Group 23

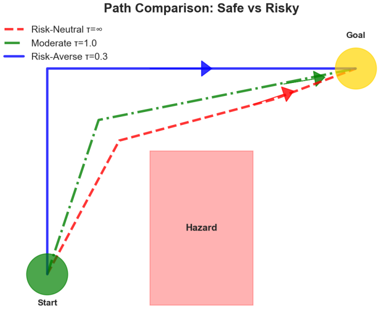
Risk-Averse Multi-Agent RL via Distribution Learning

November 13, 2025

The Problem

Standard MARL optimizes expected reward

Real-world systems need safety



Prior Work: Risk-Averse RL

Single-Agent Methods:

- **Distributional RL**
  - C51 (Bellemare et al., ICML 2017)
  - QR-DQN (Dabney et al., AAAI 2018)
  - Learn full return distribution
  - Extract risk post-hoc
- **Robust RL**
  - Iyengar, Math. OR 2005
  - Worst-case optimization
  - Very conservative
- **Mean-Variance**
  - $\max E[R] - \lambda \text{Var}[R]$
  - Hand-tune λ

Multi-Agent Methods:

- **RMIX / RiskQ**
  - RMIX (Qiu et al., NeurIPS 2021)
  - RiskQ (Shen et al., NeurIPS 2023)
  - Value factorization + CVaR
  - No equilibrium guarantees
- **Reward Shaping**
  - Manual per-environment
  - Weak theory

Gap: No tractable risk-averse equilibrium for MARL!

Risk Measures

How to extract risk-adjusted value from distribution?

Entropic Risk Measure

$$\rho_\tau(Z) = -\frac{1}{\tau} \log \mathbb{E}[\exp(-\tau Z)]$$

- $\tau \to \infty$: Risk-neutral (expected value)
- $\tau = 1.0$: Moderate risk-aversion
- $\tau = 0.3$: High risk-aversion (pessimistic)

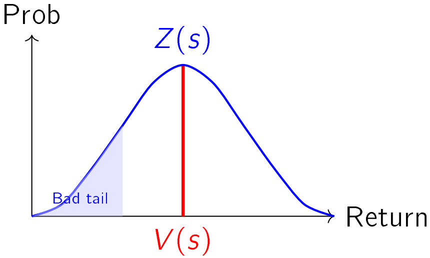
How Do We Learn the Distribution?

Standard RL:

$$V(s) = \mathbb{E}[\text{return}]$$
$$= \text{scalar}$$

Distributional RL:

$$Z(s) = \text{distribution over returns}$$
$$= [p_1, p_2, \ldots, p_{51}]$$



Distribution reveals risk

Bounded Rationality

Full PPO Loss

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CLIP}} + c_1 \cdot \mathcal{L}_{\text{VF}} - \epsilon \cdot H(\pi)$$

where $H(\pi) = -\sum_a \pi(a|s) \log \pi(a|s)$ (entropy)

- $\epsilon = 0$: Deterministic (no exploration)
- $\epsilon = 0.01$: Typical value (maintains exploration)
- ϵ is **fixed** during training (not annealed)

From behavioral economics: humans aren't perfectly rational

Theoretical Guarantee: Why Bounded Rationality?

From Mazumdar et al. (2025):

Theorem 3: Computational Tractability

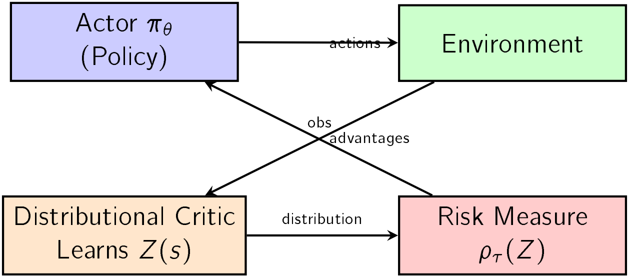
Risk-Averse QRE is **polynomial-time computable** via no-regret learning when:

$$\epsilon_1 \epsilon_2 \geq \xi_1^* \xi_2^*$$

where ϵ_i = bounded rationality, ξ_i^* = risk-aversion parameter

- Without bounded rationality ($\epsilon = 0$): Nash equilibrium is PPAD-complete (intractable)
- With bounded rationality ($\epsilon > 0$): Can use standard no-regret algorithms (tractable!)
- **Note:** In practice, entropy regularization is already standard in RL. The paper provides **theoretical justification** for this design choice.

Distributional RQE-MAPPO Architecture



Key Components:

- **Distributional Critic**: Learns return distribution $Z(s)$
- **Risk Measure**: Computes $V_\tau(s) = \rho_\tau(Z(s))$ for GAE
- **Fixed Entropy**: $\epsilon \cdot H(\pi)$ for bounded rationality

Training Objective

1. Critic Loss (Distributional Bellman)

$$\mathcal{L}_{\text{critic}} = \text{CrossEntropy}(Z_{\text{current}}, Z_{\text{target}})$$

where $Z_{\text{target}} = \text{PROJECT}[r + \gamma Z(s')]$

KL divergence minimization between distributions

2. Actor Loss (PPO + Entropy)

$$\mathcal{L}_{\text{actor}} = -\min(\text{ratio} \cdot \hat{A}, \text{clip}(\text{ratio}) \cdot \hat{A}) + \epsilon \cdot H(\pi)$$

Risk-adjusted advantages (GAE):

$$\hat{A}_t = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}$$

$$\delta_t = r_t + \gamma V_\tau(s_{t+1}) - V_\tau(s_t)$$

$$V_\tau(s) = \rho_\tau(Z(s)) = -\frac{1}{\tau} \log \mathbb{E}[\exp(-\tau Z(s))]$$

Experimental Environments

1. Risky CartPole

- Single-agent validation
- Random wind gusts
- Stochastic dynamics
- Fast iteration ($<$30 min)

Expected: Risk-averse agents survive longer under disturbances

2. Traffic Coordination (SUMO)

- Multi-agent main domain
- Intersection navigation
- Safety-critical
- Real-world relevance

Expected: Lower collision rates with risk-aversion

Baselines: Standard MAPPO ($\tau = \infty$), Reward-shaped MAPPO

Experiments

Experiment 1: Comparison Against Risk-Averse Methods

Compete against existing risk-averse MARL approaches:

- Standard MAPPO (risk-neutral baseline)
- **C51-CVaR MAPPO** (Bellemare et al., 2017 + CVaR)
- **RMIX** (Qiu et al., NeurIPS 2021)
- **Mean-Variance MAPPO** (dual critic: $E[R] - \lambda \mathrm{Var}[R]$)
- **Reward-Shaped MAPPO** (manual safety penalties)
- **RQE-MAPPO** (ours: entropic risk + bounded rationality)

Metrics: Mean return, Std dev, Collision rate, Worst 5% returns

Experiment 2: Risk-Reward Tradeoff

Sweep $\tau \in \{0.3, 0.5, 1.0, 2.0, 10.0\}$ for RQE

Expected: Pareto curve (safety vs efficiency), single τ generalizes

Experiment 3: Ablation Study

- Risk only ($\tau < \infty$, $\epsilon = 0$) vs Rationality only ($\tau = \infty$, $\epsilon > 0$) vs Both

Expected: Both components needed for best performance

Implementation Status

✓ Distributional Critic

✓ Risk Measures (entropic, CVaR, mean-var)

✓ Single-Agent PPO

✓ Risky CartPole Environment

✓ Training Pipeline

◦ Multi-Agent Integration

◦ SUMO Experiments

◦ Full Evaluation Suite

Future Direction: Policy Interpolation

Idea: Train two policies, interpolate at inference time

Approach

Training:

- Train risk-neutral policy: π_{neutral} with $\tau = \infty$
- Train risk-averse policy: π_{safe} with $\tau = 0.3$

Inference (Policy Interpolation):

$$\pi_\alpha(a|s) = \alpha \cdot \pi_{\text{safe}}(a|s) + (1 - \alpha) \cdot \pi_{\text{neutral}}(a|s)$$

- $\alpha = 1$: Fully risk-averse (safe, conservative)
- $\alpha = 0.5$: Balanced behavior
- $\alpha = 0$: Risk-neutral (efficient, aggressive)

Benefit: Tune safety-efficiency tradeoff *without retraining*

Questions?