Safe Multi-Agent Reinforcement Learning via Distributional Cost Critic and Maximum Entropy Optimization

Qiwei Liu[1,2], Ye Yuan[1], Lingyue Zhang[1], Kaitian Chen[1], Yunkai Lv[1,3], Sheng Gao[1], Huaicheng Yan[1,4*]

[1]The Key Laboratory of Smart Manufacturing in Energy Chemical Process of the Ministry of Education
East China University of Science and Technology (ECUST), Shanghai, China
[2]Shanghai Institute of Intelligent Science and Technology, Tongji University
[3]Key Laboratory of System Control and Information Processing, Shanghai Jiao Tong University
[4]Faculty of Artificial Intelligence, Shanghai University of Electric Power
{qw_liu,yye,22012221,ktchen}@mail.ecust.edu.cn, {yunkailv,sheng_gao,hcyan}@ecust.edu.cn

Abstract

Deploying multi-agent reinforcement learning (MARL) in safety-critical systems faces significant challenges due to insufficient agent exploration and inadequate safety constraint guarantees. Current approaches are constrained by two fundamental limitations: inefficient exploration leading to suboptimal policies, and expected-cost-based constraint frameworks failing to ensure full-process safety. To address these challenges, this paper proposes a novel safety-aware maximum entropy (MaxEnt) MARL framework using Conditional Value-at-Risk (CVaR) as a joint safety metric, which quantifies constraint satisfaction under worst-case scenarios for multi-agent systems. Moreover, we develop the Worst-Case Multi-Agent Soft Actor-Critic (WCMASAC) algorithm, incorporating sequential update mechanisms and maximum entropy optimization for heterogeneous agents, enhanced with distributed safety critics. Theoretically, we establish the monotonic improvement property, guaranteed constraint satisfaction, and convergence to a quantum response equilibrium for WCMASAC. Extensive experiments on safety gymnasium-based benchmarks demonstrate that WCMASAC outperforms state-of-the-art baselines in both task reward acquisition and safety constraint violation reduction, while exhibiting superior exploration efficiency and risk-aware control capabilities.

Extended version —
https://github.com/YeY-YYe/WCMASAC

1 Introduction

As a core framework for decentralized decision-making, MARL encounters substantial difficulties in scenarios demanding rigorous safety guarantees. While traditional MARL algorithms optimize cooperative efficiency through long-term return maximization, their neglect of explicit safety constraints renders them incapable of ensuring compliance with safety boundary constraints in the state-action space of agent joint actions. This limitation substantially restricts the practical applications of MARL. In particular, safe reinforcement learning (RL) for a single agent has achieved safe exploration capabilities in complex environments (Liu et al. 2022; Yang et al. 2021; Ha et al. 2021; Li et al. 2024; Yang et al. 2023; Zhang, Vuong, and Ross 2020).

However, the non-stationary nature of multi-agent systems introduces fundamental theoretical challenges that hinder the extension of such methods to multi-agent systems. Firstly, partial observability impedes individual agents from modeling the global risk distribution of joint actions. Secondly, the interactions require each agent to pay attention to its own actions and to consider the impact of other agents' decisions on safety (Gu et al. 2024). To address these challenges, several pioneering results on safe MARL have been published (Zhang et al. 2020; Lu et al. 2021; Liu et al. 2021; Gu et al. 2023; Zhang et al. 2024; Du, Gou, and Cai 2025). Dec-PG (Lu et al. 2021) integrates decentralized policy gradient with primal-duality methods to identify saddle points between task rewards and safety costs. However, its reliance on a parameter-sharing mechanism among agents may lead to suboptimal convergence. CMIX (Liu et al. 2021) extends QMIX (Rashid et al. 2020b) by incorporating peak and average constraints. Although all agents empirically satisfy these constraints in experiments, the method still lacks theoretical safety guarantees. Furthermore, it depends on value decomposition methods that are inherently limited by restrictive underlying assumptions. Based on HARL (Zhong et al. 2024), MACPO and MAPPO-Lagrangian (Gu et al. 2023) integrate constrained optimization with advantage decomposition theory (Kuba et al. 2021), ensuring both multi-agent safety constraint satisfaction and monotonic policy improvement properties, thereby establishing a novel theoretical framework for safe MARL.

Although the MACPO framework is theoretically sound and has demonstrated promising performance, it still faces several critical challenges. First, the MACPO and MAPPO-Lagrangian algorithm, which ensures agent safety during training and execution by constraining the expectation of cost, lacks quantitative modeling capabilities for tail risks associated with safety constraints. Such constraints based on the expectation of cost can regulate the average rate of constraint violations but fail to suppress instantaneous safety violations that may occur during execution (Yang et al. 2021,

*Corresponding Author.

2023). Risk-sensitive RL approaches inspired by return distribution learning offer a potential solution to this challenge (Dabney et al. 2018a; Qiu et al. 2021; Hu et al. 2022; Sun et al. 2023; Shen et al. 2023). These methods enhance agents' sensitivity to risky behaviors and demonstrate superior performance in long-term returns by learning return distributions or designing risk-sensitive functions. Regrettably, existing distributional MARL methods primarily focus on the evolution of return probability distributions while failing to explicitly consider state-action joint safety constraints in multi-agent systems, thereby being unable to guarantee execution safety for agents.

Additionally, safety-constrained RL methods based on the MACPO framework may encounter insufficient exploration issues (Liu et al. 2023; Kuba et al. 2022; Sutton, Barto et al. 1998). While policy optimization approaches like MACPO ensure the safety feasibility of iterative policies, their conservative safety constraint mechanisms significantly restrict the exploration scope of the joint action space. In sparse safety feedback scenarios, this phenomenon triggers premature policy convergence, causing systems to settle for local safe optima at the expense of global performance. A potential solution involves enabling agents to learn stochastic behaviors for effective exploration of the reward space (Haarnoja et al. 2018). This creates dynamic game conflicts between individual entropy maximization objectives and global safety constraints, potentially leading some agents to exceed local safety thresholds in pursuit of enhanced exploration. While MaxEnt safe RL has advanced in single-agent settings (Yang et al. 2021), its extension to multi-agent systems remains largely unexplored, particularly in learning stochastic policies under safety constraints. Recent work by (Liu et al. 2023) proposes a MaxEnt heterogeneous agent reinforcement learning framework (MaxEnt-HARL), which ensures convergence of stochastic policies in unconstrained scenarios through decoupling of advantage functions. However, this approach fails to address dynamically evolving feasibility boundaries induced by safety constraints, limiting its applicability to safety-critical multi-agent coordination tasks.

Inspired by the above discussion, the principal contributions of this work are fourfold:

1. We establish the first theoretically-grounded MaxEnt safe MARL framework under distributional RL principles, providing a novel paradigm for stochastic policy optimization in constrained environments. Specifically, we innovatively employ mixed CVaR as a joint safety constraint metric for multi-agent systems, constructing a risk-sensitive safety mechanism by analyzing cost tail distribution.
2. We theoretically demonstrate the existence of equilibrium policies for constrained MaxEnt MARL optimization via a dual-decoupling design (risk-aware and advantage decomposition), while establishing theoretical guarantees for monotonic policy improvement and convergence to generalized Nash equilibrium.
3. Building upon this theoretical foundation, we propose WCMASAC, the first multi-agent soft actor-critic algorithm incorporating distributional safety critics, which achieves risk-aware joint policy optimization through worst-case value estimation.
4. Comprehensive experiments in safe Multi-MuJoCo cooperative control scenarios validate WCMASAC's superior performance, demonstrating higher sample efficiency and fewer constraint violations compared to state-of-the-art baselines (Ji et al. 2023), thereby offering a reliable solution for safety-critical multi-agent deployments.

2 Preliminaries and Problem Formulation

2.1 Constrained Cooperative Markov Game

Similarly to (Gu et al. 2022), a fully cooperative safe MARL described as a decentralized constrained Markov game is considered, which is denoted as a tuple $< \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathbb{P}, R, \mathbf{C}, \mathbf{d} >$. $\mathcal{N} = \{1, 2, \ldots, n\}$ is the set of agents. $\mathcal{S}$ is the state space. $\mathcal{A}$ denotes the joint action space and there exists $|\mathcal{A}| < \infty$, where $\mathcal{A} = \prod_{i=1}^{n} \mathcal{A}_i$ is the product of the action spaces of all agents. Let $\mathbf{a}(t) = (a_1(t), \ldots, a_n(t)) \in \mathcal{A}$ be the joint action, and the joint policy can be expressed as $\pi(\mathbf{a}(t)|s(t)) = \prod_{i=1}^{n} \pi_i(a_i(t)|s(t))$. The probabilistic state transition function is presented as $\mathbb{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$. When the system employs an action $\mathbf{a}(t)$ in state $s(t)$, the probability of transferring to the next moment state $s(t+1)$ is $\mathbb{P}(s(t+1)|s(t), \mathbf{a}(t))$, and agents receive the reward $R(s(t), \mathbf{a}(t))$. The primary goal of each agent is to maximize the expectation of standard cumulative rewards $J(\pi) = \mathbb{E}_{s_{0:\infty} \sim \mathbb{P}, \mathbf{a}_{0:\infty} \sim \pi} = [\sum_{t=1}^{\infty} R(s(t), \mathbf{a}(t))]$. Each agent's action may adversely affect the overall safety of the whole multi-agent system. Therefore, the safety constraints are modeled as joint safety constraints under the actions of all agents, and $\mathbf{C} = \{J_{i_j}\}_{1 \le j \le h}^{i \in \mathcal{N}}$ can be considered as the set of cost functions J_{i_j} for agent i, where h is the number of constraints.

Then, agent $i \in \mathcal{N}$ also has the following accumulated safety cost:

$$C_{i_j}(\pi) = \mathbb{E}_{s_{0:\infty} \sim \mathbb{P}, \mathbf{a}_{0:\infty} \sim \pi}[\sum_{t=0}^{\infty} J_{i_j}(s(t), a_i(t), \mathbf{a}_{-i}(t))], \tag{1}$$

which should satisfy $C_{i_j}(\pi) \le d_{i_j}$. $\mathbf{a}_{-i}(t)$ denotes the actions of all the other agents except for the agent i. The set of constraint thresholds is given as $\mathbf{d} = \{d_{i_j}\}_{1 \le j \le h}^{i \in \mathcal{N}}$.

2.2 Safety Based on Joint CVaR

Existing safe RL methods typically ensure system safety by constraining the expected cost, where agents optimize policies to maximize the expected return while ensuring that the expected cost remains below a predefined threshold (Gu et al. 2023, 2022; Zhang et al. 2024). However, long-term safety costs exhibit stochastic characteristics under the combined effects of policy stochasticity, agent interactions inducing environmental non-stationarity, and other

factors. This renders safety constraint frameworks based on expected cost inadequate for addressing potential risks arising from heavy-tailed distributions in long-term safety costs, making them unsuitable for direct application in safety-critical systems (Yang et al. 2021).

Therefore, inspired by distributional RL (Shen et al. 2023) and WCSAC (Yang et al. 2021), we adopt the Conditional Value-at-Risk (CVaR) (Rockafellar, Uryasev et al. 2000) safety measure to replace the expected cost as the safety constraint, with the essential definitions provided as follows.

Definition 1. ($CVaR_\omega$ with Risk level ω). Let $C_{i_j} := C_{i_j}(s(t), a_i(t), \mathbf{a}_{-i}(t))$ for convenience. Then for any safe cost $C_{i_j}(s(t), a_i(t), \mathbf{a}_{-i}(t))$, its risk function $CVaR_\omega$ under policy π is expressed as $\phi^\pi_{1-\omega}(C_{i_j}) := \mathbb{E}_{\mathbf{a}_{0:\infty}\sim\pi}[C_{i_j}|C_{i_j} \geq F^{-1}_{C_{i_j}}(1-\omega)]$ where $F^{-1}_{C_{i_j}}$ denotes the inverse function of C_{i_j}'s cumulative distribution function and the scalar $\omega \in (0, 1]$ denotes the level of risk. ω closer to 0 implies stronger safety, and vice versa represents a risk-neutral case.

Definition 2. ($CVaR_\omega$-based safety constraint). Given a risk level ω and the safety threshold value d_{i_j}, if the following constraint holds for the system under the policy π:

$$\phi^\pi_{1-\omega}(C_{i_j}) \leq d_{i_j} \forall t \in [0, \infty), \forall i \in \mathcal{N}, 1 \leq j \leq h. \quad (2)$$

The joint policy π can be considered as a worst-case safe policy.

In most fully cooperative multi-agent scenarios, a local safety constraint violation by an individual agent may induce global systemic risks. For instance, abrupt torque variations in a single robotic joint may lead to overall imbalance, or unexpected acceleration/deceleration behaviors of a member in a formation may trigger cascading collisions. Therefore, this paper comprehensively considers heavy-tailed distributional safety risks and assumes, without loss of generality, that any individual agent's safety constraint violation will adversely affect the system's global safety. Specifically, we assume that the system's safety cost adheres to the following Risk-sensitive Individual-Global-Max (RIGM) criterion proposed in (Shen et al. 2023). For any joint safety cost $\phi^\pi_{1-\omega}(C_{i_j}(s(t), a_i(t), \mathbf{a}_{-i}(t))$, the following condition exists:

$$\begin{aligned} &\arg\max_{\mathbf{a}} \phi^\pi_{1-\omega}[C_{i_j}(s(t), a_i(t), \mathbf{a}_{-i}(t))] \\ &= (\arg\max_{a_1} \phi^\pi_{1-\omega}[C_{i_j}(s(t), a_1(t))], \\ &\quad \ldots, \arg\max_{a_n} \phi^\pi_{1-\omega}[C_{i_j}(s(t), a_n(t))]) \end{aligned} \quad (3)$$

where $\mathbf{a} = (a_1, \ldots, a_n)$ is the joint action.

3 Method

This section proposes a worst-case MaxEnt safe RL framework, aiming to address two critical limitations of state-of-the-art safe MARL methods: (1) suboptimal convergence caused by insufficient exploration and (2) the inability to guarantee safety constraint satisfaction during stochastic policy learning. First, we formally derive the worst-case MaxEnt MARL objective function and characterize its generalized Nash equilibrium solution in Section 3.1. Next, Section 3.2 provides a rigorous theoretical analysis of the proposed framework, focusing on monotonic policy improvement and guarantees of Nash convergence. Finally, Section 3.3 details the specific algorithmic workflow of WC-MASAC.

3.1 Worst-Case Safe MaxEnt MARL

Based on probabilistic graphical inference (Levine 2018), the MaxEnt objective function without safety constraints was successfully derived in (Liu et al. 2023). Building upon (Levine 2018; Liu et al. 2023) and (2), the constrained optimization objective for safe MaxEnt MARL can be formulated as

$$\begin{aligned} \max_\pi J(\pi) = \mathbb{E}_{s_{0:\infty}\sim\mathbb{P}, \mathbf{a}_{0:\infty}\sim\pi}\Big[\sum_{t=1}^{\infty}\gamma^t\Big(R(s(t), \mathbf{a}(t)) \\ + \beta\sum_{i=1}^{n}\mathcal{H}(\pi_i(\cdot|s(t)))\Big)\Big] \end{aligned} \quad (4)$$

$$\textbf{s.t.} \quad \phi^\pi_{1-\omega}[C_{i_j}(s, a_i, \mathbf{a}_{-i}))] \leq d_{i_j}, \ \forall i \in \mathcal{N}, \ 1 \leq j \leq h. \quad (5)$$

where $\gamma \in (0, 1]$ is the discounted factor. $\mathcal{H}$ is the entropy term of π_i, and β is the temperature-like constant that regulates the degree of exploration randomness.

When the safety constraint and the entropy term are considered in the objective function, all agents will consider the comprehensive benefit of the safety cost and the reward, assign the mass probability, and eventually converge to the Quantum Response Equilibrium (QRE) (Liu et al. 2023; McKelvey and Palfrey 1995). The following Theorem 1 provides an expression for the equilibrium policy of the game.

Theorem 1. *For all agents $i \in \mathcal{N}$, the system policy converges to the QRE equilibrium policy if no agents can increase the system reward payoff by unilaterally changing their policy. In this case, define the Q-function as*

$$\begin{aligned} Q_\pi(s(t), \mathbf{a}(t)) = R(s(t), \mathbf{a}(t)) + \mathbb{E}_{s_{0:\infty}\sim\mathbb{P}, \mathbf{a}_{0:\infty}\sim\pi} \\ \Big[\sum_{l=0}^{\infty}\gamma^l\Big(R(s(t+l), \mathbf{a}(t+l)) + \beta\sum_{i=1}^{n}\mathcal{H}(\pi_i(\cdot|s(t)))\Big)\Big] \end{aligned} \quad (6)$$

Then, the QRE equilibrium policy π^*_i of agent i is expressed as

$$\pi^*_i(a_i|s) = \frac{\Omega(a_i)}{\sum_{b_i\in\mathcal{A}_i}\Omega(b_i)}, \quad (7)$$

in which

$$\begin{aligned} \Omega(a_i) = \exp\Bigg(\beta^{-1}\Bigg(\mathbb{E}\mathbf{a}{-}i \sim \pi^{-i}\big[Q\pi(s, a_i, \mathbf{a}{-}i)\big] \\ - \sum_{j=1}^{h}\frac{\mu^{i_j}}{\omega}\mathbb{E}_{\mathbf{a}_{-i}\sim\pi^*_{-i}}\big[\hat{\phi}^{\pi_i}_{1-\omega}\big(C_{i_j}(s, a_i, \mathbf{a}_{-i})\big)\big]\Bigg)\Bigg). \end{aligned} \quad (8)$$

and

$$\hat{\phi}^{\pi_i}_{1-\omega}(C_{i_j}(s,a_i,\mathbf{a}_{-i})) := C_{i_j}(s,a_i,\mathbf{a}_{-i}) \times \mathbf{1}\{C_{i_j}(s,a_i,\mathbf{a}_{-i}) \geq F^{-1}_{C_{i_j}}(1-\omega)\} \tag{9}$$

where $\mathbf{1}\{\cdot\}$ is indicate function. μ_{i_j} is the Lagrangian multiplier of the corresponding safety constraints, which is obtained by solving the following dual optimization problem:

$$\mu^*_{i_j} = \arg\min_{\mu_{i_j}} g_i(\mu_{i_j}) = \arg\min_{\mu_{i_j}} \sum_{j=1}^{h} \mu_{i_j} d_{i_j} - \sum_{j=1}^{h} \mu_{i_j} \mathbb{E}_{\mathbf{a}_{-i}\sim\pi^*_{-\mathbf{i}}}[\hat{\phi}_{1-\omega}(C_{i_j}(s,a_i,\mathbf{a}_{-i}))]. \tag{10}$$

The proof can be seen in the Appendix A.

3.2 Theoretical Analysis of WCMASAC

In the framework of the Worst-Case Safe MaxEnt MARL algorithm, agents first perform joint soft policy evaluation and distributional cost policy evaluation for rewards and safety costs, respectively, and then perform soft policy improvement. This section briefly introduces the convergence of the policy evaluation process and then rigorously analyzes the convergence of the policy improvement process of WCMASAC in theory.

Joint Soft Policy Evaluation Under the mechanism of centralized training and decentralized execution, the multi-agent joint policy evaluation process is a direct extension of the single-agent results. For the joint soft policy evaluation concerning rewards, the application of the soft Bellman backup operator directly ensures that the policy evaluation process converges to the joint soft Q-function (Zhang et al. 2021; Liu et al. 2023). Besides, existing results (Yang et al. 2021; Tang, Zhang, and Salakhutdinov 2019; Dabney et al. 2018b) have demonstrated that the distributional Bellman operator constitutes a contraction operator when employing the p-th Wasserstein metric for joint cost distributional policy evaluation.

Joint Safe Soft Policy Improvement When performing policy improvement, it is essential to introduce two critical lemmas: the advantage decomposition lemma (Zhong et al. 2024; Kuba et al. 2021) and the distributional decomposition lemma under RIGM (Shen et al. 2023), to rigorously demonstrate the equivalence between the results of the local policy update of individual agents under a sequential update mechanism and joint soft policy updates. These lemmas establish a theoretical foundation for proving that decentralized policy adjustments align with centralized optimization objectives in risk-sensitive multi-agent settings.

Lemma 1. (Multi-Agent Advantage Decomposition) (Kuba et al. 2021) For $i \in \mathcal{N}$, let $i_{1:m}$ denote the subset of $\mathcal{N}$ consisting of m agents. Then the following equation holds for a given joint policy π and any state s:

$$A^{\pi}_{i_{1:K}}(s,\mathbf{a}_{i_{1:K}}) = \sum_{k=1}^{K} A^{\pi}_{i_k}(s,\mathbf{a}_{i_{1:k-1}},a_{i_k}) \tag{11}$$

in which the multi-agent advantage function is defined as:

$$A^{\pi}_{i_{1:K}}(s,\mathbf{a}_{i_{1:K}}) = \sum_{k=1}^{K} \left(Q^{\pi}_{i_{1:k}}(s,\mathbf{a}_{i_{1:k}}) - Q^{\pi}_{i_{1:k-1}}(s,\mathbf{a}_{i_{1:k-1}})\right).$$

Lemma 2. (Multi-Agent Joint Cost Distribution Decomposition) (Shen et al. 2023) Let $\varphi(s,\mathbf{a},\nu_m)$ be a quantile sample function at sample point ν_m with probability $p(\varphi(s,\mathbf{a},\nu_m))$ of the cost distribution. Then, any joint cost distribution $C_{i_j}(s,a_i,\mathbf{a}_{-i}),\ i\in\mathcal{N},\ 0\leq j\leq h$ satisfying RIGM:

$$C_{i_j}(s,a_i,\mathbf{a}_{-i}) = \sum_{m=1}^{M} p_m(s,\mathbf{a},\nu_m)\delta(\varphi(s,\mathbf{a},\nu_m))$$
$$\varphi(s,\mathbf{a},\nu_m) = \sum_{i=1}^{N} \varpi_i \varphi_i(s,a_i,\nu_m) \tag{12}$$

can be distributionally factorized by $[C_i(s,a_i)]^N_{i=1}$ with $CVaR_\omega$ metric $\phi_{1-\omega}$. δ is a Dirac Delta function, and M is the number of sample ν_m. ϖ_i is the weight of the weighted summation.

Lemma 1 shows that in response to task rewards, agents can perform arbitrary sequential updates while taking into account the actions of other agents, without the need for traditional value decomposition assumptions (Qiu et al. 2021; Hu et al. 2023; Rashid et al. 2020a). In addition, Lemma 2 further shows that the safe joint cost risk measure for all agent actions acting together can be decomposed into a weighted sum of each agent. As we mentioned in Section 2.2, the Lemma 2 fits most multi-agent safe cooperation scenarios. Then the following Theorem shows that sequential local policy updates can induce joint policy updates, and the jointly safe MaxEnt MARL problem can be considered as a sum of n local problems.

Theorem 2. *Given a joint policy π and a permutation $i_{1:k}$, suppose that for $\forall k = 1,\ldots,n$ the agent k update its policy by*

$$\pi^{new}_{i_k} = \arg\min_{\pi_{i_k}\in\mathcal{A}_i} \mathrm{D}_{KL}\Bigg(\pi_{ik}(\cdot_{i_k}|s) \parallel \left(\mathbb{E}_{a_{i_{1:k-1}}\sim\pi^{new}_{i_{1:k-1}}}[Z_{\pi^{old}}(s,a_{i_{1:k-1}})]\right)^{-1} \cdot \exp\Bigg(\mathbb{E}_{a_{i_{1:k-1}}\sim\pi^{new}_{i_{1:k-1}}}[\beta^{-1}Q^{\pi^{old}_i}_{i_{1:k}}(s,\mathbf{a}_{i_{1:k-1}},a_{i_k})] - \mathbb{E}_{a_{i_{1:k-1}}\sim\pi^{new}_{i_{1:k-1}}}\Big[\beta^{-1}\sum_{j=1}^{h}\frac{\mu_{i_j}}{\omega}\Big(\sum_{l=1}^{k-1}\hat{\phi}^{\pi^{new}_l}_{1-\omega} (C_{i_j}(s,a_l)) + \hat{\phi}^{\pi^{old}_i}_{1-\omega}(C_{i_j}(s,a_{i_k}))\Big)\Big]\Bigg)\Bigg) \tag{13}$$

where $Z_{\pi^{old}}$ is a partition function to normalize the distribution, and Q^{π} denotes the Q-function under the policy π.

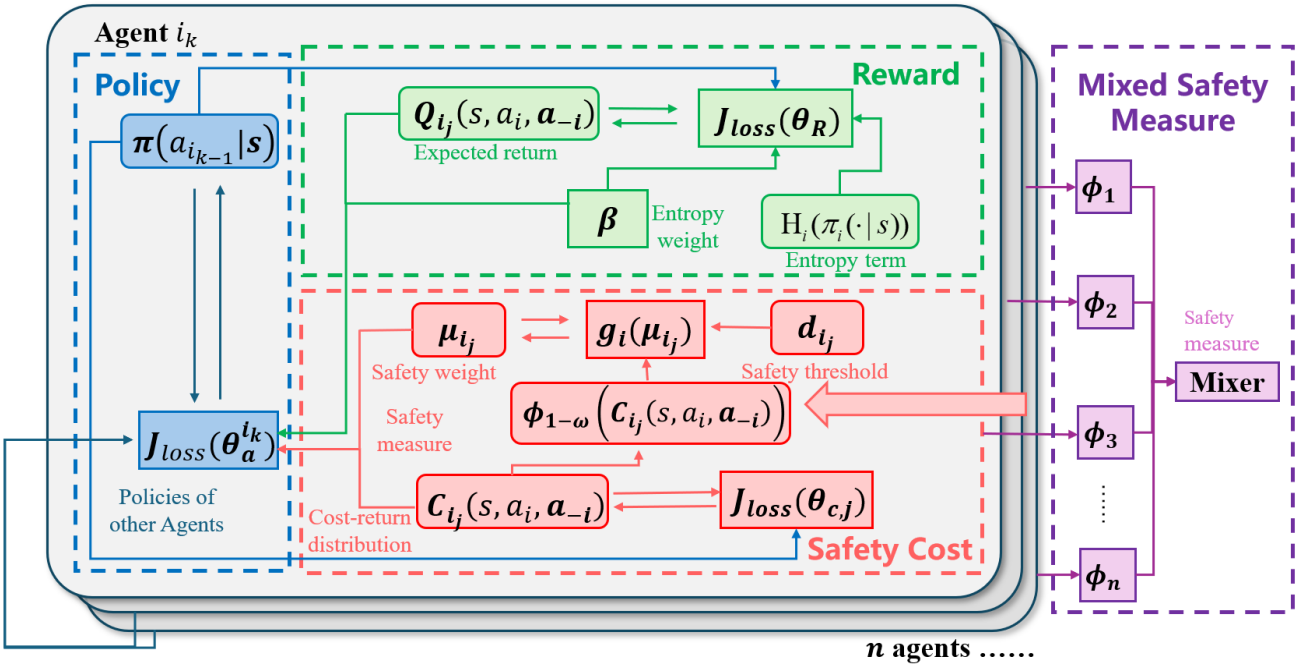


Figure 1: The framework of the WCMASAC.

Then the joint policy update process is

$$\pi^{new} = \arg\min_{\pi\in\mathcal{A}} \mathrm{D}_{KL}\Bigg(\pi(\cdot|s) \parallel \Bigg(\exp\Big(\beta^{-1}(Q^{\pi^{old}}_{i_{1:k}}(s, \mathbf{a}_{i_{1:k-1}}, a_{i_k}) - \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\mu_{i_j}}{\omega}\hat{\phi}(C_{i_j}(s,\cdot)))\Big)\Bigg) \cdot (Z_{\pi^{old}}(s))^{-1}\Bigg) \tag{14}$$

The proof can be seen in the Appendix B. Theorem 2 states that an individual agent can induce a joint soft policy by sequentially optimizing the KL divergence locally. Based on Theorem 2, the following lemma and theorem theoretically verify the monotonic improvement and QRE convergence of the policy, respectively.

Lemma 3. (Monotonic Improvement) Let $i_{1:k}$ be a permutation. Then for $\forall k = 1, \ldots, n$, if agent i_k acquires a new policy $\pi^{new}_{k_i}$ from $\pi^{old}_{i_k} \in \mathcal{A}_{i_k}$ through formula (13), then $Q_{\pi^{new}}(s,a) \geq Q_{\pi^{old}}(s,a)$ will hold and ensure that the safety constraints are satisfied.

Proof can be found in the Appendix C. Lemma 3 indicates that the agents can achieve monotonic improvement in performance while maintaining safety during policy update. Then the following Theorem shows that agents can perform policy evaluation and policy improvement alternately to satisfy the safety constraints and converge to QRE.

Theorem 3. *(QRE Convergence) If all agents perform policy evaluation and policy improvement alternately, the joint policy π eventually converges to the QRE equilibrium policy and satisfies the safe constraint.*

Proof can be found in the Appendix D. Theorem 3 provides a theoretical guarantee that agents monotonically improve their policies under safety constraints and converge to QRE, after which all agents do not unilaterally improve their own policies.

3.3 Practical Implementation of WCMASAC

In practical applications, we employ neural network-based approximators to estimate the centralized soft Q-function for rewards and decentralized policies for individual agents, while utilizing a hybrid approach combining quantile regression with a multi-head attention mechanism to approximate the risk measure of the centralized safety cost. Specifically, let θ_r, θ_c, and $\theta^{i_k}_a$ denote the network parameters of the reward critic, cost critic, and i_k's actor, respectively. These neural networks are optimized alternately via stochastic gradient descent to ensure coordinated learning of reward maximization, safety constraint satisfaction, and policy decentralization. Let $s := s(t)$, $\mathbf{a} := \mathbf{a}(t)$ and $s' = s(t+1)$, $a' = a(t+1)$ for convenience, then the centralized soft Q-function is optimized by minimizing the Bellman residual as

$$J_{loss}(\theta_r) = \mathbb{E}_{(s,\mathbf{a})\sim\mathcal{D}}\Bigg[\frac{1}{2}\Bigg(Q_{\theta_r}(s,\mathbf{a}) - \Big(R(s,\mathbf{a}) + \gamma\mathbb{E}_{s'\sim\mathbb{P},\ \mathbf{a}\sim\pi}[Q_{\theta_r}(s',\mathbf{a}') + \beta\sum_{i=1}^{n}\mathcal{H}(\pi_i(a_i{}'|s'))]\Big)\Bigg)^2\Bigg]. \tag{15}$$

Similar to the Quantile Regression (QR) loss (Dabney et al. 2018b), the j_{th} parameterized joint cost critic $C_{\theta_{c,j}}$, which gathers all agents cost C_{i_j}, achieves parameter optimization by minimizing the Huber loss between the safe cost distri-

bution and the target distribution as

$$J_{loss}(\theta_{c,j}) = H\Bigg(\mathbb{E}_{(s,\mathbf{a})\sim\mathcal{D}}\Big[C_{\theta_{c,j}}(s,\mathbf{a}) - \big(c_j + \gamma\mathbb{E}_{s'\sim\mathbb{P},\,\mathbf{a}\sim\pi}[C_{\theta_{c,j}}(s',\mathbf{a}')]\big)\Big]\Bigg) \tag{16}$$

where H means the Huber loss (Huber 1992) and c_j is the current joint cost. Then, all the agents minimize the expected KL-divergence (2) in a sequential update manner to improve their policies, which can derive the loss of the i_k's actor as follows:

$$J_{loss}(\theta_a^{i_k}) = \mathbb{E}_{(s,\mathbf{a})\sim\mathcal{D}}\Bigg[\mathbb{E}_{\mathbf{a}_{i_{1:k-1}}\sim\pi_{\hat{\theta}_a^{i_{1:k-1}}},a_i\sim\pi_{\theta_a^{i_k}}} \Big[Q_{\theta_r}(s,\ \mathbf{a}_{i_{1:k-1}},\ a_{i_k}) - \lambda_{i_k}\Big(\phi_{1-\omega}[C_{\theta_{c,j}}(s,\ \mathbf{a}_{i_{1:k-1}},\ a_{i_k})]\Big) - \alpha\log\pi_{\theta^{i_k a}}\left(a_{\theta_a^{i_k}}^{i_k}|s\right)\Big]\Bigg] \tag{17}$$

where $\hat{\theta}_a^{i_{1:k-1}}$ means the improved parameters of agents $i_{1:k-1}$. The framework can be seen in Figure 1, and the procedure as WCMASAC can be seen in the Appendix E.

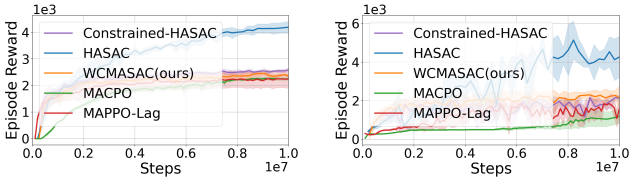
4 Experimental Setting and Results

4.1 Experimental Setting

To thoroughly evaluate the efficacy and robustness of our proposed WCMASAC algorithm, we conducted a series of experiments using the Safety-Gymnasium benchmark (Ji et al. 2023) in this section. This benchmark provides several challenging multi-agent environments specifically designed for safe RL research.

We conduct comparative experiments on six velocity-constrained tasks: `2x4Ant`, `4x2Ant`, `2x3HalfCheetah`, `6x1HalfCheetah`, `3x1Hopper`, and `2x3Walker2D`. These experiments evaluate our method against SOTA safe MARL algorithms, including MACPO, MAPPO-Lagrangian (Gu et al. 2023), and the SOTA MARL baseline HASAC (Liu et al. 2023). The tasks simulate scenarios where agents must maximize their movement speed to achieve high rewards while strictly complying with safety constraints, particularly velocity thresholds. This setup rigorously examines the trade-off between reward optimization and safety-critical constraint satisfaction, ensuring agents avoid hazardous velocity violations during operation.

Specifically, the experiments incorporate two key performance metrics, average episode reward and average episode cost, which are evaluated during the testing phase. The first metric quantifies the overall performance of the agents' tasks, while the second measures the severity of violations of the safety constraints per episode. Each agent is required to maximize cumulative rewards while ensuring that the episode cost does not exceed predefined safety thresholds. Furthermore, agents are encouraged to minimize the accumulated cost throughout the operation, thereby balancing reward-seeking behavior with strict adherence to safety-critical constraints. This dual-objective evaluation framework rigorously assesses the trade-off between task efficiency and safety compliance in multi-agent systems.



(a) Safety 2x3HalfCheetah (b) Safety 2x3Walker2d

(c) Safety 2x4Ant (d) Safety 3x1Hopper

(e) Safety 4x2Ant (f) Safety 6x1HalfCheetah

Figure 2: Reward comparisons on selected tasks of multiple benchmarks.

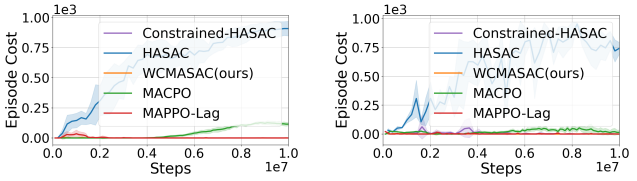
Moreover, Constrained-HASAC is developed to ensure fair comparisons. Unlike WCMASAC, it does not constrain the CVaR. Instead, it directly constrains the joint expectation of all agents.

All algorithms are implemented in a standard framework. Details of network architectures, hyperparameters, training procedures, and specific choices of optimizers, learning rates, and batch sizes are provided in the Appendix F.1.

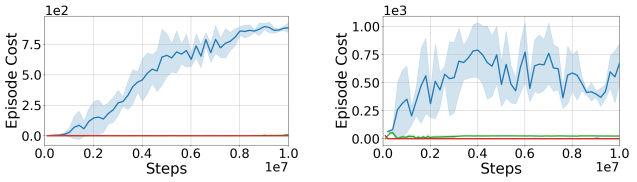
4.2 Experimental Results

Reward Comparison: As shown in the comparison of reward with SOTA safe MARL algorithms and the unconstrained SOTA MARL algorithms in Figure 2, our proposed WCMASAC and Constrained-HASAC show superior task reward performance compared to MACPO and MAPPO-Lag in all the six safety environments, though trailing HASAC. Notably, HASAC exhibits multiple risk-prone behaviors due to the absence of safety constraints (see Figure 3). Furthermore, WCMASAC achieves faster learning convergence rates, as its intrinsic exploration-encouraging mechanism enables more efficient exploration of the reward landscape, thereby accelerating the acquisition of optimal behaviors.

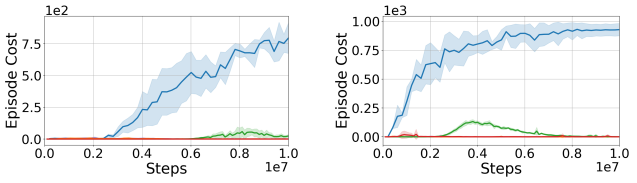
Cost Comparison: Figure 3 presents the cumulative safety costs and constraint violation statistics of different algorithms under identical environments. First, due to the absence of safety constraints, the SOTA MARL algorithm HASAC exhibits cumulative costs that exceed the safety threshold by a substantial margin and are significantly higher than those of safety-aware MARL baselines. This indicates that HASAC's policy execution compromises system



(a) Safety 2x3HalfCheetah (b) Safety 2x3Walker2d

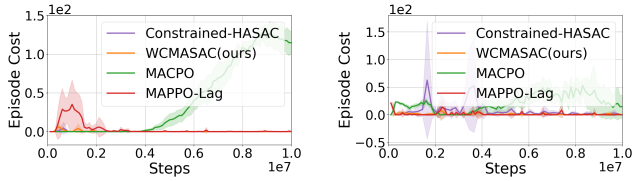


(c) Safety 2x4Ant (d) Safety 3x1Hopper

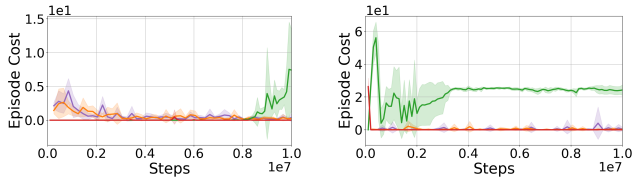


(e) Safety 4x2Ant (f) Safety 6x1HalfCheetah

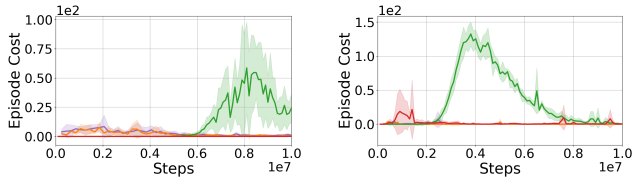
Figure 3: Cost comparisons on selected tasks of multiple benchmarks.



(a) Safety 2x3HalfCheetah (b) Safety 2x3Walker2d



(c) Safety 2x4Ant (d) Safety 3x1Hopper



(e) Safety 4x2Ant (f) Safety 6x1HalfCheetah

Figure 4: Cost comparisons on selected tasks of multiple safety benchmarks(without HASAC).

safety through hazardous behaviors, confirming the critical necessity of explicit safety constraints in real-world deployments.

Compared to MACPO, the proposed WCMASAC achieves lower costs in all environments while achieving higher task rewards with reduced execution risks, as evidenced in Figure 2-3. Notably, MACPO violates safety constraints in most environments, except the `Safety2x4AntVelocity` scenario, indicating dangerous agent behaviors during task execution and its failure to guarantee full-process safety. In contrast, WCMASAC and Constrained-HASAC ensure system safety throughout all operational phases in every environment, demonstrating that modeling tail risk distributions of agent costs effectively reduces the probability of selecting high-risk actions.

Figure 4 provides further safety performance comparisons between all the algorithms with safety constrains: MAPPO-Lag, MACPO, Constrained-HASAC and WCMASAC. Overall, WCMASAC and Constrained-HASAC demonstrate comparable safety performance to MAPPO-Lag but with a much lower probability of constraint violations. Specifically, from Figure 4a and Figure 4f we can see that WCMASAC and Constrained-HASAC achieve lower costs than MAPPO-Lag in the `2x3HalfCheetah` and `6x1HalfCheetah` environments, which means that our algorithms have better safety performance. In `2x3Ant` (Figure 4c), `3x1Hopper` (Figure 4d), and `4x2Ant` (Figure 4e) environments, WCMASAC exhibits slightly higher costs than MAPPO-Lag due to exploration encouragement. Notably, in the `2x3HalfCheetah`, `2x3Walker2D`, and `3x1Hopper` environments, MAPPO-Lag occasionally violates safety constraints, indicating that agents execute risk-prone behaviors that jeopardize system safety. In contrast, the proposed WCMASAC maintains strict constraint compliance across all environments, validating the effectiveness of the distributional safe critic in mitigating high-risk action selection. Moreover, WCMASAC outperforms Constrained-HASAC in most environments with lower costs in most environments(see Figure 4b,Figure 4c,Figure 4d and Figure 4e), which demonstrates the effectiveness of the CVaR constraint.

More detailed information such as safety thresholds and analysis can be found in the Appendix G.

Analysis of risk-level parameters: We investigate the impact of the risk-level parameter ω on task rewards and safety costs in the system. As $1-\omega$ approaches 1, agents adopt more conservative policies to ensure higher safety levels. We evaluate cumulative rewards and safety performance across all environments under $1-\omega$ values of $0.01, 0.5, 0.99$. Results reveal a declining trend in overall task rewards as $1-\omega$ increases, accompanied by substantial improvements in system safety (see Figure 5 above and Figure 2 in the Appendix G for detalied experiment results). This demonstrates the feasibility of adjusting ω based on different task requirements to balance reward maximization and safety assurance.

5 Conclusion

This work proposes a MaxEnt safe MARL framework with a safe distributional critic, enabling multi-agent systems to learn stochastic policies for enhanced exploration while guaranteeing safety constraint compliance throughout the

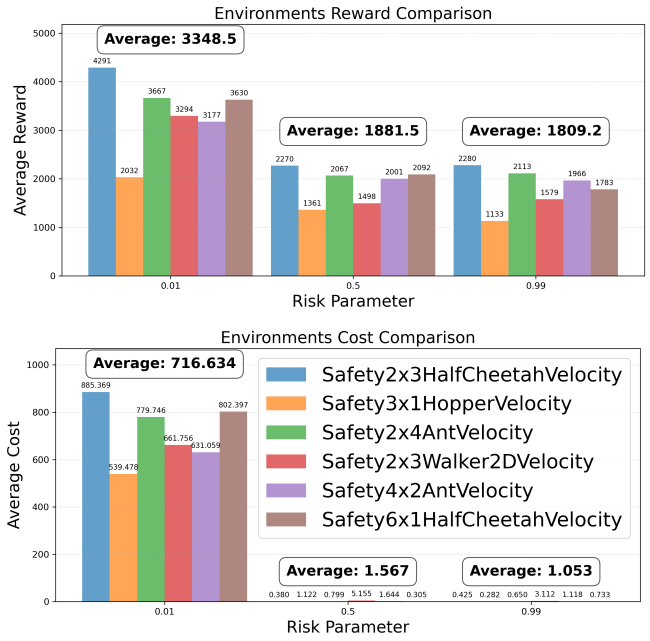


Figure 5: Average reward and average cost under different risk_level parameter ω.

entire operation. The proposed method effectively improves the system's final task reward and significantly reduces the probability of agents adopting high-risk behaviors. Theoretically, the framework provides closed-form expressions for game-theoretic equilibrium policies through rigorous mathematical proofs, ensuring monotonicity and convergence during policy improvement. Furthermore, we implement the algorithm as WCMASAC and demonstrate its superior performance in both task efficiency and safety compliance within the Safe Multi-Mujoco benchmark environments. Future work will focus on enhancing the algorithm's scalability in large-scale networks and exploring its deployment in practical multi-agent collaborative scenarios.

Acknowledgments

This work is supported by in part by the National Natural Science Foundation of China (62333005, 62403200, 62503174, 62473133, 62533002), in part by the Shanghai International Science and Technology Cooperation Project (24510714000), in part by the Shanghai Natural Science Foundation (24ZR1416200), in part by the China University-Industry-Research Innovation Founds (2024IT032), in part by the opening project of the State Key Laboratory of Autonomous Intelligent Unmanned Systems (ZZKF2025-1-24), and in part by the Fundamental Research Funds for the Central Universities.

References

Dabney, W.; Ostrovski, G.; Silver, D.; and Munos, R. 2018a. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, 1096–1105. PMLR.

Dabney, W.; Rowland, M.; Bellemare, M.; and Munos, R. 2018b. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Du, H.; Gou, F.; and Cai, Y. 2025. Scalable Safe Multi-Agent Reinforcement Learning for Multi-Agent System. *arXiv preprint arXiv:2501.13727*.

Gu, S.; Kuba, J. G.; Chen, Y.; Du, Y.; Yang, L.; Knoll, A.; and Yang, Y. 2023. Safe multi-agent reinforcement learning for multi-robot control. *Artificial Intelligence*, 319: 103905.

Gu, S.; Yang, L.; Du, Y.; Chen, G.; Walter, F.; Wang, J.; and Knoll, A. 2022. A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint arXiv:2205.10330*.

Gu, S.; Yang, L.; Du, Y.; Chen, G.; Walter, F.; Wang, J.; and Knoll, A. 2024. A Review of Safe Reinforcement Learning: Methods, Theories, and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 11216–11235.

Ha, S.; Xu, P.; Tan, Z.; Levine, S.; and Tan, J. 2021. Learning to Walk in the Real World with Minimal Human Effort. In *Conference on Robot Learning*, 1110–1120. PMLR.

Haarnoja, T.; Zhou, A.; Hartikainen, K.; Tucker, G.; Ha, S.; Tan, J.; Kumar, V.; Zhu, H.; Gupta, A.; Abbeel, P.; et al. 2018. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*.

Hu, J.; Sun, Y.; Chen, H.; Huang, S.; Chang, Y.; Sun, L.; et al. 2022. Distributional reward estimation for effective multi-agent deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 12619–12632.

Hu, T.; Luo, B.; Yang, C.; and Huang, T. 2023. MO-MIX: Multi-Objective Multi-Agent Cooperative Decision-Making With Deep Reinforcement Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 12098–12112.

Huber, P. J. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, 492–518. Springer.

Ji, J.; Zhang, B.; Zhou, J.; Pan, X.; Huang, W.; Sun, R.; Geng, Y.; Zhong, Y.; Dai, J.; and Yang, Y. 2023. Safety gymnasium: A unified safe reinforcement learning benchmark. *Advances in Neural Information Processing Systems*, 36: 18964–18993.

Kuba, J. G.; Feng, X.; Ding, S.; Dong, H.; Wang, J.; and Yang, Y. 2022. Heterogeneous-agent mirror learning: A continuum of solutions to cooperative marl. *arXiv preprint arXiv:2208.01682*.

Kuba, J. G.; Wen, M.; Meng, L.; Zhang, H.; Mguni, D.; Wang, J.; Yang, Y.; et al. 2021. Settling the variance of multi-agent policy gradients. *Advances in Neural Information Processing Systems*, 34: 13458–13470.

Levine, S. 2018. Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review. arXiv:1805.00909.

Li, Z.; Hu, C.; Wang, Y.; Yang, Y.; and Li, S. E. 2024. Safe Reinforcement Learning With Dual Robustness. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 10876–10890.

Liu, C.; Geng, N.; Aggarwal, V.; Lan, T.; Yang, Y.; and Xu, M. 2021. Cmix: Deep multi-agent reinforcement learning with peak and average constraints. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I 21*, 157–173. Springer.

Liu, J.; Zhong, Y.; Hu, S.; Fu, H.; Fu, Q.; Chang, X.; and Yang, Y. 2023. Maximum entropy heterogeneous-agent reinforcement learning. *arXiv preprint arXiv:2306.10715*.

Liu, Z.; Cen, Z.; Isenbaev, V.; Liu, W.; Wu, S.; Li, B.; and Zhao, D. 2022. Constrained variational policy optimization for safe reinforcement learning. In *International Conference on Machine Learning*, 13644–13668. PMLR.

Lu, S.; Zhang, K.; Chen, T.; Başar, T.; and Horesh, L. 2021. Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 8767–8775.

McKelvey, R. D.; and Palfrey, T. R. 1995. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1): 6–38.

Qiu, W.; Wang, X.; Yu, R.; Wang, R.; He, X.; An, B.; Obraztsova, S.; and Rabinovich, Z. 2021. RMIX: Learning risk-sensitive policies for cooperative reinforcement learning agents. *Advances in Neural Information Processing Systems*, 34: 23049–23062.

Rashid, T.; Farquhar, G.; Peng, B.; and Whiteson, S. 2020a. Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 33: 10199–10210.

Rashid, T.; Samvelyan, M.; De Witt, C. S.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2020b. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178): 1–51.

Rockafellar, R. T.; Uryasev, S.; et al. 2000. Optimization of conditional value-at-risk. *Journal of risk*, 2: 21–42.

Shen, S.; Ma, C.; Li, C.; Liu, W.; Fu, Y.; Mei, S.; Liu, X.; and Wang, C. 2023. RiskQ: risk-sensitive multi-agent reinforcement learning value factorization. *Advances in Neural Information Processing Systems*, 36: 34791–34825.

Sun, W.-F.; Lee, C.-K.; See, S.; and Lee, C.-Y. 2023. A unified framework for factorizing distributional value functions for multi-agent reinforcement learning. *Journal of Machine Learning Research*, 24(220): 1–32.

Sutton, R. S.; Barto, A. G.; et al. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Tang, Y. C.; Zhang, J.; and Salakhutdinov, R. 2019. Worst cases policy gradients. *arXiv preprint arXiv:1911.03618*.

Yang, Q.; Simão, T. D.; Tindemans, S. H.; and Spaan, M. T. 2021. WCSAC: Worst-case soft actor critic for safety-constrained reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10639–10646.

Yang, Q.; Simão, T. D.; Tindemans, S. H.; and Spaan, M. T. 2023. Safety-constrained reinforcement learning with a distributional safety critic. *Machine Learning*, 112(3): 859–887.

Zhang, K.; Sun, T.; Tao, Y.; Genc, S.; Mallya, S.; and Basar, T. 2020. Robust multi-agent reinforcement learning with model uncertainty. *Advances in neural information processing systems*, 33: 10571–10583.

Zhang, L.; Li, L.; Wei, W.; Song, H.; Yang, Y.; and Liang, J. 2024. Scalable Constrained Policy Optimization for Safe Multi-agent Reinforcement Learning. *Advances in Neural Information Processing Systems*, 37: 138698–138730.

Zhang, T.; Li, Y.; Wang, C.; Xie, G.; and Lu, Z. 2021. Fop: Factorizing optimal joint policy of maximum-entropy multi-agent reinforcement learning. In *International conference on machine learning*, 12491–12500. PMLR.

Zhang, Y.; Vuong, Q.; and Ross, K. 2020. First order constrained optimization in policy space. *Advances in Neural Information Processing Systems*, 33: 15338–15349.

Zhong, Y.; Kuba, J. G.; Feng, X.; Hu, S.; Ji, J.; and Yang, Y. 2024. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, 25(32): 1–67.