Risk-Sensitive Multi-Agent Reinforcement Learning in Network Aggregative Markov Games

Extended Abstract

Hafez Ghaemi
University of Tehran, School of ECE
hafez.ghaemi@ut.ac.ir

Hamed Kebriaei
University of Tehran, School of ECE
kebriaei@ut.ac.ir

Alireza Ramezani Moghaddam
University of Tehran, School of ECE
a.ramezany@ut.ac.ir

Majid Nili Ahamdabadi
University of Tehran, School of ECE
mnili@ut.ac.ir

ABSTRACT

Classical multi-agent reinforcement learning (MARL) assumes risk neutrality and complete objectivity for agents. However, in settings where agents need to consider or model human economic or social preferences, a notion of risk must be incorporated into the RL optimization problem. This will be of greater importance in MARL where other human or non-human agents are involved, possibly with their own risk-sensitive policies. In this work, we consider risk-sensitive and non-cooperative MARL with cumulative prospect theory (CPT), a non-convex risk measure and a generalization of coherent measures of risk. CPT is capable of explaining loss aversion in humans and their tendency to overestimate/underestimate small/large probabilities. We propose a distributed sampling-based actor-critic (AC) algorithm with CPT risk for network aggregative Markov games (NAMGs), which we call Distributed Nested CPT-AC. Under a set of assumptions, we prove the convergence of the algorithm to a subjective notion of Markov perfect Nash equilibrium in NAMGs. The experimental results show that subjective CPT policies obtained by our algorithm can be different from the risk-neutral ones, and agents with a higher loss aversion are more inclined to socially isolate themselves in an NAMG.[1]

KEYWORDS

Multi-agent reinforcement learning, actor-critic, aggregative games, risk sensitivity, cumulative prospect theory

ACM Reference Format:
Hafez Ghaemi, Hamed Kebriaei, Alireza Ramezani Moghaddam, and Majid Nili Ahamdabadi. 2024. Risk-Sensitive Multi-Agent Reinforcement Learning in Network Aggregative Markov Games: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

[1]Code available at https://github.com/hafezgh/risk-sensitive-marl-namg

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

1 INTRODUCTION

Markov game (MG) is a theoretical framework for studying multi-agent systems (MAS) and multi-agent reinforcement learning (MARL) [21, 41]. Conventional risk-neutral MARL in MGs has seen great advances in recent years [1, 11, 13, 17, 22, 24, 30, 34, 40, 51]. Due to their internal preferences, agents can integrate a measure of risk into their RL objective, ushering into the realm of risk-sensitive RL. Risk in RL can be categorized into two main types based on the risk-sensitive objective [31]. Implicit risks impose a constraint on the RL stochastic optimization problem, e.g., variance as risk [32, 36, 45, 46] and chance constraints [7], while explicit risks directly incorporate risk into the objective function, e.g., entropic risk predicated on exponential return [4, 12, 25, 27, 43], coherent risk measures [2, 8], such as conditional value at risk (CVaR) [37], and cumulative prospect theory (CPT). Risk-sensitive MDPs governed by Markov coherent risk measures fall under the domain of robust MDPs [28], and dynamic programming and policy gradient (PG) techniques have been proposed for them [5, 6, 16, 26, 35, 38, 44, 47, 52]. CPT [50] is a non-convex generalization of coherent risk measures and an alternative to expected utility theory for modeling human decision making. It applies weighting functions to cumulative probabilities, separately for positive and negative outcomes, and uses non-linear utility functions to explains loss aversion in humans and their tendency to overestimate/underestimate small/large probabilities.

Given a real-valued r.v. X with distribution $\mathbb{P}(X)$, a reference point x_0, two monotonically non-decreasing weighting functions, $\omega^+ : [0,1] \rightarrow [0,1], \omega^- : [0,1] \rightarrow [0,1]$, utility functions $u^+ : \mathbb{R}^+ \rightarrow \mathbb{R}^+, u^- : \mathbb{R}^- \rightarrow \mathbb{R}^+$, and appropriate integrability assumptions, we can define the CPT value using Choquet integrals as $\mathbb{CPT}_{\mathbb{P}}[X] := \int_0^\infty \omega^+(\mathbb{P}(u^+((X-x_0)_+) > x))dx - \int_0^\infty \omega^-(\mathbb{P}(u^-((X-x_0)_-) > x))dx.$, where $(.)_+ = max(0,.)$ and $(.)_+ = -min(0,.)$. For a definition on a discrete r.v., see the complete version of the paper [14]. Conventional representations of CPT weighting and utility functions are $\omega^+(p) = \frac{p^\gamma}{(p^\gamma+(1-p)^\gamma)^{(1/\gamma)}}$, $\omega^-(p) = \frac{p^\delta}{(p^\delta+(1-p)^\delta)^{(1/\delta)}}$, and $u^+(x) = x^\alpha$ if $x \geq 0$ and $u^-(x) = \lambda(-x)^\beta$ if $x < 0$ [50]. The parameters $\gamma, \delta, \alpha, \beta$, and λ are subjective model parameters that can differ from person to person based on individual characteristics.

In this work, we consider risk-sensitive MARL with CPT risk measure in network aggregative Markov games (NAMGs). We derive a policy gradient theorem for CPT MARL as a generalization of previous PG algorithms for coherent risk measures [6, 44], and propose a distributed actor-critic algorithm to find CPT-sensitive policies for each agent with theoretical convergence guarantees, and the potential of convergence to a CPT-sensitive Markov perfect Nash equilibrium (MPNE).

Related Works. In the context of Markov risk measures in MDPs, CPT is articulated through two distinct formulations. The first one is the nested structure, wherein the CPT operator is applied to the cumulative return after each step (action taken) [18–20], which ensures the existence of a Bellman optimality equation. Recently, Tian et al. [49] extended this nested formulation to a multi-agent setting but restricted their approach to deterministic policies and a centralized value-iteration algorithm. In the second formulation, the CPT operator is applied solely to the agent's final cumulative return at the end of each episode [15, 33] and does not accept a Bellman equation, and is therefore approached by gradient-based policy optimization via offline Monte Carlo sampling [15, 23]. In this work, we opt for the nested formulation and an AC framework to learn risk-sensitive policies in a distributed manner in NAMGs (see [14] for a justification).

Network Aggregative Markov Games. An NAMG [9, 10, 29, 39, 42, 48] is an MG denoted by $M = (S, N, A, R, P, \mathcal{G}, \gamma, p_{s_0})$, where $\mathcal{G}(\mathcal{N}, \mathcal{E})$ is a communication graph of agents, and the reward function is a function of agent's own action and an aggregative function of the neighbors' actions, $R^i(s, a^i, a^{-i}) = R^i(s, a^i, \sigma^i(a^{-i}))$, where $\sigma^i(a^{-i}) = \sum_{j \in \mathcal{N} \setminus i} \omega_{ij} a^j$, with w_{ij} denoting the weight of the edge from j to i.

CPT Risk-Sensitive MARL Objective in NAMGs. Using the nested CPT formulation, the objective of the risk-sensitive agent i in an NAMG will be equivalent to

$$\max_{\pi^i} V_\pi^i(s_0) = \max_{\pi^i} \mathbb{CPT}_{\pi^i(a_0^i|s_0) \times \mathbb{P}(\sigma_0^{-i}|s_0) \times \mathbb{P}(s_1|s_0,a_0)} \left[R^i(s_0, a_0) + \gamma V_\pi^i(s_1)\right]. \tag{1}$$

2 DISTRIBUTED NESTED CPT ACTOR-CRITIC

We derive a gradient expression for the Markov dynamic CPT risk measure in NAMGs, $\nabla V_{\pi_\theta}^i(s_0)$ (the proof of theorems are available in the complete version [14]).

Theorem 1. *(Nested CPT Policy Gradient) Given Assumption 1 (see [14]), the gradient of the CPT return for agent i, $V_{\pi_\theta}^i(s_0)$, with respect to the policy parameter θ^i is*

$$\nabla V_{\pi_\theta}^i(s_0) \propto \mathbb{E}_{\mu_{cpt}^i(s)} \Big[\sum_{a,s'} \frac{\partial \phi}{\partial(\pi_\theta^i(a^i|s)\, \mathbb{P}(\sigma^{-i}|s)\, \mathbb{P}(s'|s,a))} \mathbb{P}(\sigma^{-i}|s)\, \mathbb{P}(s'|s,a) (\nabla \pi_{\theta^i}(a^i|s)) u(R^i(s, a^i, \sigma^{-i}, s') + \gamma V_{\pi_\theta}^i(s')) \Big], \tag{2}$$

where distribution μ_{cpt}^i is a subjective *steady-state probability distribution of the MDP.*

For the approximation scheme to estimate the subjective steady-state distribution and the gradient based on Algorithm 1 in Jie et al. [15] see the complete version [14]. Having a policy gradient theorem and a corresponding gradient approximation scheme, we propose Algorithm (1) to learn CPT-sensitive policies in NAMGs. **Convergence.** Convergence of the critic follows from Theorem 6 of Lin et al. [19], as the $TD(0)$ CPT operator,
$T_{cpt} V_{\pi_\theta}(s) = \mathbb{CPT}_{\pi_\theta(.|s) \times \mathbb{P}(.|s,a)} \left[R(s, a, s') + \gamma V_{\pi_\theta}(s')\right]$ is a sup-norm contraction (see [14] for assumptions and details).

Algorithm 1 Distributed Nested CPT Actor-Critic

For each agent n, repeat until convergence:
  Sample a_t^n from $\pi_{\theta_t^n}(.|s_t)$. Execute a_t^n and observe r_t^n, s_{t+1}, and σ_t^{-n}. Push $(r_t, s_{t+1}, \sigma_t^{-n})$ to $ExpDict^n(s_t, a_t^n, \sigma_t^{-n})$.
  Critic value estimation:
  for each $i = 1, 2, ..., n_{max}$, **do**
    Sample $\hat{a}_t^n$ from $\pi_{\theta_t^n}(.|s_t)$ and construct $\hat{\sigma}_t^{-n}$ by observing neighbors. Sample $(\hat{r}_t^n, \hat{s}_{t+1})$ from $ExpDict(s_t, \hat{a}_t^n, \hat{\sigma}_t^{-n})$ or a simulator of the environment.
    Let $X_i = \hat{r}_t^n + \gamma V_{\pi_\theta}^n(\hat{s}_{t+1})$. If the sample came from a simulator, push $(\hat{r}_t^n, \hat{s}_{t+1})$ to $ExpDict(s_t, \hat{a}_t^n, \hat{\sigma}_t^{-n})$.
  end for
  Estimate $\hat{V}_{\pi_{\theta_t}}^n(s_t)$ using array of X and Algorithm 1 in [15].
  Critic step:
  $\delta_t := \hat{V}_{\pi_{\theta_t}}^n(s_t) - V_{\pi_{\theta_t}}^n(s_t), \quad V_{\pi_{\theta_t}}^n(s_t) \leftarrow V_{\pi_{\theta_t}}^n(s_t) + \alpha_{cr,t}\delta_t.$
  Actor step: Compute $\nabla V_{\pi_{\theta_t}}^n(s_0)$ using the gradient estimation scheme and then $\theta_{t+1}^n := \theta_t^n + \alpha_{ac,t} \nabla V_{\pi_{\theta_t}}^n(s_0)$.

Theorem 2. *(**Convergence of the actor**) Given Assumptions 4 and 5 in [14] and learning steps such that, $\sum_{t=0}^\infty \alpha_{ac,t} = \infty$, $\sum_{t=0}^\infty \alpha_{cr,t} = \infty$, $\sum_{t=0}^\infty \alpha_{cr,t}^2 < \infty$, $\sum_{t=0}^\infty \alpha_{ac,t}^2 < \infty$, $\lim_{t\to\infty} \frac{\alpha_{ac,t}}{\alpha_{cr,t}} = 0$, Algorithm (1) converges to the unique CPT-sensitive Markov perfect Nash equilibrium of the NAMG, asymptotically.*

Given the asymptotic proofs, we apply Theorem 1.1 of Borkar [3], which implies asymptotic convergence of the AC algorithm. Note that Assumptions 4 and 5 [14] are hard to verify and if they do not hold, we can only ensure convergence to locally optimal policies.

3 NUMERICAL EXPERIMENT

We construct a risk-sensitive NAMG with an interpretable design to measure the effect of loss aversion on CPT-sensitive agents. In the NAMG ($\mathcal{N} = 4, \mathcal{S} = \{0, 1, 2, 3, 4\}, \mathcal{A} = \{0, 1, 2\}$), the reward function is defined as $R^i(s, a^i, \sigma^i(a^{-i})) = R_{self}^i(s) + \sigma^i(a^{-i}) R_{com}^i(s) a^i$, with $R_{self}(s, a^i) \sim N(0.5, 0.1)$ and $R_{com}^i(s) \sim 5 \cdot U(-0.5, 0.5)$, and $\sigma^i(a^{-i}) = \frac{1}{N-1}(\sum_{j \in \mathcal{N} \setminus i} a_j)$. This setup implies a high risk for the agent if it decides to take an action greater than $a^i = 0$, become socially involved with its neighboring community and tie its received reward to their actions. Figure 1 shows the convergence results and the probability of choosing $a = 0$ (a quantitative indicator of social conservatism) which is proportional to the loss-aversion level of the agents in the community.

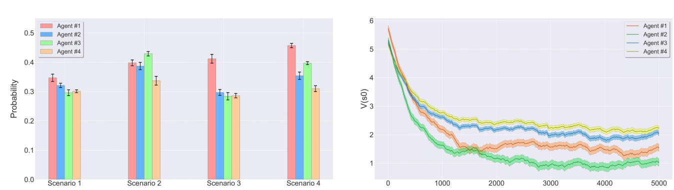

Figure 1: Left: mean converged policies over eight independent runs for different loss aversion scenarios. Scenario 1: all agents risk-neutral, scenario 2: all agents risk-sensitive ($\lambda = 2.6$), scenario 3: only Agent 1 is risk-sensitive ($\lambda = 2.6$), scenario 4: Agent 1 has a higher loss aversion coefficient ($\lambda = 3.2$) than others ($\lambda = 2.6$). Right: the state value of s_0 for scenario 2 over iterations.

REFERENCES

[1] Ahmet Alacaoglu, Luca Viano, Niao He, and Volkan Cevher. 2022. A natural actor-critic framework for zero-sum Markov games. In *International Conference on Machine Learning*. PMLR, 307–366.

[2] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. 1999. Coherent measures of risk. *Mathematical finance* 9, 3 (1999), 203–228.

[3] Vivek S Borkar. 1997. Stochastic approximation with two time scales. *Systems & Control Letters* 29, 5 (1997), 291–294.

[4] Vivek S Borkar. 2001. A sensitivity formula for risk-sensitive cost and the actor–critic algorithm. *Systems & Control Letters* 44, 5 (2001), 339–346.

[5] Ozlem Cavus and Andrzej Ruszczynski. 2014. Risk-averse control of undiscounted transient Markov models. *SIAM Journal on Control and Optimization* 52, 6 (2014), 3935–3966.

[6] Yinlam Chow and Mohammad Ghavamzadeh. 2014. Algorithms for CVaR optimization in MDPs. *Advances in neural information processing systems* 27 (2014).

[7] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. 2017. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research* 18, 1 (2017), 6070–6120.

[8] Freddy Delbaen. 2002. Coherent risk measures on general probability spaces. *Advances in finance and stochastics: essays in honour of Dieter Sondermann* (2002), 1–37.

[9] Zhenhua Deng. 2019. Distributed algorithm design for resource allocation problems of second-order multiagent systems over weight-balanced digraphs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 51, 6 (2019), 3512–3521.

[10] Zhenhua Deng. 2021. Distributed algorithm design for aggregative games of Euler–Lagrange systems and its application to smart grids. *IEEE Transactions on Cybernetics* 52, 8 (2021), 8315–8325.

[11] Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo Jovanovic. 2022. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*. PMLR, 5166–5220.

[12] Yingjie Fei, Zhuoran Yang, Yudong Chen, and Zhaoran Wang. 2021. Exponential bellman equation and improved regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 20436–20446.

[13] Roy Fox, Stephen M Mcaleer, Will Overman, and Ioannis Panageas. 2022. Independent natural policy gradient always converges in Markov potential games. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4414–4425.

[14] Hafez Ghaemi, Hamed Kebriaei, Alireza Ramezani Moghaddam, and Majid Nili Ahamdabadi. 2024. Risk-Sensitive Multi-Agent Reinforcement Learning in Network Aggregative Markov Games. arXiv:2402.05906 [cs.LG]

[15] Cheng Jie, LA Prashanth, Michael Fu, Steve Marcus, and Csaba Szepesvári. 2018. Stochastic optimization in a cumulative prospect theory framework. *IEEE Trans. Automat. Control* 63, 9 (2018), 2867–2882.

[16] Prashanth La and Mohammad Ghavamzadeh. 2013. Actor-critic algorithms for risk-sensitive MDPs. *Advances in neural information processing systems* 26 (2013).

[17] Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. 2021. Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969* (2021).

[18] Kun Lin. 2013. *Stochastic systems with cumulative prospect theory.* Ph.D. Dissertation. University of Maryland, College Park.

[19] Kun Lin, Cheng Jie, and Steven I Marcus. 2018. Probabilistically distorted risk-sensitive infinite-horizon dynamic programming. *Automatica* 97 (2018), 1–6.

[20] Kun Lin and Steven I Marcus. 2013. Dynamic programming with non-convex risk-sensitive measures. In *2013 American Control Conference*. IEEE, 6778–6783.

[21] Michael L Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*. Elsevier, 157–163.

[22] Chinmay Maheshwari, Manxi Wu, Druv Pai, and Shankar Sastry. 2022. Independent and decentralized learning in markov potential games. *arXiv preprint arXiv:2205.14590* (2022).

[23] Jared Markowitz, Marie Chau, and I-Jeng Wang. 2021. Deep CPT-RL: Imparting Human-Like Risk Sensitivity to Artificial Agents.. In *SafeAI@ AAAI*.

[24] David H Mguni, Yutong Wu, Yali Du, Yaodong Yang, Ziyi Wang, Minne Li, Ying Wen, Joel Jennings, and Jun Wang. 2021. Learning in nonzero-sum stochastic games with potentials. In *International Conference on Machine Learning*. PMLR, 7688–7699.

[25] Mehrdad Moharrami, Yashaswini Murthy, Arghyadip Roy, and Rayadurgam Srikant. 2022. A Policy Gradient Algorithm for the Risk-Sensitive Exponential Cost MDP. *arXiv preprint arXiv:2202.04157* (2022).

[26] Md Shirajum Munir, Sarder Fakhrul Abedin, Nguyen H Tran, Zhu Han, Eui-Nam Huh, and Choong Seon Hong. 2021. Risk-aware energy scheduling for edge computing with microgrid: A multi-agent deep reinforcement learning approach. *IEEE Transactions on Network and Service Management* 18, 3 (2021), 3476–3497.

[27] Erfaun Noorani and John S Baras. 2022. Risk-attitudes, Trust, and Emergence of Coordination in Multi-agent Reinforcement Learning Systems: A Study of Independent Risk-sensitive REINFORCE. In *2022 European Control Conference (ECC)*. IEEE, 2266–2271.

[28] Takayuki Osogami. 2012. Robustness and risk-sensitivity in Markov decision processes. *Advances in neural information processing systems* 25 (2012).

[29] Francesca Parise, Sergio Grammatico, Basilio Gentile, and John Lygeros. 2020. Distributed convergence to Nash equilibria in network and average aggregative games. *Automatica* 117 (2020), 108959.

[30] Julien Perolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. 2015. Approximate dynamic programming for two-player zero-sum Markov games. In *International Conference on Machine Learning*. PMLR, 1321–1329.

[31] LA Prashanth, Michael C Fu, et al. 2022. Risk-Sensitive Reinforcement Learning via Policy Gradient Search. *Foundations and Trends® in Machine Learning* 15, 5 (2022), 537–693.

[32] LA Prashanth and Mohammad Ghavamzadeh. 2016. Variance-constrained actor-critic algorithms for discounted and average reward MDPs. *Machine Learning* 105 (2016), 367–417.

[33] LA Prashanth, Cheng Jie, Michael Fu, Steve Marcus, and Csaba Szepesvári. 2016. Cumulative prospect theory meets reinforcement learning: Prediction and control. In *International Conference on Machine Learning*. PMLR, 1406–1415.

[34] Shuang Qiu, Xiaohan Wei, Jieping Ye, Zhaoran Wang, and Zhuoran Yang. 2021. Provably efficient fictitious play policy optimization for zero-sum Markov games with structured transitions. In *International Conference on Machine Learning*. PMLR, 8715–8725.

[35] Wei Qiu, Xinrun Wang, Runsheng Yu, Rundong Wang, Xu He, Bo An, Svetlana Obraztsova, and Zinovi Rabinovich. 2021. RMIX: Learning risk-sensitive policies for cooperative reinforcement learning agents. *Advances in Neural Information Processing Systems* 34 (2021), 23049–23062.

[36] D Sai Koti Reddy, Amrita Saha, Srikanth G Tamilselvam, Priyanka Agrawal, and Pankaj Dayama. 2019. Risk averse reinforcement learning for mixed multi-agent environments. In *Proceedings of the 18th international conference on autonomous agents and multiagent systems*. 2171–2173.

[37] R Tyrrell Rockafellar, Stanislav Uryasev, et al. 2000. Optimization of conditional value-at-risk. *Journal of risk* 2 (2000), 21–42.

[38] Andrzej Ruszczyński. 2010. Risk-averse dynamic programming for Markov decision processes. *Mathematical programming* 125 (2010), 235–261.

[39] Mohsen Saffar, Hamed Kebriaei, and Dusit Niyato. 2017. Pricing and rate optimization of cloud radio access network using robust hierarchical dynamic game. *IEEE Transactions on Wireless Communications* 16, 11 (2017), 7404–7418.

[40] Muhammed Sayin, Kaiqing Zhang, David Leslie, Tamer Basar, and Asuman Ozdaglar. 2021. Decentralized Q-learning in zero-sum Markov games. *Advances in Neural Information Processing Systems* 34 (2021), 18320–18334.

[41] Lloyd S Shapley. 1953. Stochastic games. *Proceedings of the national academy of sciences* 39, 10 (1953), 1095–1100.

[42] Mohammad Shokri and Hamed Kebriaei. 2020. Leader–follower network aggregative game with stochastic agents' communication and activeness. *IEEE Trans. Automat. Control* 65, 12 (2020), 5496–5502.

[43] Mehdi Naderi Soorki, Walid Saad, Mehdi Bennis, and Choong Seon Hong. 2021. Ultra-reliable indoor millimeter wave communications using multiple artificial intelligence-powered intelligent surfaces. *IEEE Transactions on Communications* 69, 11 (2021), 7444–7457.

[44] Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. 2015. Policy gradient for coherent risk measures. *Advances in neural information processing systems* 28 (2015).

[45] Aviv Tamar, Dotan Di Castro, and Shie Mannor. 2012. Policy gradients with variance related risk criteria. In *Proceedings of the twenty-ninth international conference on machine learning*. 387–396.

[46] Aviv Tamar, Dotan Di Castro, and Shie Mannor. 2013. Temporal difference methods for the variance of the reward to go. In *International Conference on Machine Learning*. PMLR, 495–503.

[47] Aviv Tamar, Shie Mannor, and Huan Xu. 2014. Scaling up robust MDPs using function approximation. In *International conference on machine learning*. PMLR, 181–189.

[48] Shaolin Tan, Yaonan Wang, and Athanasios V Vasilakos. 2021. Distributed population dynamics for searching generalized nash equilibria of population games with graphical strategy interactions. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 52, 5 (2021), 3263–3272.

[49] Ran Tian, Liting Sun, and Masayoshi Tomizuka. 2021. Bounded risk-sensitive markov games: Forward policy design and inverse reward learning with iterative reasoning and cumulative prospect theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6011–6020.

[50] Amos Tversky and Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty* 5 (1992), 297–323.

[51] Kaiqing Zhang, Sham Kakade, Tamer Basar, and Lin Yang. 2020. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems* 33 (2020), 1166–1178.

[52] Ziqing Zhu, Ka Wing Chan, Siqi Bu, Bin Zhou, and Shiwei Xia. 2022. Nash Equilibrium Estimation and Analysis in Joint Peer-to-Peer Electricity and Carbon Emission Auction Market With Microgrid Prosumers. *IEEE Transactions on Power Systems* (2022).