the residual function Z_r is modeled as $-|\sum_{i=1}^N w_i Z_i|$.

Mask Function The mask $w_r(\boldsymbol{\tau}, \boldsymbol{u})$ requires knowing the maximal joint actions for the joint state-action value function. However, it is intractable to obtain the optimal actions without searching over the entire action space. For practical implementation, we use approximation, and w_r is defined as follows.

$$w_r(\boldsymbol{\tau}, \boldsymbol{u}) = \begin{cases} 0 & \boldsymbol{u} = \widetilde{u} \quad \text{(13a)} \\ 1 & \boldsymbol{u} \neq \widetilde{u} \quad \text{(13b)} \end{cases}$$

where $\widetilde{u} = \arg\max_u Q_{tot}(\tau, u)$ for the expectation case. For the stochastic case, $\widetilde{u} = \arg\max_u \mathbb{E}[Z_{tot}(\boldsymbol{\tau}, \boldsymbol{u})]$.

Loss For the expected-value case, the loss L of ResQ consists of two temporal-difference losses: L^* and L^{jt}. $L = L^* + L^{jt}$. $L^* = \sum_{k=1}^{b}(y^k - Q_{jt}(\tau^k, u^k, \theta))^2$, where $y^k = r + \gamma Q_{jt}(\tau^{k+1}, \widetilde{u}, \theta^-)$, $\widetilde{u} = [\widetilde{u_i}]_{i=1}^N$, $\widetilde{u_i} = \arg\max Q_i(\tau_i^{k+1}, u_i)$. θ^- are the parameters of a target network. L^{jt} measure the distance between Q_{jt} and $Q_{tot} + w_r Q_r$. It is defined as $L^{jt} = \sum(Q_{jt} - Q_{tot} - w_r Q_r)^2$. For the distributional case, $L = L^{z*} + L^{zjt}$. L^{z*} and L^{zjt} are similar to L^* and L^{jt}, respectively. L^{z*} and L^{zjt} use the pair-wise sampled temporal difference error instead of the standard temporal error, and the Huber quantile regression loss [10] rather than the mean square error loss.

5 Evaluation

We evaluate the performance of ResQ on one-step matrix games, the StarCraft II Multi-Agent Challenge benchmark (SMAC) [11], and predator-prey with multiple algorithms. The experimental results show that ResQ can obtain *better results* than state-of-the-art methods, and it satisfies both the *IGM and DIGM principle* together without representation limitation. The ablation study shows that ResQ can *improve the performance* of multiple value factorization methods through using the residual function. Please refer to the supplement materials for detailed setup and more results.

5.1 Experimental Setup

The studied algorithms are listed in Table 1. ResZ is the distributional version of ResQ factorized according to (12). DAtten, designed by us, integrates distributional RL [10] and the attention mechanism. Other algorithms are configured with their default parameters, and the number of rollouts is one. Each experiment is repeated at least 5 times with different seeds. The configuration of ResQ follows WQMIX [8].

Table 1: Studied Algorithms

Value type	Algorithms	# Alg.
Expected value	ResQ, QMIX [4], CW QMIX [8], OW QMIX [8], QTran [5], QTran++ [18], QPlex [9], QAtten [16], VDN [6]	9
Stochastic value	ResZ, DAtten, DMIX [7], DDN [7]	4

5.2 Matrix Game

We study the representation power of multiple methods for a matrix game (depicted in Fig. 1 and Table 2). All the algorithms are ran through a full exploration $\epsilon = 1$ for ϵ-greedy conducted over 20,000 steps. This setting guarantees the exploration of all possible actions.

As it is observed in Table 2, only ResQ/ResZ, and CW QMIX can obtain the optimal policy, DMIX, DDN, QTran, and QPlex learn the second-best policy, and OW QMIX learns a wrong policy. Albeit CW QMIX can find the optimal action, its learned Q_{jt} has a high approximation error, as it focuses on the optimal actions only. Further, we modify the matrix into a distributional matrix by adding a nomral distribution (mean = 0, std=1) value into all action-value. Again, ResQ/ResZ can find the optimal policy while most of the algorithms cannot. Detailed results are shown in the appendix. These results demonstrate the ability of ResQ to factorize difficult state-action value functions with low approximation errors and high sample efficiency. Moreover, we find that ResZ has a lower approximation error than ResQ, which suggests that using stochastic functions is helpful for estimating values.

5.3 StarCraft II Multi-Agent Challenge (SMAC)

In SMAC [11], two teams of agents fight against each other. For the SMAC tasks, we train each algorithm for 1 to 2 million steps, and for every 10,000 steps, the learned policies are evaluated. The test win rate, which measures the average win ratio of the agents controlled by MARL algorithms for each 32 test episodes, is reported.

As shown in Figure 3 (a), ResQ and ResZ achieve the best performance, and QMIX has the third-best results in the MMM2 scenario. The performance improvement over QMIX comes from the fact that ResQ/ResZ can model better the non-monotonic state-action function of the MMM2 scenario. For the MMM scenario, ResQ/ResZ has the best test win rate.

Table 2: Payoff matrix of a one-step matrix game and reconstructed Q_{jt}. Boldface means greedy actions. Red color indicates *wrongly* estimated optimal actions, whereas blue color represents the opposite.

u_1 \ u_2	**A**	B	C
A	**8**	-12	-12
B	-12	0	0
C	-12	0	7.999

(a) Game Payoff matrix.

Q_1 \ Q_2	**0.038 (A)**	-0.07 (B)	0.037 (C)
0.038(A)	**8.10**	-9.10	-10.01
-0.08(B)	-9.03	-6.90	-9.03
0.037(C)	-10.01	-9.03	-10.02

(b) ResQ: Q_1, Q_2, Q_{jt}

Z_1 \ Z_2	**1.34 (A)**	-0.75(B)	1.33(C)
1.34(A)	**8.01**	-1.19	-1.19
-0.75(B)	-1.20	-2.48	-3.98
1.33(C)	-1.20	-1.17	7.99

(c) ResZ: $\mathbb{E}[Z_{tot}], \mathbb{E}[Z_1], \mathbb{E}[Z_2]$

Q_1 \ Q_2	-6.33(A)	-0.08(B)	**0.03(C)**
-6.19(A)	-10.22	-9.71	-9.67
-0.07(B)	-9.59	-0.31	0.05
0.03(C)	-9.53	0.03	**7.71**

(d) DMIX: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	-6.21(A)	-0.19(B)	**1.48(C)**
-6.23(A)	-12.44	-6.40	-4.71
–0.19(B)	-6.43	-0.39	1.29
1.49(C)	-4.75	1.29	**2.97**

(e) DDN: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	1.587(A)	-0.024(B)	**1.589**(C)
-1.301(A)	13.08	-6.94	-6.95
-0.033(B)	-6.94	5.08	5.06
1.304(C)	-6.96	5.05	**13.06**

(f) QTran: $Q_1, Q_2, Q_{\mathrm{jt}}$

Q_1 \ Q_2	-0.90(A)	-52.28(B)	**0.12(C)**
-0.96(A)	7.48	-12.54	-12.04
-52.2(B)	-12.26	-0.54	-0.38
0.16(C)	-12.46	-0.51	**7.49**

(g) QPlex: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	**-0.03(A)**	-90.5(B)	-3.23(C)
-0.03(A)	**-1.07**	106.64	-1.95
-90.9(B)	17.13	390.93	17.22
-3.23(C)	-2.43	106.81	7.01

(h) CW QMIX: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	-0.087(A)	-35.27(B)	**0.305(C)**
0.36(A)	-9.15	-0.878	**0.856**
-34.26(B)	-3.81	45.30	-3.933
-0.15(C)	-2.61	-0.88	-6.64

(i) OW QMIX: Q_1, Q_2, Q_{jt}

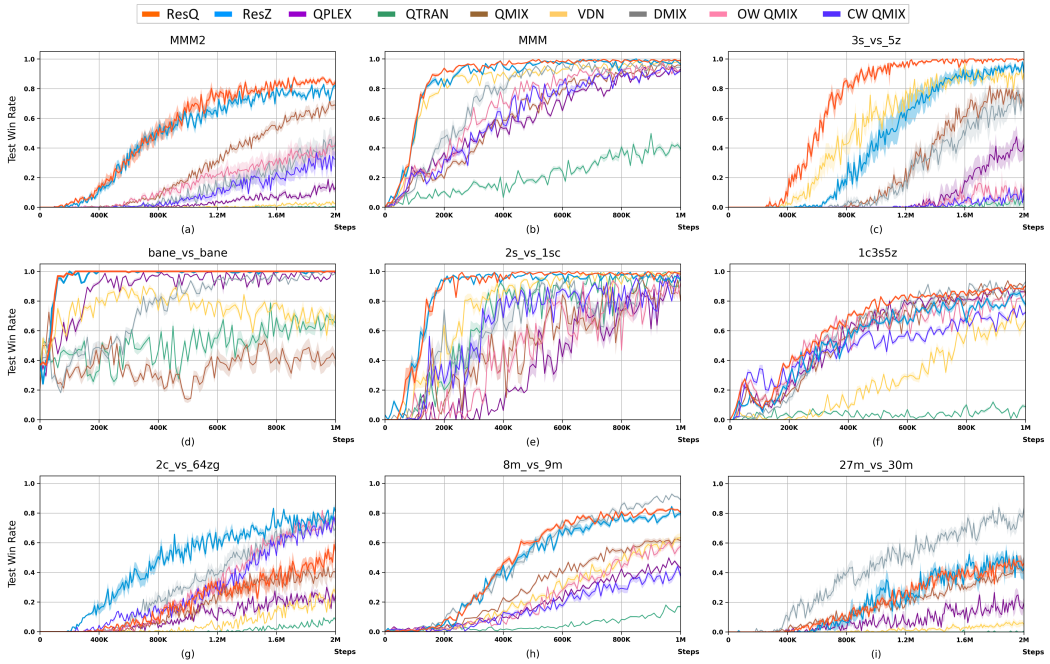

Figure 3: The Test Win Rate of the SMAC benchmarks. The horizontal axis shows the training time step.

The results for the 3s_vs_5z tasks are depicted in Fig. 3 (c). ResQ can obtain close-to-optimal performance after 1 million steps, while others cannot. For the bane_vs_bane scenario, in Fig. 3 (d), ResQ and ResZ can learn the optimal policy in less than 0.2 million steps, while the other methods cannot. The results indicate that through masking state-action value pair from Q_{jt}/Z_{jt}, ResQ/ResZ can learn the optimal policy in short time, and the performance is *stable*. Albeit QPlex and DIMX can obtain close-to-optimal performance at a much later time, but their performance is unstable. The results for CW/OW QMIX for bane_vs_bane and 27m_vs_30m are not shown due to out-of-memory error on an NVIDIA Geforce RTX 3090 GPU.

ResQ/ResZ can obtain the close-to-optimal performance after 0.25 million steps in the 2s_vs_1sc scenario, shown in Fig. 3 (e). Although the other methods can learn close-to-optimal policy with more training steps, their performance is not stable. For the 1c3s5z scenario, depicted in Figure 2 (f), ResQ/ResZ are among the best-performing algorithms.

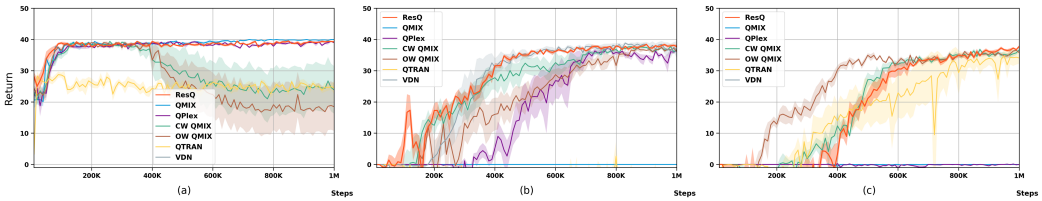

Figure 4: The return of Predator Prey with punishment: $p = 0$ (Left), $p = -2$ (Middle) and $p = -4$ (Right).

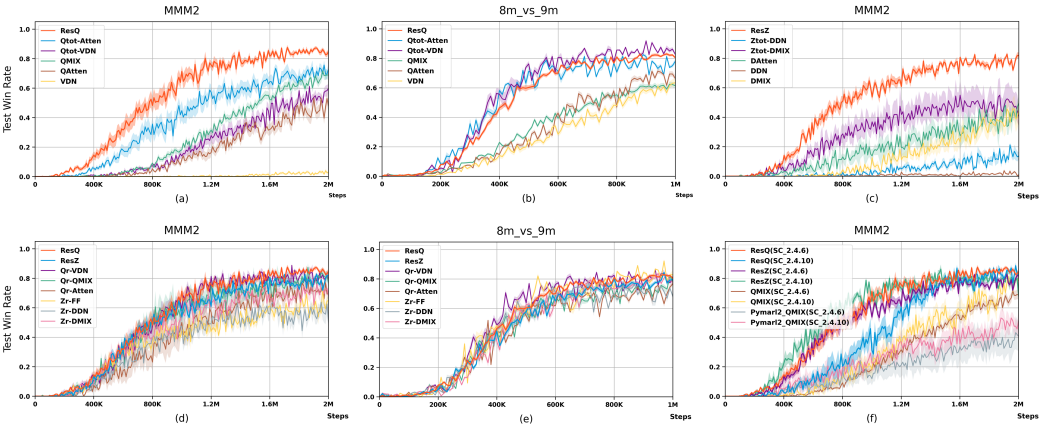

Figure 5: Impacts of the main functions (a), (b), (c), the residual functions (d) (e), and StarCraft versions (f)

In the 2c_vs_64zg scenario, 2 Colossi agents fights against 64 enemy agents. It has a large action space, and the reward of each action varies rather than is constant as in other scenarios. As it is depicted in Fig. 3 (g), ResZ can obtain the best performance. This indicates ResZ's ability to model distributional reward. The results for 8m_vs_9m and 27m_vs_30m are depicted in Fig. 3 (h) and (f), respectively. ResQ and ResZ are the second-best performing algorithms. DMIX is good at scenarios which consist of homogeneous agents (e.g., 8m_vs_9m), but it under-performs for scenarios with heterogeneous agents (e.g., MMM2, 3s_vs_5z), where agent interactions are complex.

5.4 Predator Prey

In the predator prey environment [23], 8 agents hunt 8 preys in a 10×10 grid world. Each agent has 6 actions: moves in either 4 directions, stands still, and catches prey. If two adjacent agents execute simultaneously the *catch* action to a nearby prey, a reward $r = 10$ is given to the agents, but a failed catch by a single agent is punished by a reward $p \leq 0$. The more negative p is, the higher level of coordination is needed for the agents. Following [8], the ϵ of the ϵ-greedy strategy for QPlex, ResQ, and CW/OW QMIX is gradually annealed from 1 to 0.05 within 1 million steps.

The test return for three punishments: $p = 0$, $p = -2$, and $p = -4$ are shown in Fig. 4. For the most easy scenario $p = 0$, ResQ, QMIX, and QPlex can learn the optimal policies quickly. The performance of CW/OW QMIX drops in the middle of training. For the difficult $p = -2$ case, ResQ can obtain the best returns. CW QMIX and OW QMIX have the second and third best scores. QTran and QMIX can not obtain any positive returns. For $p = -4$, ResQ can obtain the best returns in the end. These results indicate ResQ's ability for challenging cooperation scenarios.

5.5 Ablation Study

We study the impact of different implementations for the main function and the residual function, and Starcraft versions on the SMAC benchmark. We find that *ResQ can improve the performance of multiple value factorization approaches through the use of the residual function* Q_r.

Main Functions For ResQ, the results of using QMIX, QAtten, and VDN as the main function Q_{tot} are depicted as ResQ, Qtot-QAtten, Qtot-VDN in Fig. 5 (a) and (b). As it is shown, for the MMM2 and 8m_vs_9m scenarios, ResQ, Qtot-QAtten, and Qtot-VDN perform better than QMIX, QAtten, and VDN, respectively. This indicates that ResQ can *improve* the performance of value factorization

approaches (the main function) through the use of residual functions and mask. Moreover, we find that QMIX is the preferred main function.

For the distributional case, shown in Fig. 5 (c). The results of using DAtten, DDN, and DMIX as the main function Z_{tot} are depicted as ResZ, Ztot-DDN, and Ztot-DMIX. DAtten is designed by us; it integrates IQN [10] with QAtten. ResZ, Ztot-DDN, and Ztot-DMIX perform better than DAtten, DDN, and DMIX, respectively. This indicates that ResZ can *improve* the performance of stochastic value factorization methods through residual functions.

Residual Functions The results of using different mixers (factorization methods) for the residual functions are shown in the Fig. 5 (d) and (e). The performance of using a feed-forward network (FF) as Q_r (denoted as ResQ) is similar to that of using advanced mixers such as QMIX (Qr-QMIX) and VDN (Qr-VDN), and higher than that of QAtten (Qr-QAtten). In ResQ, we do not restrict the function classes of Q_r. This suggests that using FF may be sufficient for Q_r. For the distributional case, the performance of using DAtten, FF, DDN, DMIX is depicted as ResZ, Zr-FF, Zr-DDN, Zr-DMIX. We find that using DAtten as the residual function Z_r is preferred.

The Starcraft II version used in this work is SC2.4.6, which is the same version used as QMIX and WQMIX. Some methods (e.g.,[7]) uses the version SC2.4.10. We compare ResQ with Pymarl2_QMIX [14], which is a fine-tuned QMIX, across two versions. Figure 5 (f) shows that ResQ/ResZ perform better than QMIX and Pymarl2_QMIX for both two versions in the MMM2 scenario. As Pymarl2_QMIX performs better than most algorithms in most case for SC2.4.10 [14], this suggests that ResQ/ResZ may perform better than other algorithms in SC2.4.10.

6 Conclusion

In this work, we propose, ResQ, a residual function-based approach for Multi-Agent Reinforcement Learning (MARL) value function factorization. ResQ recovers the optimal policy for any joint state-action value function by masking out state-action value pairs from the value function. We show that ResQ can satisfy the individual-global-max (IGM) principle and the distributional IGM principle without representation limitations. Through extensive experiments on multiple MARL tasks, we show that ResQ can obtain promising results.

References

[1] Pablo Hernandez-Leal, Bilal Kartal, and Matthew E. Taylor. Is multiagent deep reinforcement learning the answer or the question? A brief survey. *CoRR*, abs/1810.05587, 2018.

[2] Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P. How, and John Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *ICML*, pages 2681–2690, 2017.

[3] Frans A. Oliehoek, Matthijs T. J. Spaan, and Nikos A. Vlassis. Optimal and approximate q-value functions for decentralized pomdps. *J. Artif. Intell. Res.*, 32:289–353, 2008.

[4] Tabish Rashid, Mikayel Samvelyan, Christian Schröder de Witt, Gregory Farquhar, Jakob N. Foerster, and Shimon Whiteson. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning. In *ICML*, 2018.

[5] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Hostallero, and Yung Yi. QTRAN: learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *ICML*, 2019.

[6] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinícius Flores Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *AAMAS*, 2018.

[7] Wei-Fang Sun, Cheng-Kuang Lee, and Chun-Yi Lee. DFAC framework: Factorizing the value function via quantile mixture for multi-agent distributional q-learning. In *ICML*, 2021.

[8] Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. Weighted QMIX: expanding monotonic value function factorisation for deep multi-agent reinforcement learning. In *NeurIPS*, 2020.

[9] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. QPLEX: duplex dueling multi-agent q-learning. In *ICLR*, 2021.

[10] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In Jennifer G. Dy and Andreas Krause, editors, *ICML*, volume 80, pages 1104–1113, 2018.

[11] Mikayel Samvelyan, Tabish Rashid, Christian Schröder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob N. Foerster, and Shimon Whiteson. The starcraft multi-agent challenge. In *AAMAS*, pages 2186–2188, 2019.

[12] Frans A. Oliehoek and Christopher Amato. *A Concise Introduction to Decentralized POMDPs*. Springer Briefs in Intelligent Systems. 2016.

[13] Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In Doina Precup and Yee Whye Teh, editors, *ICML*, volume 70, pages 449–458, 2017.

[14] Jian Hu, Siyang Jiang, Seth Austin Harding, Haibin Wu, and Shih-wei Liao. Rethinking the implementation tricks and monotonicity constraint in cooperative multi-agent reinforcement learning, 2021.

[15] Wei Qiu, Xinrun Wang, Runsheng Yu, Rundong Wang, Xu He, Bo An, Svetlana Obraztsova, and Zinovi Rabinovich. RMIX: learning risk-sensitive policies for cooperative reinforcement learning agents. In *NeurIPS*, pages 23049–23062, 2021.

[16] Yaodong Yang, Jianye Hao, Ben Liao, Kun Shao, Guangyong Chen, Wulong Liu, and Hongyao Tang. Qatten: A general framework for cooperative multiagent reinforcement learning. *CoRR*, abs/2002.03939, 2020.

[17] Shariq Iqbal, Christian A. Schröder de Witt, Bei Peng, Wendelin Boehmer, Shimon Whiteson, and Fei Sha. Randomized entity-wise factorization for multi-agent reinforcement learning. In *ICML*, 2021.

[18] Kyunghwan Son, Sungsoo Ahn, Roben Delos Reyes, Jinwoo Shin, and Yung Yi. Qtran++: Improved value transformation for cooperative multi-agent reinforcement learning, 2020.

[19] Rafael Pina, Varuna De Silva, Joosep Hook, and Ahmet Kondoz. Residual q-networks for value function factorizing in multi-agent reinforcement learning, journal = CoRR, volume = abs/2205.15245, year = 2022.

[20] Anuj Mahajan, Tabish Rashid, Mikayel Samvelyan, and Shimon Whiteson. MAVEN: multi-agent variational exploration. In *NeurIPS*, pages 7611–7622, 2019.

[21] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In *NeurIPS*, 2016.

[22] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *NeurIPS*, pages 2137–2145, 2016.

[23] Wendelin Böhmer, Vitaly Kurin, and Shimon Whiteson. Deep coordination graphs. *ICML*, 2020.

[24] Bo Liu, Qiang Liu, Peter Stone, Animesh Garg, Yuke Zhu, and Anima Anandkumar. Coach-player multi-agent reinforcement learning for dynamic team composition. In *ICML*, 2021.

[25] Tonghan Wang, Jianhao Wang, Chongyi Zheng, and Chongjie Zhang. Learning nearly decomposable value functions via communication minimization. In *ICLR*, 2020.

[26] Chenghao Li, Chengjie Wu, Tonghan Wang, Jun Yang, Qianchuan Zhao, and Chongjie Zhang. Celebrating diversity in shared multi-agent reinforcement learning. In *NeurIPS*, 2021.

[27] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *NeurIPS*, pages 6379–6390, 2017.

[28] Shariq Iqbal and Fei Sha. Actor-attention-critic for multi-agent reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *ICML*, 2019.

[29] Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *AAAI*, 2018.

[30] Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre M. Bayen, and Yi Wu. The surprising effectiveness of MAPPO in cooperative, multi-agent games. *CoRR*, abs/2103.01955, 2021.

[31] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

[32] Siyi Hu, Fengda Zhu, Xiaojun Chang, and Xiaodan Liang. Updet: Universal multi-agent reinforcement learning via policy decoupling with transformers. In *ICLR*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] The complete proofs are included in the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] in the supplemental material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] The standard errors are shown.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Each experiment is ran with a NVIDIA 3090 GPU.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We use some of the code from Pymarl, Pymarl2, and DFAC
   (b) Did you mention the license of the assets? [Yes] It is source code released by the authors of papers. Their code licensed under the Apache License v2.0.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The source code of this work is included in the supplemental material, and it will be open-sourced.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

Thanks for your time and interest to read this supplementary document. In Section A.1, we provide the ResQ theorems and proofs. Section A.2 describes the experimental setup in detail, shows another payoff matrix that cannot represent well by state-of-the-art algorithms, and depicts more results about the SMAC benchmark. The source code of ResQ is included in the supplementary materials.

A.1 Theorems

In this section, we show the proof of all the Theorems of this work.

Theorem 1. *A joint state-action function*

$$Q_{jt}(\tau, u) = Q_{tot}(\tau, u) + w_r(\tau, u) Q_r(\tau, u) \tag{1}$$

is factorized by $[Q_i(\tau_i, u_i)]_{i=1}^N$*, if* $Q_r(\tau, u) \leq 0$*,* $Q_{tot}(\tau, u)$ *and* $[Q_i(\tau_i, u_i)]_{i=1}^N$ *satisfy the monotonicity conditions (2), and*

$$w_r(\tau, u) = \begin{cases} 0 & u = \bar{u}, \quad \text{(2a)} \\ 1 & u \neq \bar{u}, \quad \text{(2b)} \end{cases}$$

Proof. Theorem 1 shows that if condition (2) holds and $Q_r(\tau, u) \leq 0$, then $[Q_i(\tau_i, u_i)]_{i=1}^N$ satisfy the IGM principle for Q_{jt}. We will show that $\arg\max_u Q_{\text{jt}}(\tau, u) = \bar{u}$, where $\bar{u}_i = \arg\max_{u_i} Q_i(\tau_i, u_i)$ and $\bar{u} = [\bar{u}_i]_{i=1}^N$.

$$\begin{aligned} Q_{jt}(\tau, \bar{u}) &= Q_{tot}(\tau, \bar{u}) \quad \text{(From (1)(2a))} && (3) \\ &\geq Q_{tot}(\tau, u) \quad \forall u \neq \bar{u} \quad \bar{u}\text{maximize} Q_{tot} && (4) \\ &\geq Q_{tot}(\tau, u) + w_r(\tau, u) Q_r(\tau, u) \quad \forall u \neq \bar{u} \quad \text{(From(2))} && (5) \\ &= Q_{\text{jt}}(\tau, u) \quad \forall u \neq \bar{u} \quad \text{(From (1))} && (6) \end{aligned}$$

(5) comes from $w_r(\tau, u) = 1 \; \forall u \neq \bar{u}$ and $Q_r(\tau, u) \leq 0$. (3) to (6) mean that $\bar{u} = [\bar{u}_i]_{i=1}^N$ maximizes $Q_{jt}(\tau, u)$. Thus $[Q_i(\tau_i, u_i)]_{i=1}^N$ satisfies IGM for $Q_{jt}(\tau, u)$. □

Lemma 1. *For any joint state-action function* $Q(\tau, u)$*, we can find a* Q_{tot} *that satisfy the monotonicity conditions, and they share the same optimal policy.*

Proof. Theorem 1 of [8] shows that Weighted QMIX can always find a monotonically increasing function Q_{tot} that shares the same optimal policy as the true value function of $Q(s, u)$. This indicates that for any $Q(s, u)$, there exists a monotonically increasing function $Q_{tot}(s, u)$ that shares the same optimal policy as $Q(s, u)$. □

Theorem 2. *For any joint state-action function* $Q(\tau, u)$*, we can find* $Q_{jt}(\tau, u) = Q_{tot}(\tau, u) + w_r(\tau, u) Q_r(\tau, u)$ *that*

$$\bar{u} = \arg\max_u Q(\tau, u) = \arg\max_u Q_{jt}(\tau, u) \tag{7}$$

$$Q(\tau, u) = Q_{jt}(\tau, u) \quad \forall u \neq \bar{u} \tag{8}$$

$Q_{tot}(\tau, u)$ *is a monotonic increasing function with respect to* $[Q_i(\tau, u)]_{i=1}^N$*,* $w_r(\tau, u)$ *satisfies the condition (2), and* $Q_r(\tau, u) \leq 0$*.*

Proof. Lemma (1) shows that for any joint state-action function $Q(\tau, u)$, we can find a Q_{tot} that satisfy the monotonicity conditions and $\bar{u} = \arg\max Q_{tot}(\tau, u) = \arg\max Q(\tau, u)$. Let

$$\Delta = Q(\tau, \bar{u}) - Q_{tot}(\tau, \bar{u}) \tag{9}$$

$$Q_{tot}^{'}(\tau, u) = Q_{tot}(\tau, u) + \Delta \quad \forall u \tag{10}$$

Obviously, the new function $Q_{tot}^{'}(\tau, u)$ satisfy the monotonic conditions, and $\arg\max_u Q(\tau, u) = \arg\max_u Q_{tot}^{'}(\tau, u)$. Define the residual Q function as

$$Q_r(\tau, u) = Q(\tau, u) - Q_{tot}^{'}(\tau, u) \tag{11}$$

Then

$$\begin{aligned} Q_r(\tau, \bar{u}) &= Q(\tau, \bar{u}) - Q_{tot}^{'}(\tau, \bar{u}) && (12)\\ &= Q(\tau, \bar{u}) - Q_{tot}(\tau, \bar{u}) - \Delta && (13)\\ &= 0 \quad \text{(From(9))} && (14) \end{aligned}$$

We can write $Q(\tau, u)$ as follows.

$$\begin{aligned} Q(\tau, u) &= Q_{tot}^{'}(\tau, u) + Q_r(\tau, u) && (15)\\ &= Q_{tot}^{'}(\tau, u) + w_r(\tau, u) Q_r(\tau, u) && (16) \end{aligned}$$

where

$$w_r(\tau, u) = \begin{cases} 0 & u = \bar{u}, \quad (17a)\\ 1 & u \neq \bar{u}, \quad (17b) \end{cases}$$

Let

$$Q_r^{'}(\tau, u) = \begin{cases} 0 & u = \bar{u}, \quad (18a)\\ Q_r(\tau, u) - \max_u Q_r(\tau, u) & u \neq \bar{u}, \quad (18b) \end{cases}$$

Through definition, $Q_r^{'}(\tau, u) \leq 0$. Let

$$Q_{tot}^{''}(\tau, u) = Q_{tot}^{'}(\tau, u) + \max_u Q_r(\tau, u) \tag{19}$$

Obviously, $Q_{tot}^{''}(\tau, u)$ satisfy the monotonic conditions with $[Q_i]_{i=1}^N$. $\arg\max_u Q(\tau, u) = \arg\max_u Q_{tot}^{'}(\tau, u) = \arg\max_u Q_{tot}^{''}(\tau, u)$. Combining (18) and (19), we can find a function

$$Q_{jt}(\tau, u) = Q_{tot}^{''}(\tau, u) + w_r(\tau, u) Q_r^{'}(\tau, u) \tag{20}$$

Thus, we show that for any joint state action function $Q_(\tau, u)$, it can find a function Q_{jt} that can be expressed in the form of ResQ (1), and share the same optimal actions with Q.

Further, we analysis the property of $Q_{jt}(\tau, u) \quad \forall u \neq \bar{u}$.

$$\begin{aligned} Q_{jt}(\tau, u) &= Q_{tot}^{''}(\tau, u) + w_r(\tau, u) Q_r^{'}(\tau, u) \quad \forall u \neq \bar{u} && (21)\\ &= Q_{tot}^{'}(\tau, u) + w_r(\tau, u) Q_r(\tau, u) \quad \forall u \neq \bar{u} && (22)\\ &= Q(\tau, u) \quad \forall u \neq \bar{u} && (23) \end{aligned}$$

Then, we analysis the property of $Q_{jt}(\tau, \bar{u})$.

$$\begin{aligned} Q_{jt}(\tau, \bar{u}) &= Q_{tot}^{''}(\tau, \bar{u}) + w_r(\tau, \bar{u}) Q_r^{'}(\tau, \bar{u}) && (24)\\ &= Q_{tot}^{''}(\tau, \bar{u}) && (25)\\ &= Q_{tot}^{'}(\tau, \bar{u}) + \max_u Q_r(\tau, u) && (26)\\ &= Q(\tau, \bar{u}) + \max_u Q_r(\tau, u) && (27) \end{aligned}$$

Combining (23) and (27), we reach the conclusion that the two functions $Q(\tau, u)$ and $Q_{jt}(\tau, u)$ differ only in the maximal values. And $Q_{jt}(\tau, u)$ can be expressed as the form of the addition of one monotonicity function $Q_{tot}(\tau, u)$ and a residual state-action function $Q_r(\tau, u)$ □

Theorem 3. *A stochastic joint state-action function*

$$Z_{jt}(\tau, u) = Z_{dmix}(\tau, u) + w_r(\tau, u) Z_r(\tau, u) \tag{28}$$

is factorized by $[Z_i(\tau_i, u_i)]_{i=1}^N$*, if* $Z_r(\tau, u) \leq 0$ *and* $w_r(\tau, u) = 0$ *when* $u = \bar{u}$*, otherwise* 1. $Z_{dmix}(\tau, u)$ *is the factorization method proposed in [7].* $Z_{dmix}(\tau, u) = Z_{mean}(\tau, u) + Z_{shape}(\tau, u)$*,* $\mathbb{E}[Z_{shape}(\tau, u)] = 0$*,* $Q_i = \mathbb{E}[Z_i(\tau_i, u_i)]$*.* $Z_{mean}(\tau, u)$ *is a monotonically increasing function with respect to* Q_i *or the sum of* $[Q_i]_1^N$*.*

Proof. We will show that $\arg\max_u \mathbb{E}[Z_{jt}(\tau, u)] = \bar{u}$, where $\bar{u}_i = \arg\max_{u_i} \mathbb{E}[Z_i(\tau_i, u_i)]$ and $\bar{u} = [\bar{u}_i]_{i=1}^N$.

$$\begin{aligned}
\mathbb{E}[Z_{jt}(\tau, \bar{u})] &= \mathbb{E}[Z_{dmix}(\tau, \bar{u})] \quad \text{(From (28) } w(\tau, \bar{u}) = 0) && (29)\\
&= \mathbb{E}[Z_{mean}(\tau, \bar{u}) + Z_{shape}(\tau, \bar{u})] && (30)\\
&= \mathbb{E}[Z_{mean}(\tau, \bar{u})] \quad \text{(From } \mathbb{E}[Z_{shape}(\tau, u)] = 0) && (31)\\
&= \mathbb{E}[Q_{mix}(\tau, \bar{u})] \quad \text{(Definition)} && (32)\\
&\geq \mathbb{E}[Q_{mix}(\tau, u)] \quad \forall u \neq \bar{u} \quad \text{(Mononicity relation)} && (33)\\
&= \mathbb{E}[Z_{mean}(\tau, u)] \quad \forall u \neq \bar{u} \quad \text{(Definition)} && (34)\\
&= \mathbb{E}[Z_{mean}(\tau, u) + Z_{shape}(\tau, u)] \quad \forall u \neq \bar{u} \quad \text{(From } \mathbb{E}[Z_{shape}(\tau, u)] = 0) && (35)\\
&= \mathbb{E}[Z_{dmix}(\tau, u)] \quad \forall u \neq \bar{u} \quad \text{(Definition)} && (36)\\
&\geq \mathbb{E}[Z_{dmix}(\tau, u)] + w_r(\tau, u)\mathbb{E}[Z_r(\tau, u)] \quad \forall u \neq \bar{u} && (37)\\
&= \mathbb{E}[Z_{dmix}(\tau, u) + w_r(\tau, u) Z_r(\tau, u)] \quad \forall u \neq \bar{u} && (38)\\
&= \mathbb{E}[Z_{jt}(\tau, u)] \quad \forall u \neq \bar{u} \quad \text{(From (28))} && (39)
\end{aligned}$$

(37) comes from $w_r(\tau, u) = 1 \ \forall u \neq \bar{u}$ and $Z_r(\tau, u) \leq 0$. (29) to (39) mean that $\bar{u} = [\bar{u}_i]_{i=1}^N$ maximizes $\mathbb{E}[Z_{jt}(\tau, u)]$. Thus $[Z_i(\tau_i, u_i)]_{i=1}^N$ satisfies DIGM for $Z_{jt}(\tau, u)$. □

Theorem 4. *A stochastic joint state-action function*

$$Z_{jt}(\tau, u) = Z_{tot}(\tau, u) + w_r(\tau, u) Z_r(\tau, u) \tag{40}$$

is factorized by $[Z_i(\tau_i, u_i)]_{i=1}^N$*, if* $Z_r(\tau, u) \leq 0$*,* $Z_{tot}(\tau, u) = \sum_{i=1}^N k_i Z_i(\tau_i, u_i) \ \ k_i \geq 0$ *and* $w_r(\tau, u) = 0$ *when* $u = \bar{u}$*, otherwise* 1.

Proof. Theorem 4 shows that if $Z_r(\tau, u) \leq 0$, $Z_{tot}(\tau, u) = \sum_{i=1}^N k_i Z_i(\tau_i, u_i) \ \ k_i \geq 0$ and $w_r(\tau, u) = 0$ when $u = \bar{u}$, otherwise 1, then $[Z_i(\tau_i, u_i)]_{i=1}^N$ satisfy the DIGM principle for $Z_{jt}(\tau, u)$. We will show that $\arg\max_u \mathbb{E}[Z_{jt}(\tau, u)] = \bar{u}$, where $\bar{u}_i = \arg\max_{u_i} \mathbb{E}[Z_i(\tau_i, u_i)]$ and $\bar{u} = [\bar{u}_i]_{i=1}^N$.

$$\begin{aligned}
\mathbb{E}[Z_{jt}(\tau, \bar{u})] &= \mathbb{E}[Z_{tot}(\tau, \bar{u})] \quad \text{(From (40) } w(\tau, \bar{u}) = 0) && (41)\\
&\geq \mathbb{E}[Z_{tot}(\tau, u)] \quad \forall u \neq \bar{u} && (42)\\
&\geq \mathbb{E}[Z_{tot}(\tau, u)] + w_r(\tau, u)\mathbb{E}[Z_r(\tau, u)] \quad \forall u \neq \bar{u} && (43)\\
&= \mathbb{E}[Z_{tot}(\tau, u) + w_r(\tau, u) Z_r(\tau, u)] \quad \forall u \neq \bar{u} && (44)\\
&= \mathbb{E}[Z_{jt}(\tau, u)] \quad \forall u \neq \bar{u} \quad \text{(From (40))} && (45)
\end{aligned}$$

Because $Z_{tot}(\tau, u) = \sum_{i=1}^N k_i Z_i(\tau_i, u_i) \, k_i \geq 0$, the inequality formula (42) is established. (43) comes from $w_r(\tau, u) = 1 \ \forall u \neq \bar{u}$ and $Q_r(\tau, u) \leq 0$. (44) comes from $w_r(\tau, u) Z_r(\tau, u) = \mathbb{E}[w_r(\tau, u) Z_r(\tau, u)]$. (41) to (45) mean that $\bar{u} = [\bar{u}_i]_{i=1}^N$ maximizes $\mathbb{E}[Z_{jt}(\tau, u)]$. Thus $[Z_i(\tau_i, u_i)]_{i=1}^N$ satisfies DIGM for $Z_{jt}(\tau, u)$. □

Example: Monotonic increasing function fails. In Theorem 4, the main function Z_{tot} is a positive weighted sum of Z_i instead of a monotonic mixing of Z_i, because not all monotonic mixing functions satisfy the DIGM principle (e.g., $Z_{jt} = Z_{tot} = \sum_{i=1}^N Z_i^2$), it could lead to incorrect estimation of the optimal actions. We show an example that $Z'_{tot} = Z_1^2 + Z_2^2$ fails. Let $Z_1(\tau, a) = 2$ for 100% of the time, and $Z_1(\tau, b) = 3$ for 50% of the time, $Z_1(\tau, b) = 0$ for 50% of the time. Clearly $\mathbb{E}Z_1(\tau, a) = 2$ and $\mathbb{E}Z_1(\tau, b) = 1.5$. Let $Z_2 = Z_1$. If Z_{jt} and Z_i satisfy the DIGM theorem, then the optimal action for Z_{jt} should be $(a, a) = (\arg\max \mathbb{E}Z_1(\tau, u), \arg\max \mathbb{E}Z_2(\tau, u))$. However, $\mathbb{E}Z_1^2(\tau, a) = 4$ and $\mathbb{E}Z_1^2(\tau, b) = 4.5$. $\arg\max Z_{jt} = (b, b)$ rather than (a, a)

Table 1: Payoff matrix of a didactic **distributional** matrix game and reconstructed Q_{jt}. Boldface means greedy actions. Red color indicates wrongly estimated optimal actions, whereas blue color represents the opposite.

u_1 \ u_2	A	B	C
A	**(8, 1)**	(-12,1)	(-12,1)
B	(-12,1)	(0,1)	(0,1)
C	(-12,1)	(0,1)	(7.999,1)

(a) Game Payoff matrix.

Q_1 \ Q_2	**0.001 (A)**	-0.003 (B)	0.0003 (C)
0.001(A)	**8.56**	-6.54	-6.53
-0.003(B)	-6.54	-6.54	-6.54
0.0003(C)	-6.53	-6.54	-6.53

(b) ResQ Q_1, Q_2, Q_{jt}

Z_1 \ Z_2	**0.81 (A)**	-1.07(B)	0.80(C)
0.81(A)	**8.53**	-11.61	-11.46
-1.07(B)	-11.44	0.43	0.49
0.80(C)	-11.23	0.46	8.46

(c) ResZ: $\mathbb{E}[Z_{tot}], \mathbb{E}[Z_1], \mathbb{E}[Z_2]$

Q_1 \ Q_2	-6.89(A)	-0.08(B)	**0.04(C)**
-7.12(A)	-10.1	-9.63	-9.58
-0.09(B)	-9.78	0.15	0.54
0.46(C)	-9.76	0.49	**8.84**

(d) DMIX: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	-6.27(A)	0.03(B)	**1.64(C)**
-6.29(A)	-12.56	-6.23	-4.62
0.04(B)	-6.25	-0.07	1.68
1.64(C)	-4.64	1.68	**3.29**

(e) DDN: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	4.19(A)	0.54(B)	**4.20**
4.22(A)	8.68	-11.22	-11.18
0.56(B)	-11.25	0.75	0.71
4.23(C)	-11.17	0.71	**8.52**

(f) QTran: $Q_1, Q_2, Q_{\mathrm{jt}}$

Q_1 \ Q_2	0.076(A)	-120(B)	**0.096(C)**
0.076(A)	-5.8	-20.76	-14.10
-120(B)	-20.14	-9.69	-9.90
0.096(C)	-14.45	-9.51	**-5.73**

(g) QPlex: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	**-0.17(A)**	-387.41(B)	-0.36(C)
0.19(A)	**-2.84**	2.24	-9.03
-431.8(B)	2.38	301.72	2.37
-0.39(C)	-9.28	2.25	0.51

(h) CW QMIX: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	-3.38(A)	-140(B)	**0.29(C)**
0.20(A)	-8.79	5.12	**-2.73**
-8.25(B)	1.02	5.24	1.03
-0.01(C)	-1.36	5.13	-6.64

(i) OW QMIX: Q_1, Q_2, Q_{jt}

A.2 Experimental Setup

We adopt the PyMARL [11] implementation of QMIX,VDN, QAtten, QTran, OW QMIX, CW QMIX, QPlex, and DMIX from their open-source repositories[2][3][4]. Their code are released under the Apache License V2.0. Their hyper-paramters are the same as that in PyMARL. ResQ is designed based on PyMARL as well. All the algorithms in this submission is trained with 1 rollout process for 2 million steps (the default setup for Weighted QMIX). At each 10,000 training step, the algorithm is tested on test episodes to evaluate their performance. For ResQ, the RMSProp optimizer ($l_r = 1e^{-3}$) is used. The batch size and buffer size are 32 and 5,000, respectively. Q_{tot} has 32 hidden dimensions, equiped with 2 hypernets, each with 64 dimensions. The ϵ used in $\epsilon-$greedy annealed from 1 to 0.05 within 100K steps. TD lambda is used, and $\lambda = 0.6$. All the experiments are ran within a local cluster. Each experiment is ran within a NVIDIA 3090 GPU.

A.2.1 A distributional one-step Matrix Game

We study the representation power of multipe methods for a distributional matrix game. The matrix game is modified from the matrix depicted in Figure 1. We add a nomral distribution (mean = 0, std=1) value into all action-value of the matrix. The payoff matrix of the game is depicted in Table 1 (a). All the cells show the mean and the variance of a normal distributed reward. For example, (8,1) indicates that the mean and the variance of the reward distribution are 8 and 1, respectively.

In this matrix game, all the algorithm are ran through a full exploration $\epsilon = 1$ for ϵ-greedy conducted over 20,000 steps. This setting guarantees the exploration of all possible actions.

As it is observed in Table 1, only ResQ/ResZ, and CW QMIX can obtain the optimal policy, DMIX, DDN, QTran, and QPlex learn the second-best policy, and OW QMIX learns a wrong policy. Albeit CW QMIX can find the optimal action, its learned Q_{jt} has a high approximation, as it focus on the optimal actions only.

These result demonstrates the ability of ResQ to factorize difficult state-action value functions with low approximation error and high sample efficiency. Moreover, we find that ResZ has lower approximation error than ResQ.

[2]https://github.com/oxwhirl/pymarl
[3]https://github.com/oxwhirl/wqmix
[4]https://github.com/wjh720/QPLEX

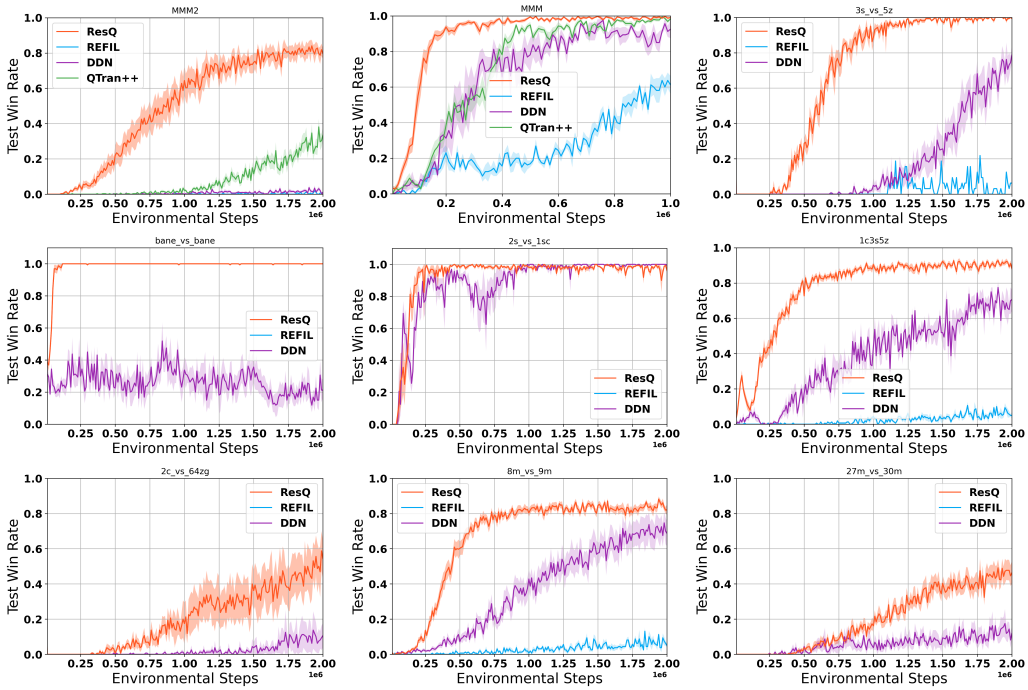

Figure 1: The test win rate of ResQ, DDN and REFIL for the SMAC benchmark.

A.2.2 StarCraft II Multi-Agent Challenge (SMAC)

We study the performance of ResQ and other algorithms on a well known MARL benchmark, the StarCraft II Multi-Agent Challenge (SMAC). In each environment of SMAC, two teams of agents fight against each other. One team is controlled by the built-in game artificial intelligence, which is handcrafted carefully. We choose the highest difficult built-in game script. Another team of agents are controlled by the decentralized policies learned by the MARL algorithms. Each agent can receive circular observation around the agent, and attack enemies close-by. In each episode, if the agents eliminate all the enemies within a limited time, the game episode is counted as won. Otherwise, the game is counted as a failure. For the SMAC tasks, the agents must learn the policies which maximize their expected total rewards. The reward of SMAC is based on the damage dealt, enemy agents killed, and the game won. We adopt the default SMAC reward schema.

The Starcraft II version used in ResQ is 2.4.6.2.69232, which is the same version used in QMIX and Weighted QMIX. For pymarl2, it used another version SC 2.4.10. As it is indicated in Weighted QMIX and confirmed by us, "*performance is not comparable across versions*". We have shown in Figure 5 (f), ResQ/ResZ perform better than the fine-tuned QMIX [14] for SC2.4.10 in the MMM2 scenario.

We have shown the results for ResQ, ResZ, QPLEX, QTran, QMIX, VDN, DMIX, OW QMIX, and CW QMIX. We have also evaluated the performance of DDN [7], REFIL [17], and QTran++. Their results are ploted in Figure 1. Due to the poor performance of REFIL, we only evaluate its performance on the MMM2, MMM, 3s_vs_5z, 1c3s5z, 8m_vs_9m scenarios. For QTran++, we only evaluate its performance on the MMM2 and MMM scenarios due to its poor performance and computation resource limitation.

Figure 2 depicts the results for three super-hard scenarios (Corridor, 3s5z_vs_3s6z, 6h_vs_8z) of the SMAC benchmark. AS it is shown in the figure, all algorithms perform poorly for these super-hard scenarios. This indicates that better exploration techniques should be used together with the value factorization methods.

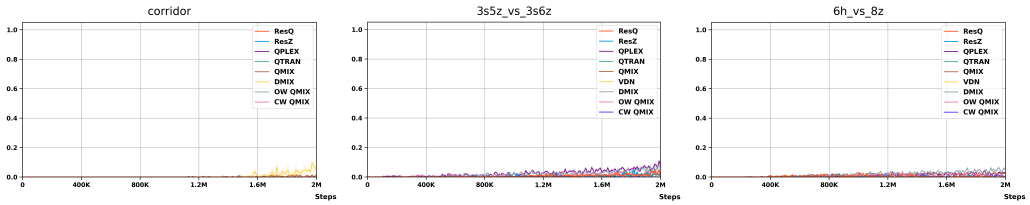

Figure 2: The test win rate of different algorithms for the SMAC benchmark: Corridor (Left), 3s5z_vs_3s6z (Middle), and 6h_vs_8z (Right).

Table 2: Payoff matrix of a one-step matrix game and reconstructed Q_{jt}. Boldface means greedy actions. Red color indicates *wrongly* estimated optimal actions, whereas blue color represents the opposite.

u_1 \ u_2	A	B	C
A	8	-12	-12
B	-12	0	0
C	-12	0	**7.9**

(a) Game Payoff matrix.

Q_1 \ Q_2	0.12 (A)	-0.39 (B)	0.11
0.12(A)	**8.06**	-7.8	0.51
-0.39(B)	-3.71	0.29	-3.72
0.11(C)	0.51	-4.2	0.50

(b) ResQ: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	**2.63(A)**	-0.07(B)	2.60(C)
3.08(A)	**10.29**	-9.71	-9.70
-0.02(B)	-9.71	2.28	2.29
3.04(C)	-9.70	2.29	10.18

(c) QTran: $Q_1, Q_2, Q_{\mathrm{jt}}$

Q_1 \ Q_2	-0.61(A)	-39(B)	**0.52(C)**
-0.51(A)	9.10	-10.22	-14.74
-39(B)	-12.1	1.97	1.39
0.54(C)	-15.79	1.69	**9.16**

(d) QPlex: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	**0.17(A)**	-4.80(B)	-4.80(C)
1.99(A)	**-1.67**	-0.98	-0.98
-53.2(B)	-21.47	49.2	49.1
-53.4(C)	1918	49.5	49.3

(e) CW QMIX: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	0.38(A)	-28.3(B)	**0.93(C)**
0.96(A)	3.10	-3.32	**7.72**
-33.1(B)	-5.70	68.5	-5.81
0.47(C)	0.26	-3.33	-0.38

(f) OW QMIX: Q_1, Q_2, Q_{jt}

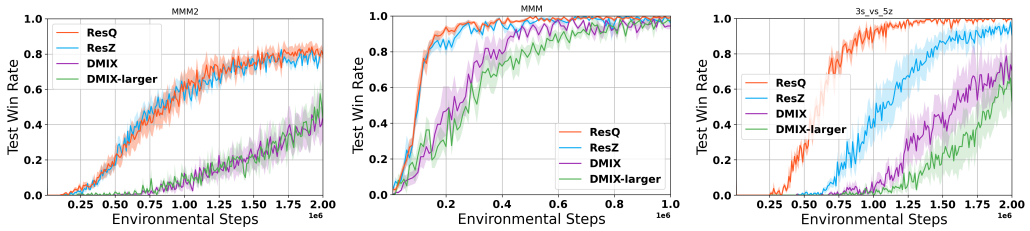

Figure 3: The test win rate of ResQ, ResZ, DMIX and DMIX-larger for the SMAC benchmark.

B Response to Reviewers

Table 2 depicts a non-monotonic pay-off matrix, ResQ, QTran, and CW QMIX can recover the optimal action. Table 3 shows another difficult pay-off matrix. In this matrix, only ResQ and QTran can recover the optimal action, the other algorithms cannot. Table 4 shows a difficult non-monotonic pay-off matrix, only ResQ recovers the optimal action.

ResQ and ResZ use more parameters than DMIX with the residual functions. By modifying the dimension of the last layer of the hypernet inside DMIX, we have increased the number of parameters of DMIX from 85K to 350K, which is bigger than that of ResQ (320K) and that of ResZ(316K). Let's denote DMIX with more parameters as DMIX-larger, and test its performance in the MMM2, the MMM, and the 3s_vs_5z scenarios. The experimental results are depicted in Figure 3. DMIX-larger performs slightly better than DMIX in the MMM2 scenario but worse than DMIX in the MMM and the 3s_vs_5z scenarios. ResQ and ResZ performs better than DMIX and DMIX-larger in these scenarios. This means that the performance improvement of ResQ over DMIX does not come from the use of more parameters.

To study the reason why ResQ performs better than QTran, We use two variants of ResQ: Qtot-VDN and Qtot-VDN-MSE. Qtot-VDN uses VDN as the main function same as QTran. The difference between Qtot-VDN between Qtran is the residual function, the mask, and the inequality conditions. ResQ requires $Q_r \leq 0$ and QTran requires that $Q_{tran}(\tau, u) \geq Q(\tau, u) \quad \forall u \neq \bar{u}$. Qtot-VDN-MSE uses MSE loss to implement $Q_r \leq 0$, and it uses VDN as its main function. Their results are depicted in Figure 4. For MMM2, Qtot-VDN can obtain a win rate of 0.6. The win rate for QTran is 0. For the

Table 3: Payoff matrix of a one-step matrix game and reconstructed Q_{jt}. Boldface means greedy actions. Red color indicates *wrongly* estimated optimal actions, whereas blue color represents the opposite.

u_1 \ u_2	**A**	B	C
A	2.5	0	-100
B	0	2	0
C	-100	-100	**3**

(a) Game Payoff matrix.

Q_1 \ Q_2	-0.27 (A)	-1.77 (B)	**0.29 (C)**
-0.34(A)	0.56	2.26	-14.6
-2.63(B)	1.91	-0.74	1.786
0.43(C)	-5.25	3.43	**2.96**

(b) ResQ: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	1.32(A)	1.24(B)	**1.39**(C)
1.67(A)	2.5	0.01	-99.40
1.56(B)	0.02	2.00	0.02
1.78(C)	-99.48	-99.72	**3.30**

(c) QTran: $Q_1, Q_2, Q_{\mathrm{jt}}$

Q_1 \ Q_2	1.98(A)	1.86(B)	**1.99(C)**
2.82(A)	1.80	-1.51	-5.78
2.83(B)	1.70	1.79	**1.80**
2.81(C)	-11.32	-11.32	1.79

(d) QPlex: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	**0.07(A)**	-0.35(B)	-156(C)
0.06(A)	**-70.41**	-34.57	723
-0.31(B)	-21.47	2.59	726
-134(C)	1918	1917	1805

(e) CW QMIX: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	**-0.13(A)**	-1.64(B)	-12.81(C)
6.44(A)	64.94	72.90	65.91
11.75(B)	**130.16**	141.05	154.39
-12.22(C)	17.15	40.16	206.21

(f) OW QMIX: Q_1, Q_2, Q_{jt}

Table 4: Payoff matrix of a one-step matrix game and reconstructed Q_{jt}. Boldface means greedy actions. Red color indicates *wrongly* estimated optimal actions, whereas blue color represents the opposite.

u_1 \ u_2	**A**	B	C
A	**8**	-12	-120
B	-12	7.8	7.7
C	-120	-130	7.9

(a) Game Payoff matrix.

Q_1 \ Q_2	3.93 (A)	3.92 (B)	3.93
3.97(A)	**7.95**	6.05	6.02
3.92(B)	5.88	6.01	5.98
3.93(C)	5.89	6.02	5.99

(b) ResQ: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	1.09(A)	1.11(B)	**1.12(C)**
4.64(A)	10.24	-9.58	-118
4.74(B)	-10.00	10.06	9.6
4.77(C)	-117	-127	**10.23**

(c) QTran: $Q_1, Q_2, Q_{\mathrm{jt}}$

Q_1 \ Q_2	0.56(A)	-1.68(B)	**4.62(C)**
1.71(A)	7.90	-12.39	-116
5.89(B)	-11.7	7.90	**7.91**
2.59(C)	-118	-128	7.89

(d) QPlex: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	**0.35(A)**	0.09(B)	-0.89(C)
0.77(A)	-78	-104	-124
1.90(B)	**-12.16**	-33.68	-111
-7.75(C)	64.2	59.1	38.3

(e) CW QMIX: Q_1, Q_2, Q_{jt}

Q_1 \ Q_2	**10.77(A)**	5.04(B)	3.24(C)
-6.58(A)	2.98	-81.0	-106
3.19(B)	**194**	103	54.4
-22.85(C)	-154	-51.6	-34.98

(f) OW QMIX: Q_1, Q_2, Q_{jt}

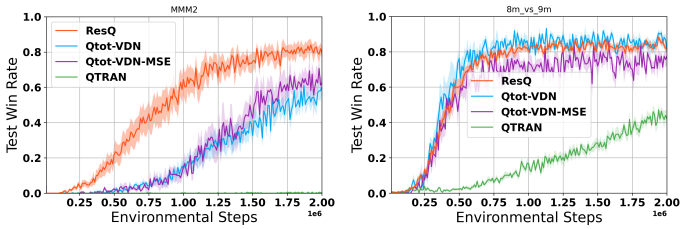

Figure 4: The test win rate of ResQ, Qtot-VDN, Qtot-VDN-MSE, and QTran for the SMAC benchmark.

8m_vs_9m scenario, Qtot-VDN can obtain a win rate of 0.8. And QTran can obtain a win rate of 0.4. The performance gap does not come from the implementation of the main function; it comes from the residual function, the mask function, and the inequality conditions. Further, Qtot-VDN-MSE performs similarly to Qtot-VDN, and it performs better than QTran. This suggests that the performance gap between ResQ and QTran does not come from the implementation of inequality conditions for ResQ.

ResQ does not significantly affect by inequality violations. We have evaluated the performance of a variant of ResQ, ResQ-MSE, which uses MSE loss to implement $Q_r \leq 0$. Further, we have studied the performance of ResQ-MSE in the MMM2, MMM, and 8m_vs_9m scenarios. The results are depicted in Figure 5. The performance of ResQ-mse and ResQ are similar in these three scenarios. This indicates that the violation of inequality conditions does not significantlyaffect the performance for these two scenarios.

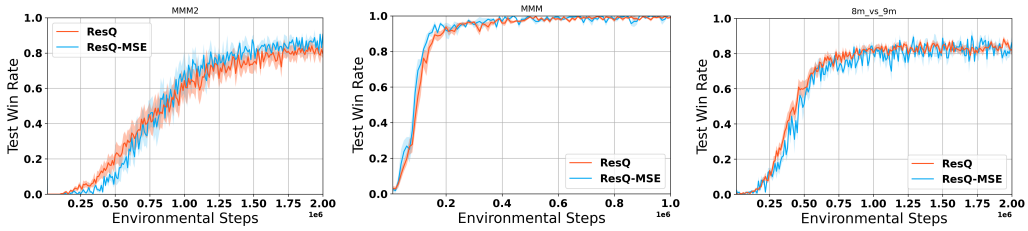

Figure 5: The test win rate of ResQ and ResQ-MSE for the SMAC benchmark.