RiskQ : Risk-sensitive Multi-Agent Reinforcement Learning Value Factorization

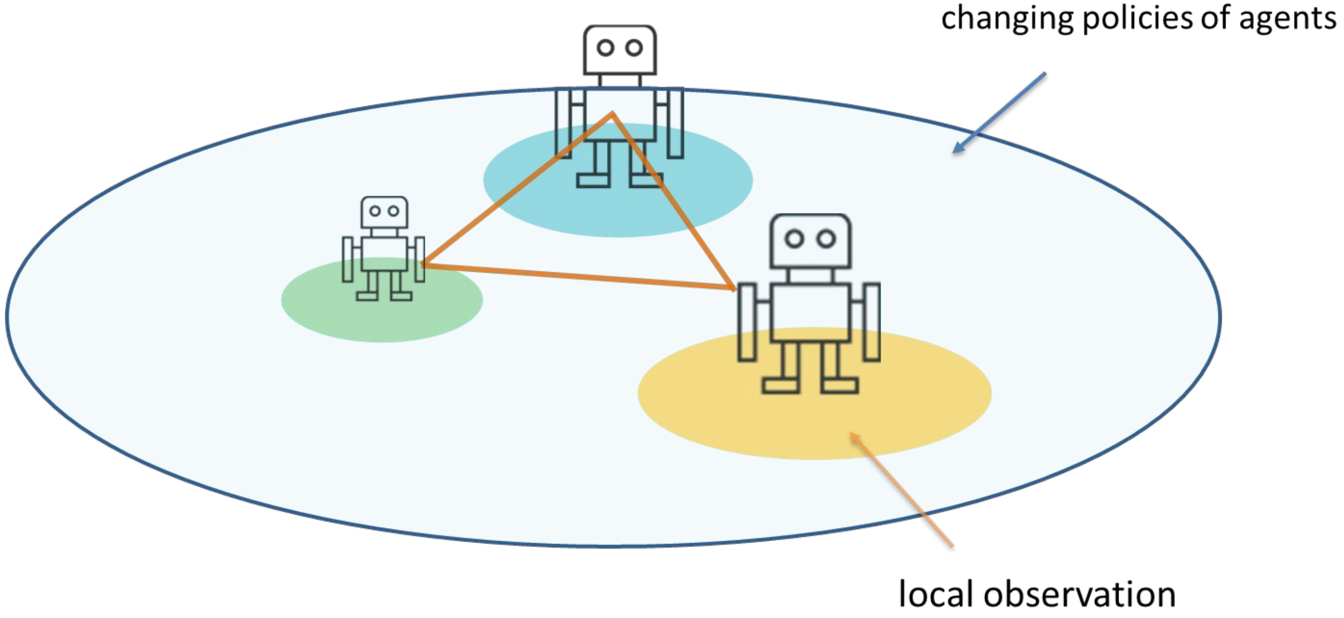
Siqi Shen†, **Chennan Ma**†, Chao Li†, Weiquan Liu†,
Yongquan Fu‡∗, Songzhu Mei‡, Xinwang Liu‡, Cheng Wang†

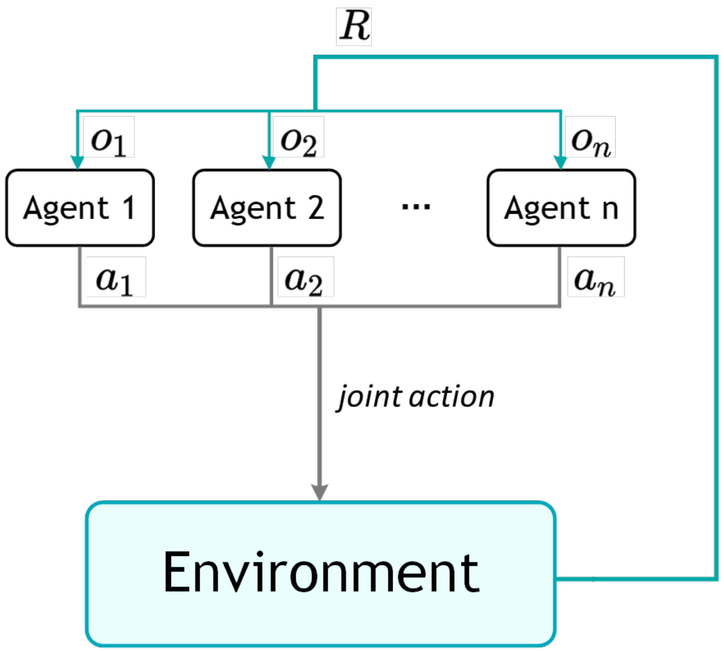
siqishen@xmu.edu.cn, chennanma@stu.xmu.edu.cn, chaoli@stu.xmu.edu.cn,
yongquanf@nudt.edu.cn

† Xiamen University
‡ National University of Defense Technology

NeurIPS 2023
https://arxiv.org/pdf/2311.01753.pdf

Challenges in MARL





Centralized Training with Decentralized Execution paradigm (CTDE)

Value Factorization

- ***Individual-Global-Max (IGM) principle***

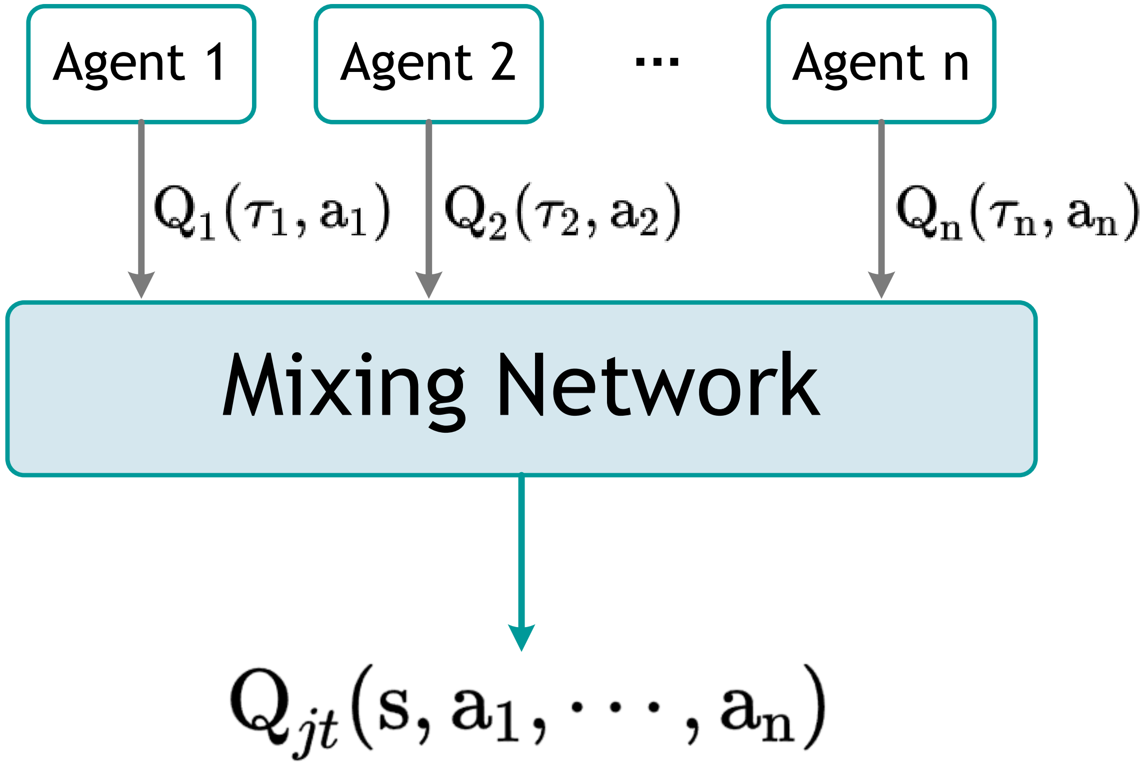
$$\arg\max_{\mathbf{u}} Q_{jt}(\tau, u) = \begin{pmatrix} \arg\max_{\mathbf{u}_1} Q_1(\tau_1, u_1) \\ \vdots \\ \arg\max_{\mathbf{u}_n} Q_n(\tau_n, u_n) \end{pmatrix}$$

VDN

$$Q_{\mathrm{jt}}(\boldsymbol{\tau}, \boldsymbol{u}) = \sum_{i=1}^{N} Q_i(\tau_i, u_i)$$

QMIX

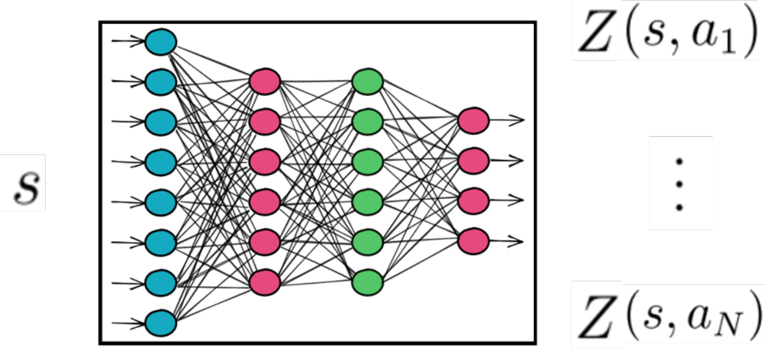
$$\frac{\partial Q_{\mathrm{jt}}(\boldsymbol{\tau}, \boldsymbol{u})}{\partial Q_i(\tau_i, u_i)} \geq 0, \quad \forall i \in \mathcal{N}$$

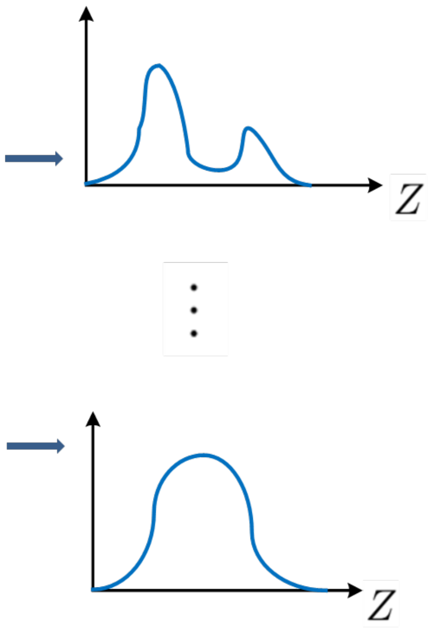


Sunehag et al. Value-decomposition networks for cooperative multi-agent learning based on team reward. In AAMAS, 2018.
Rashid et al. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning. In ICML, 2018.

Distributional RL

Value is actually a distribution





$$\theta_Z(\boldsymbol{\tau}, \boldsymbol{u}, \omega) = \inf\{z \in \mathcal{R} : \omega \leq CDF_Z(z)\}, \quad \forall \omega \in [0, 1]$$

$$Z(\tau, u) = \sum_{i=1}^{n} p_i(\tau, u, \omega_i)\delta_{\theta(\boldsymbol{\tau}, \boldsymbol{u}, \omega_i)}$$

- ***Distributional IGM (DIGM) principle***

$$\arg\max_{\mathbf{u}} \mathbb{E}[Z_{jt}(\boldsymbol{\tau}, \boldsymbol{u})]$$
$$= (\arg\max_{u_1} \mathbb{E}[Z_1(\tau_1, u_1)], \ \ldots, \ \arg\max_{u_N} \mathbb{E}[Z_N(\tau_N, u_N)])$$

Sun et al. DFAC Framework: Factorizing the Value Function via Quantile Mixture for Multi-Agent Distributional Q-Learning. In ICML, 2021.

Risk-sensitive RL

Risk-sensitive RL aims to optimize a risk measure based on a return distribution, rather than the expectation.

$$\pi_{\psi_\alpha}(s) = \arg\max_u \psi_\alpha[Z(s,u)]$$

Risk measures:

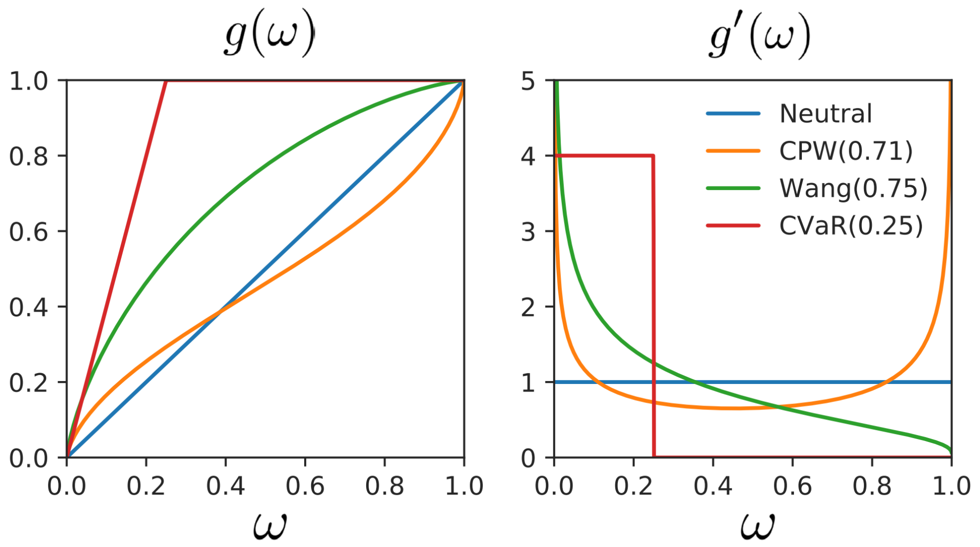
- Value-at-risk (VaR) $VaR_\alpha(Z(\boldsymbol{\tau},\boldsymbol{u})) = \theta(\boldsymbol{\tau},\boldsymbol{u},\alpha)$

Distorted risk measure (DRM) $\psi(Z) = \int_0^1 g'(\omega)\theta(\omega)d\omega$

- Conditional Value at Risk (CVaR)

$$CVaR_\alpha(Z) = \mathbb{E}_Z[z|z \leq \theta(\alpha)]$$

- Wang $g(\omega) = \Phi(\Phi^{-1}(\omega) + \alpha)$
- CPW $g(\omega) = \omega^\alpha/(\omega^\alpha + (1-\omega)^\alpha)^{\frac{1}{\alpha}}$



Motivation

➢ **Risk-sensitive scenarios**

Risk and Return

Risk-seeking

Risk-averse

➢ Most of the existing MARL value factorization methods do not extensively consider ***risk***, which could impact their performance negatively in some **risk-sensitive scenarios**.

➢ How to **effectively** combine risk-sensitive reinforcement learning with MARL value factorization?

RiskQ

Risk-sensitive Individual-Global-Max (RIGM) Principle

Definition 6 (RIGM). *Given a risk metric ψ_α, a set of individual return distribution utilities $[Z_i(\tau_i, u_i)]_{i=1}^N$, and a joint state-action return distribution $Z_{jt}(\boldsymbol{\tau}, \mathbf{u})$, if the following conditions are satisfied:*

$$\arg\max_{\mathbf{u}} \psi_\alpha[Z_{jt}(\boldsymbol{\tau}, \mathbf{u})] = (\arg\max_{u_1}[\psi_\alpha[Z_1(\tau_1, u_1)], \ \ldots, \ \arg\max_{u_N}[\psi_\alpha[Z_N(\tau_N, u_N)]), \quad (7)$$

where $\psi_\alpha : Z \times R \to R$ is a risk metric such as the VaR or a distorted risk measure, α is its risk level. Then, $[Z_i(\tau_i, u_i)]_{i=1}^N$ satisfy the RIGM principle with risk metric ψ_α for Z_{jt} under under τ. We can state that $Z_{jt}(\boldsymbol{\tau}, \mathbf{u})$ can be distributionally factorized by $[Z_i(\tau_i, u_i)]_{i=1}^N$ with risk metric ψ_α.

The RIGM principle is a generalization of the DIGM and the IGM principle.

- $\psi = CVaR$ and $\alpha = 1$, RIGM principle ➡ DIGM principle .

$$\arg\max_{\mathbf{u}} \mathbb{E}[Z_{jt}(\boldsymbol{\tau}, \mathbf{u})] = (\arg\max_{u_1} \mathbb{E}[Z_1(\tau_1, u_1)], \ \ldots, \ \arg\max_{u_N} \mathbb{E}[Z_N(\tau_N, u_N)])$$

- If Z_i is a single Dirac Delta Distribution(value distribution Z_i becomes a single value, i.e., Q_i), and in this case ($\psi = CVaR$ and $\alpha = 1$), RIGM principle ➡ IGM principle .

$$\arg\max_{\mathrm{u}} Q_{jt}(\tau, u) = \begin{pmatrix} \arg\max_{\mathrm{u}_1} Q_1(\tau_1, u_1) \\ \vdots \\ \arg\max_{\mathrm{u}_n} Q_n(\tau_n, u_n) \end{pmatrix}$$

RiskQ

Current value factorization methods can not satisfy RIGM principle

Theorem 1. *Given a deterministic joint action-value function Q_{jt}, a stochastic joint action-value function Z_{jt}, and a factorization function Φ for deterministic utilities:*

$$Q_{jt}(\tau, u) = \Phi(Q_1(\tau_1, u_1), ..., Q_n(\tau_n, u_n)) \tag{7}$$

such that $[Q_i]_{i=1}^n$ satisfy IGM for Q_{jt} under τ, the following risk-sensitive distributional factorization:

$$Z_{jt}(\tau, u) = \Phi(Z_1(\tau_1, u_1), ..., Z_n(\tau_n, u_n)) \tag{8}$$

is insufficient to guarantee that $[Z_i]_{i=1}^n$ satisfy RIGM for $Z_{jt}(\tau, u)$ with risk metric ψ_α.

Theorem 2. *Given a stochastic joint action-value function Z_{jt}, and a distributional factorization function Φ for the stochastic utilities which satisfy the DIGM theorem, the following risk-sensitive distributional factorization:*

$$Z_{jt}(\tau, u) = \Phi(Z_1(\tau_1, u_1), ..., Z_n(\tau_n, u_n)) \tag{9}$$

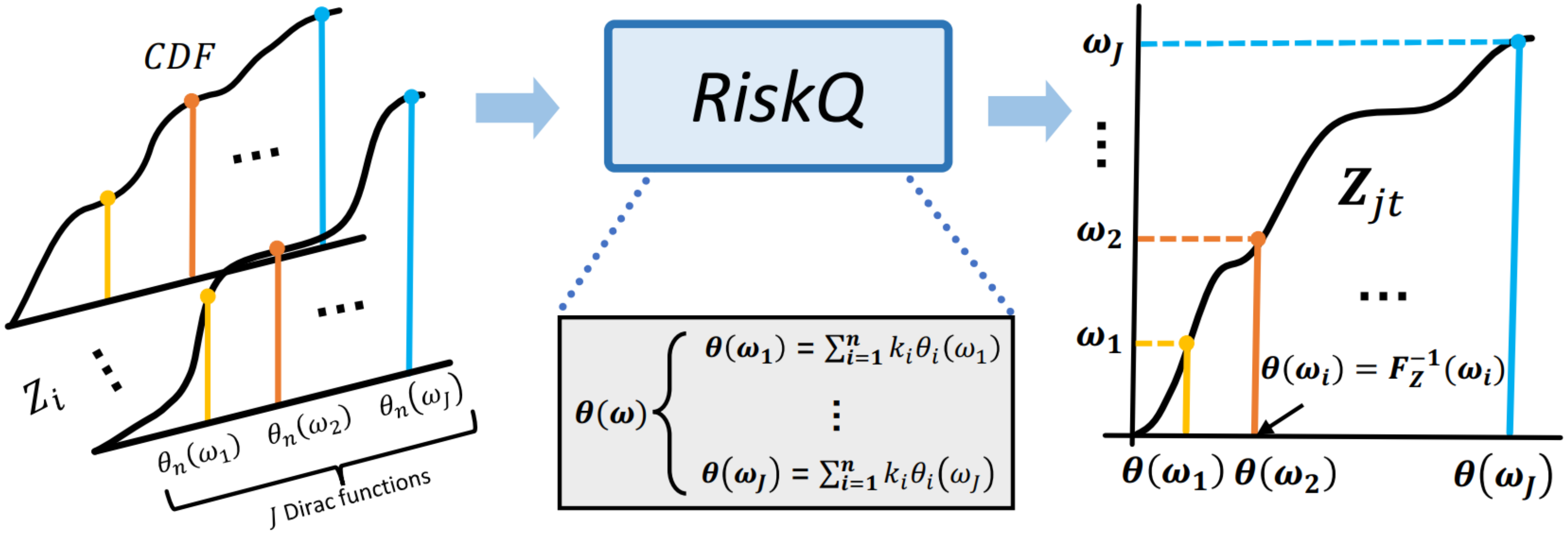
is insufficient to guarantee that $[Z_i]_{i=1}^n$ satisfy RIGM for $Z_{jt}(\tau, u)$ with risk metric VaR_α.

Theorem 3. *DRIMA [14] does not guarantee adherence to the RIGM principle for CVaR metric.*

RiskQ

RiskQ satisfies RIGM principle

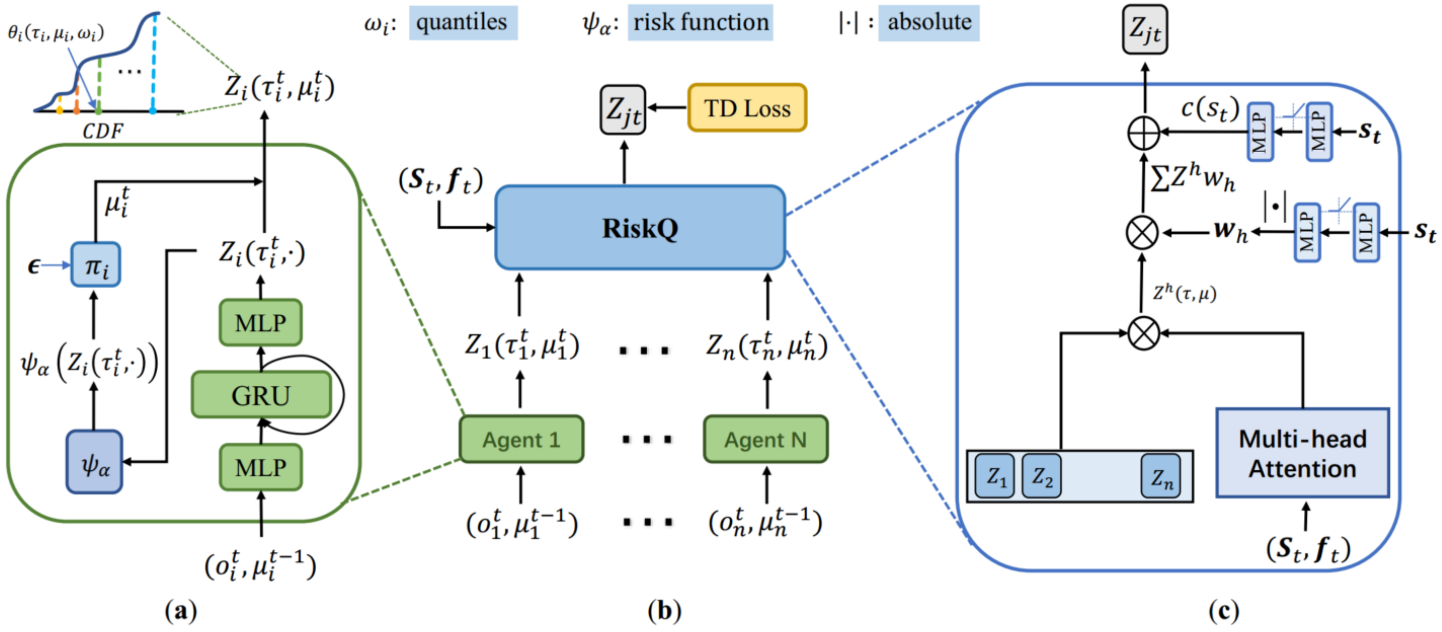
$$Z_{jt}(\boldsymbol{\tau}, \boldsymbol{u}) = \sum_{j=1}^{J} p_j(\boldsymbol{\tau}, \boldsymbol{u}, \omega_j)\delta_{\theta(\boldsymbol{\tau}, \boldsymbol{u}, \omega_j)} \qquad \theta(\boldsymbol{\tau}, \boldsymbol{u}, \omega_j) = \sum_{i=1}^{N} k_i \theta_i(\tau_i, u_i, \omega_j)$$



RiskQ overview: quantiles mixing for Z_{jt}

RiskQ

$$Z_{jt}(\boldsymbol{\tau}, \boldsymbol{u}) = \sum_{j=1}^{J} p_j(\boldsymbol{\tau}, \boldsymbol{u}, \omega_j)\delta_{\theta(\boldsymbol{\tau}, \boldsymbol{u}, \omega_j)} \qquad \theta(\boldsymbol{\tau}, \boldsymbol{u}, \omega_j) = \sum_{i=1}^{N} k_i \theta_i(\tau_i, u_i, \omega_j)$$



The framework of RiskQ

target distribution:

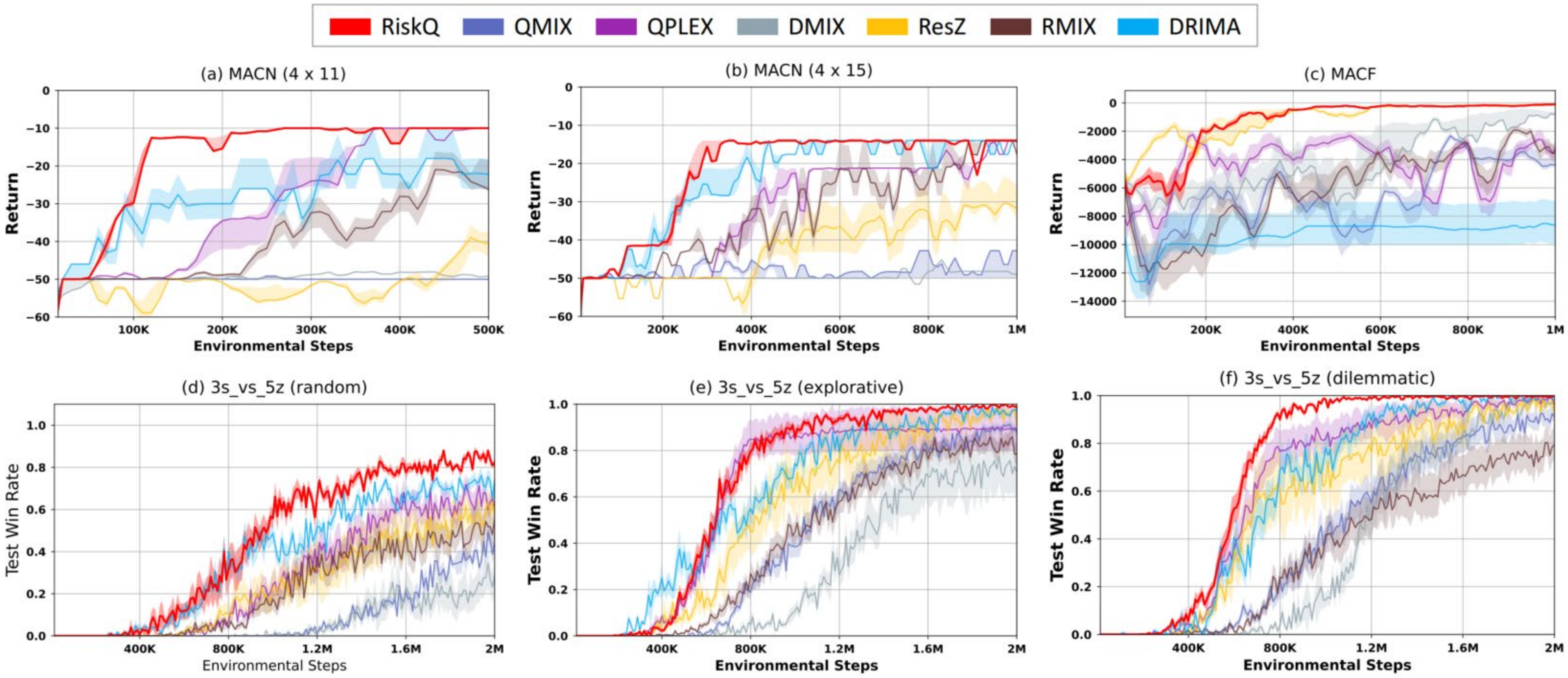
$$y^k(\boldsymbol{\tau}^k, \boldsymbol{u}^k, \sigma) \triangleq r + \gamma Z_{jt}(\boldsymbol{\tau}^{k+1}, \widetilde{\boldsymbol{u}}, \sigma^-)$$

$$\widetilde{\boldsymbol{u}} = [\widetilde{u}_i]_{i=1}^{N}$$

$$\widetilde{u}_i = \arg\max_{u_i} \psi_\alpha[Z_i(\tau_i^{k+1}, u_i)]$$

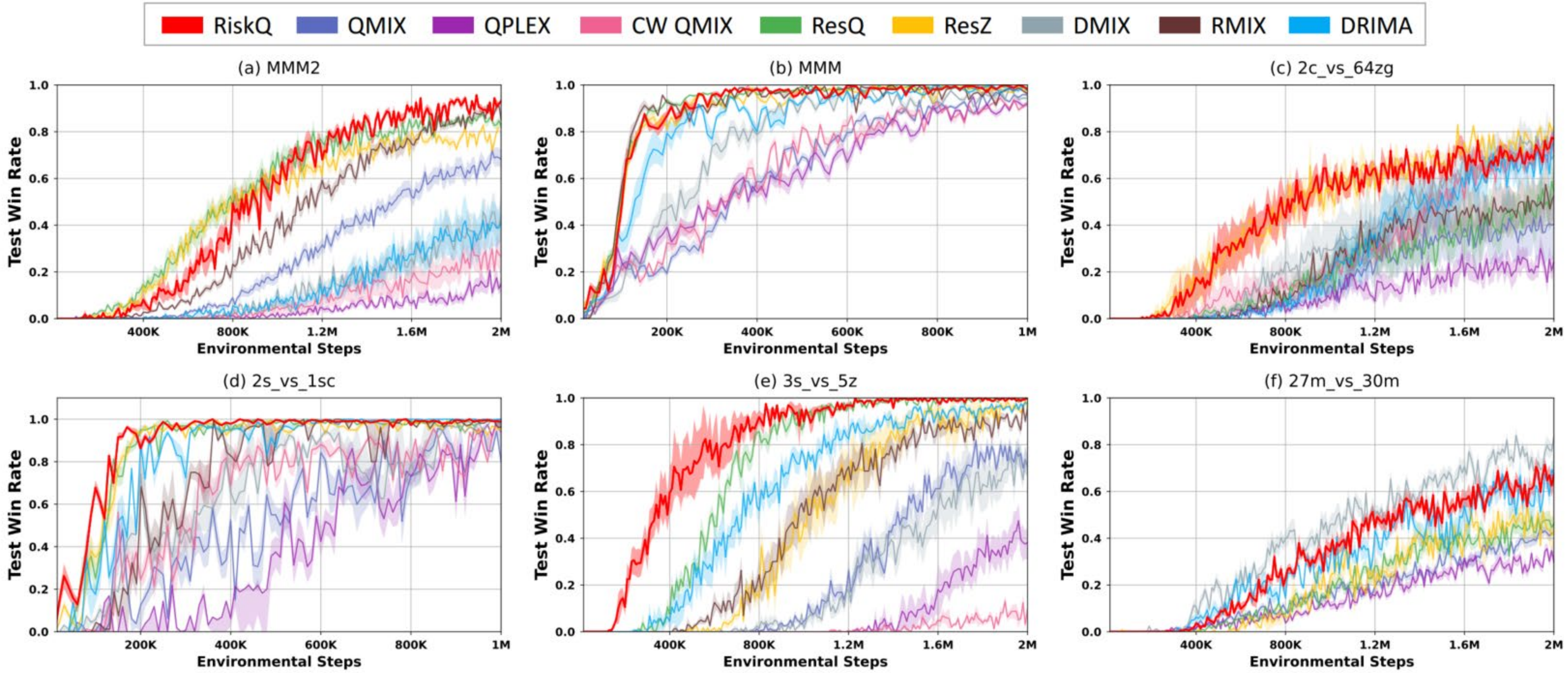
Experiments —— Risk-sensitive environments

RiskQ has better performance than other baselines in risk-sensitive settings



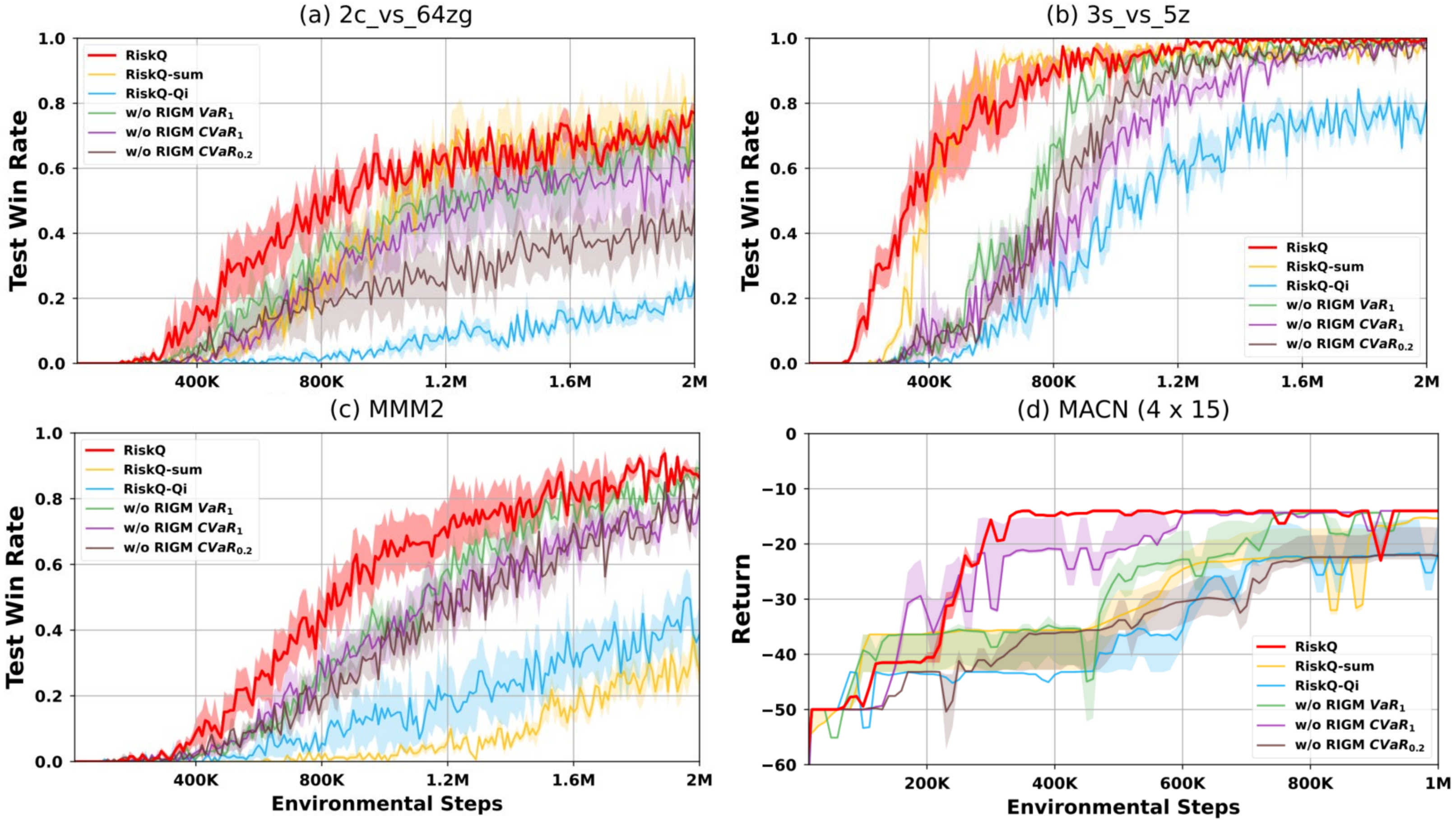
Experiments —— Starcraft II Multi-agent Challenge(SMAC)

RiskQ reaches the best win rate in most scenarios



Experiments —— ablations

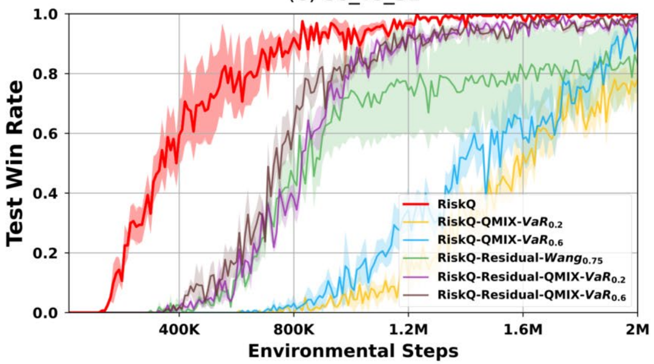
- It is important to satisfy the RIGM principle



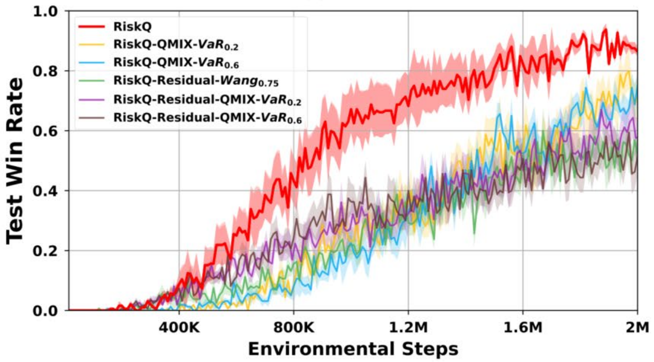
Experiments —— ablations

- The representation limitations of RiskQ do not significantly impact its performance

(e) 3s_vs_5z

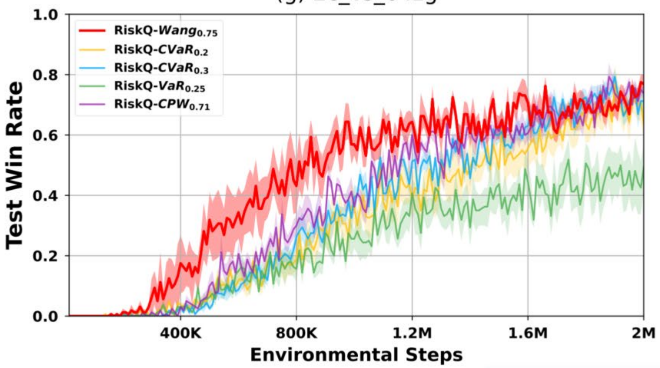


(f) MMM2

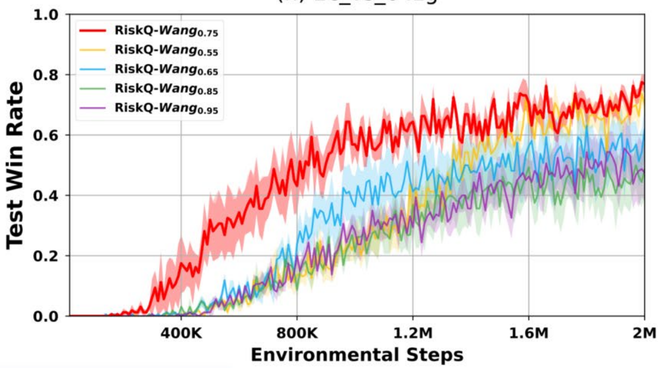


- Evaluate the impact of different risk metrics, risk levels and number of percentiles

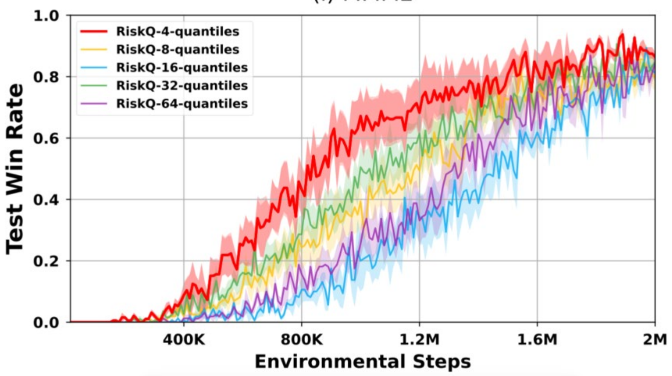
(g) 2c_vs_64zg



(h) 2c_vs_64zg



(i) MMM2

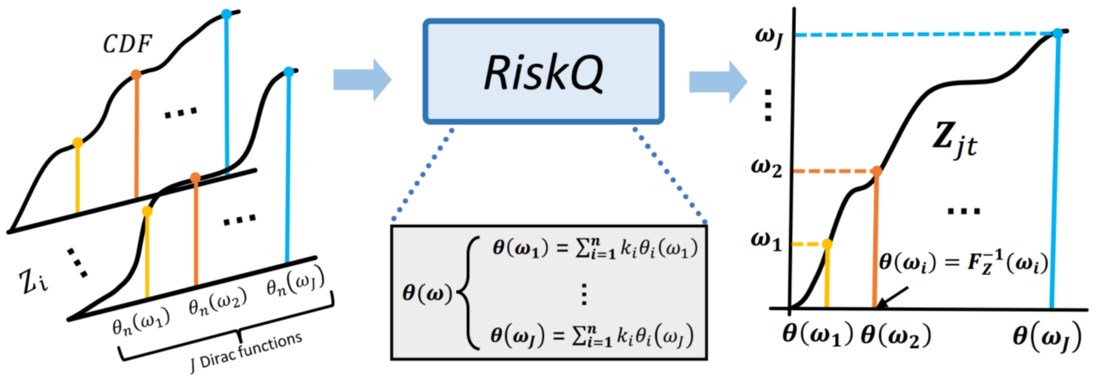


Summary

- RIGM principle, a generalization of IGM and DIGM principles.
- RiskQ, a value distribution factorization approach satisfying RIGM principle for Risk-sensitive Multi-Agent Reinforcement Learning problems
- Through extensive experiments, we show that RiskQ can obtain promising results.

For more details, please check our project page:
https://github.com/xmu-rl-3dv/RiskQ

Contact us:
siqishen@xmu.edu.cn
chennanma@stu.xmu.edu.cn
chaoli@stu.xmu.edu.cn
yongquanf@nudt.edu.cn



Thanks for your attention!