Bounded Risk-Sensitive Markov Games: Forward Policy Design and Inverse Reward Learning with Iterative Reasoning and Cumulative Prospect Theory

Ran Tian[*]   **Liting Sun**[*]   **Masayoshi Tomizuka**

{rantian, litingsun, tomizuka}@berkeley.edu (*Equal contribution)
University of California, Berkeley

Abstract

Classical game-theoretic approaches for multi-agent systems in both the forward policy design problem and the inverse reward learning problem often make strong rationality assumptions: agents perfectly maximize expected utilities under uncertainties. Such assumptions, however, substantially mismatch with observed humans' behaviors such as satisficing with sub-optimal, risk-seeking, and loss-aversion decisions. In this paper, we investigate the problem of bounded risk-sensitive Markov Game (BRSMG) and its inverse reward learning problem for modeling human realistic behaviors and learning human behavioral models. Drawing on iterative reasoning models and cumulative prospect theory, we embrace that humans have *bounded intelligence* and maximize *risk-sensitive* utilities in BRSMGs. Convergence analysis for both the forward policy design and the inverse reward learning problems are established under the BRSMG framework. We validate the proposed forward policy design and inverse reward learning algorithms in a navigation scenario. The results show that the behaviors of agents demonstrate both risk-averse and risk-seeking characteristics. Moreover, in the inverse reward learning task, the proposed bounded risk-sensitive inverse learning algorithm outperforms a baseline risk-neutral inverse learning algorithm by effectively recovering not only more accurate reward values but also the intelligence levels and the risk-measure parameters given demonstrations of agents' interactive behaviors.

1 Introduction

Markov Game (MG), as an approach to model interactions and decision-making processes of agents in multi-agent systems, have been employed in many domains such as economics (Amir 2003), games (Silver et al. 2017), and human-robot/machine interaction (Bu et al. 2008; Fisac et al. 2019). In classical MGs, agents are commonly assumed to be perfectly rational in obtaining their interaction policies. For instance, in a two-player Markov Game, agent 1 is assumed to make decisions based on his/her belief in agent 2's behavioral model in which agent 2 is also assumed to behave according to his/her belief in agent 1's model . . . and both agents are maximizing their expected rewards based on the infinite levels of mutual beliefs. If the beliefs match the actual models, perfect Markov strategies of all agents may be found by

solving the Markov-perfect equilibrium (MPE) of the game where a Nash equilibrium is reached. Under such assumptions, we can either solve for humans' optimal strategies with pre-defined rewards (forward policy design) or recover humans' rewards by observing their behaviors (inverse reward learning).

However, real human behaviors often significantly deviate from such "perfectly rational" assumptions from two major aspects (Goeree and Holt 2001). First, mounting evidence has shown that rather than spending a great amount of effort to hunt for the best action, humans often choose actions that are satisfying (i.e., actions that are above their pre-defined thresholds according to certain criteria) and relatively quick and easy to find. Simon (Simon 1976) formulated such a decision-making strategy as bounded rationality. Among the many developed models to capture such bounded rationality, iterative reasoning models from behavioral game theory (Camerer 2011) are some of the most prominent paradigms. These models do not assume humans as perfect players with infinite layers of strategic thinking during interactions, but model them as agents with finite levels of intelligence (rationality). Second, instead of optimizing the risk-neutral expected rewards, humans demonstrate strong tendency towards risk-sensitive measures when evaluating the outcomes of their actions under uncertainties. They are risk-seeking in terms of gains and risk-averse for losses. Such deviations make it difficult to model real humans' interactive behaviors using classical MGs.

We aim to establish a new game-theoretic framework, i.e., the bounded risk-sensitive Markov Game (BRSMG) that considers the two aspects of realistic human behaviors discussed above. The integration of bounded rationality and risk-sensitive measure have failed the well-established Nash equilibrium strategies, and fundamental questions such as the convergence of policy design and reward learning have to be re-visited. Hence, our goal is to develop general solutions to both the forward policy design problem and inverse reward learning problem in the BRSMG framework.

More specifically, in the forward strategy design problem, we model humans' cognitive limitation via iterative reasoning models, and model the influence of humans' risk sensitivity via cumulative prospect theory (CPT) (Tversky and Kahneman 1992) - a non-expected utility theory. In the inverse reward learning problem, we develop a bounded risk-
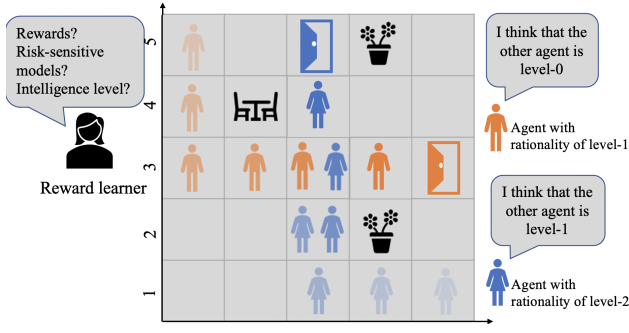
Figure 1: Modeling interactions between humans as a bounded risk-sensitive Markov Game: two human agents plan to exit the room through specified doors without collisions with obstacles and each other. We aim to answer two questions: 1) assuming both humans have bounded intelligence levels and risk-sensitive performance measures, how will their optimal policies differ from that in classical MGs? and 2) how to recover the rewards, intelligence levels, and risk-sensitivity parameters given their trajectories?

sensitive inverse learning algorithm which can recover not only the nominal rewards of agents but also the intelligence levels and the parameters in their risk measures directly from their demonstrated behaviors. *To our best knowledge, our work is the first to consider both bounded rationality and risk-sensitivity in general-sum MGs for both the forward and the inverse problems.*

Contributions. In summary, our contributions are threefold:

1. We propose a novel game-theoretic framework (BRSMG) that captures bounded rationality and risk-sensitivity in humans' reasoning process.

2. The proposed framework makes the first attempt to establish a bridge between inverse reward learning and risk-sensitive iterative reasoning models for decision-making.

3. In contrast to previous risk-neutral reward learning algorithms that learn humans' rewards under equilibrium solution concepts, we exploit an alternative paradigm based on non-equilibrium solution concepts and offer a solution that simultaneously learns humans' rewards, intelligence levels, and parameters in their risk-sensitive measure. Therefore, our solution provides an interpretable and heterogeneous human behavioral model, which is of critical importance for the development of human-robot interaction such as autonomous vehicles and intelligent robots.

2 Related Work

Bounded rationality. The influence of bounded rationality in forward strategy design problems has been studied in both single-agent and multi-agent settings. One group of studies formulate such influence by introducing additional computational costs to agents' actions (Ben-Sasson, Kalai, and Kalai 2007; Halpern 2008; Halpern and Pass 2015). Another group focuses on models that can explicitly capture the bounded reasoning processes of humans. Examples include the Boltzmann rationality model (Von Neumann and Morgenstern 2007), the quantal response equilibrium solution (QRE) (McKelvey and Palfrey 1995), and various iterative reasoning models (Costa-Gomes, Crawford, and Broseta

2001; Camerer, Ho, and Chong 2004; Stahl II and Wilson 1994). The Boltzmann model and QRE model formulate irrational behaviors of humans via sub-optimality, while iterative reasoning models emphasize more on the reasoning depths of humans. Instead of assuming humans to perform infinite levels of strategic reasoning, iterative reasoning models only allow for a finite number. The above models have been exploited for modeling human behaviors in many application domains, including normal-form zero-sum games (Tian et al. 2020), aerospace (Yildiz, Agogino, and Brat 2014; Kokolakis, Kanellopoulos, and Vamvoudakis 2020), cyber-physical security (Kanellopoulos and Vamvoudakis 2019), and human-robot interaction (Li et al. 2018; Fisac et al. 2019; Tian et al. 2018). It is shown in (Wright and Leyton-Brown 2014) that compared to QRE, iterative reasoning models can achieve better performance in predicting human behaviors in simultaneous move games. More specifically, (Wright and Leyton-Brown 2017) suggests that the quantal level-k model is the state-of-the-art among various iterative reasoning models.

Risk measure. Many risk measures have been proposed to model humans' decision process under uncertainties. Beyond expectation, value-at-risk (VaR) and conditional value-at-Risk (CVaR) (Pflug 2000) are two well-adopted risk measures. In addition, the cumulative prospect theory (CPT) (Tversky and Kahneman 1992) formulates a model that can well explain a substantial amount of human risk-sensitive behaviors. In the light of those risk measures, many risk-aware decision-making and reward learning algorithms have been proposed for single-agent settings (Lin and Marcus 2013; Chow et al. 2015; Mazumdar et al. 2017; Jie et al. 2018; Kwon et al. 2020; Majumdar et al. 2017; Ratliff and Mazumdar 2019). The risk-sensitive inverse reward learning problem for multi-agents was studied in (Sun et al. 2019) with a Stackelberg Game assumption (Simaan and Cruz 1973). The ego agent was assumed to be the leader and other agents were followers. Such a leader-follower assumption treats all agents as homogeneous agents and thus can not capture the diversity of humans in terms of their reasoning capabilities.

Inverse reward learning in games. The inverse reward learning problem for modeling the interactions among humans has attracted great attention from researchers, starting from simplified open-loop game formulations (Sadigh et al. 2016; Sun et al. 2018) to closed-loop games (Yu, Song, and Ermon 2019; Gruver et al. 2020). The concept of QRE was first adopted by (Yu, Song, and Ermon 2019) in inverse reward learning in multi-agent games to develop a maximum-entropy multi-agent inverse reinforcement learning algorithm. (Gruver et al. 2020) further extended the idea for better efficiency and scalability by introducing a latent space in the reward network. Though (Wright and Leyton-Brown 2014) suggested that iterative reasoning models can predict human behaviors more accurately in simultaneous move games than QRE, the multi-agent reward learning problem with iterative reasoning models and risk sensitivity has never been addressed. In this work, we propose the BRSMG framework to fill the gap.

3 Preliminaries

3.1 Classical Markov Game

In classical two-player MGs, each agent is represented by a Markov decision process. We denote a MG as $\mathcal{G} \triangleq \langle \mathcal{P}, \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \tilde{\gamma} \rangle$, where $\mathcal{P}=\{1,2\}$ is the set of agents in the game; $\mathcal{S}=S^1 \times S^2$ and $\mathcal{A}=A^1 \times A^2$ are, respectively, the joint state and action spaces of the two agents; $\mathcal{R}=(R^1, R^2)$ is the set of agents' one-step reward functions with $R^i : \mathcal{S} \times A^i \times A^{-i} \to \mathbb{R}$ $(-i = \mathcal{P} \setminus \{i\})$ represents the opponent of agent i); $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ represents the state transition of the game (we consider deterministic state transitions in this paper); and $\tilde{\gamma}$ is the reward discount factor.

We let $\pi^i : \mathcal{S} \to A^i$ denote a deterministic policy of agent i. At step t, given the current state s_t, each agent tries to maximize his/her expected total discounted reward under uncertainties in opponent's responses. Namely, the optimal policy $\pi^{*,i}$ is given by $\pi^{*,i} = \arg\max_{\pi^i} V^{i,\pi^i}(s_t)$, where $V^{i,\pi^i}(s_t) = \mathbb{E}_{\pi^{-i}} \left[\sum_{\tau=0}^{\infty} \tilde{\gamma}^{\tau} R^i(s_{t+\tau}, a_{t+\tau}^i, a_{t+\tau}^{-i}) \right]$ represents the value of s_t, i.e., the expected total reward collected by i starting from s_t under policy π^i. The notations $a_{t+\tau}^{-i}$ and $s_{t+\tau}$, respectively, represent the predicted future action and state of the opponent at $t + \tau$. In the MPE, both agents achieve their optimal policies. Due to the mutual influence between the value functions of both agents, finding the MPE is normally NP-hard.

3.2 Quantal Level-k Model

The quantal level-k model is one of the most effective iterative reasoning models in predicting human behaviors in simultaneous move games (Wright and Leyton-Brown 2017). It assumes that each human agent has an *intelligence level* that defines his/her reasoning capability. More specifically, the level-0 agents do not perform any strategic reasoning, while quantal level-k $(k \geq 1)$ agents make strategic decisions by treating other agents as quantal level-$(k-1)$ agents. As shown in Fig. 1, the orange agent is a level-1 agent who believes that the blue agent is a level-0 agent. Meanwhile, the blue agent, who is in fact a level-2 agent, treats the orange agent as a level-1 agent when making decisions. The quantal level-k model has therefore reduced the complex circular strategic thinking in classical MGs to finite levels of iterative optimizations. On the basis of an anchoring level-0 policy, the quantal level-k policies of all agents can be defined for all $k = 1, \ldots, k_{\max}$ through sequential and iterative process.

3.3 Cumulative Prospect Theory

The cumulative prospect theory (CPT) is a non-expected utility measure that describes the risk-sensitivity of humans' decision-making processes (Kahneman and Tversky 2013). It can explain many systematic biases of human behaviors that deviate from risk-neutral decisions, such as risk-avoiding/seeking and framing effects.

Definition 1 (CPT value). *For a discrete random variable X satisfying $\sum_{i=-m}^{n} \mathbb{P}(X=x_i)=1$, $x_i \geq x^0$ for $i=0, \cdots, n$, and $x_i < x^0$*

for $i=-m, \cdots, -1$, then the CPT value of X is defined as

$$\mathbb{CPT}(X) = \sum_{i=0}^{n} \tilde{\rho}^+ \left(\mathbb{P}(X=x_i) \right) u^+ (X - x^0)$$
$$- \sum_{i=-m}^{-1} \tilde{\rho}^- \left(\mathbb{P}(X=x_i) \right) u^- (X - x^0), \quad (1a)$$

$$\tilde{\rho}^+ \left(\mathbb{P}(X=x_i) \right) = \left[w^+ \left(\sum_{j=i}^{n} \mathbb{P}(X=x_j) \right) \right.$$
$$\left. - w^+ \left(\sum_{j=i+1}^{n} \mathbb{P}(X=x_j) \right) \right], \quad (1b)$$

$$\tilde{\rho}^- \left(\mathbb{P}(X=x_j) \right) = \left[w^- \left(\sum_{j=-m}^{i} \mathbb{P}(X=x_j) \right) \right.$$
$$\left. - w^- \left(\sum_{j=-m}^{i-1} \mathbb{P}(X=x_j) \right) \right]. \quad (1c)$$

The functions $w^+ : [0,1] \to [0,1]$ and $w^- : [0,1] \to [0,1]$ are two continuous non-decreasing functions and are referred as the probability decision weighting functions. They describe the characteristics of humans to deflate high probabilities and inflate low probabilities. The two functions $u^+ : \mathbb{R}^+ \to \mathbb{R}^+$ and $u^- : \mathbb{R}^- \to \mathbb{R}^+$ are concave utility functions which are, respectively, monotonically non-decreasing and non-increasing. The notation x^0 denotes a "reference point" that separates the value X into gains $(X \geq x^0)$ and losses $(X < x^0)$. Without loss of generality, we set $x^0 = 0$ and omit x^0 in the rest of this paper.

Many experimental studies have shown that representative functional forms for u and w are: $u^+(x)=(x)^{\alpha}$ if $x \geq 0$, and $u^+(x)=0$ otherwise; $u^-(x)=\lambda(-x)^{\beta}$ if $x < 0$, and $u^-(x)=0$ otherwise; $w^+(p)=\frac{p^{\gamma}}{(p^{\gamma}+(1-p)^{\gamma})^{1/\gamma}}$ and $w^-(p) = \frac{p^{\delta}}{(p^{\delta}+(1-p)^{\delta})^{1/\delta}}$. The parameters $\alpha, \beta, \gamma, \delta \in (0,1]$ are model parameters. We adopt these two representative functions in this paper. Section A of the supplementary material illustrates the probability weighting functions and the utility functions.

4 Bounded Risk-Sensitive Markov Game

In this section, we investigate the agents' policies in a new general-sum two-player MG, i.e., the bounded risk-sensitive MG (BRSMG), in which each agent is bounded-rational with a risk-sensitive performance measure.

4.1 Bounded Risk-Sensitive Policies

According to the quantal level-k model described in Section 3.2, a quantal level-k $(k \in \mathbb{N}^+)$ agent *A* obtains its closed-loop game policy by assuming its opponents as level-$(k-1)$ agents and predicting their responses according to the closed-loop quantal level-$(k-1)$ policy in *A*'s decision-making process. Similarly, the closed-loop quantal level-$(k-1)$ policy can be obtained by assuming other agents following a closed-loop quantal level-$(k-2)$ policy. Such an iterative reasoning process traces back to the quantal level-0 policy which is a pure responsive policy. Therefore, on the basis of a selected quantal level-0 policy[1], π^0, we can iteratively solve for the closed-loop quantal policies with higher levels of intelligence, i.e., π^k with $k \geq 1$.

---

[1]Note that the selection of quantal level-0 policy can be different according to applications. We use the notation π^0 to represent a generic quantal level-0 policy and describe the exemplary quantal level-0 policy in Section 6.

If we strictly consider positive rewards and set $x^0=0$, we have the CPT value in (1) reduced to include only $u^+ : \mathbb{R}^+ \to \mathbb{R}^+$ and $\tilde{\rho}^+ : [0,1] \to [0,1]$. In (Lin 2013), it is proved that under such condition, the $\mathbb{CPT}$ measure is a reward transition mapping (Theorem 3.2). Thus, following Section 2 in (Lin 2013), given current state s_t, the discounted future cumulative prospects that a risk-sensitive quantal level-k agent i tries to maximize can be expressed as:

$$\max_{\pi^{i,k}} J_{\pi^{i,k}}(s_t) = \max_{\pi^{i,k}} \mathbb{CPT}_{\pi^{*,-i,k-1}}\Big[R^i(s_t, a_t^i, a_t^{-i}) + \cdots$$
$$+ \tilde{\gamma}^{\tau} \mathbb{CPT}_{\pi^{*,-i,k-1}}\Big[R^i(s_{t+\tau}, a_{t+\tau}^i, a_{t+\tau}^{-i}) + \dots \Big]\Big], \quad (2)$$

where $\pi^{*,-i,k-1}:\mathcal{S}\times A^{-i}\to[0,1]$ denotes the optimal risk-sensitive quantal level-$(k-1)$ policy of agent $-i$ whose level of intelligence is believed to be $(k-1)$ from agent i's perspective. The action $a_{t+\tau}^{-i}$ denotes the predicted action of agent $-i$ sampled from $\pi^{*,-i,k-1}$ at time step $t+\tau$.

We define $\pi^{*,i,k}$ as the optimal policy for (2), i.e., $\pi^{*,i,k}=\arg\max J_{\pi^{i,k}}(s_t)$, and $V^{*,i,k}(s_t)\triangleq J_{\pi^{*,i,k}}(s_t)$ as the optimal CPT value that i could collect following $\pi^{*,i,k}$ starting from s_t. Then, the optimal CPT value at any $s\in\mathcal{S}$ satisfies (Ruszczyński 2010; Lin and Marcus 2013):

$$V^{*,i,k}(s) = \max_{a^i \in A^i} \mathbb{CPT}_{\pi^{*,-i,k-1}}\Big[R^i(s, a^i, a^{-i}) +$$
$$\tilde{\gamma} V^{*,i,k}(s')\Big], s' = \mathcal{T}_{a^{-i}\sim\pi^{*,-i,k-1}}(s, a^i, a^{-i}). \quad (3)$$

We also define the optimal CPT Q-value of agent i as $Q^{*,i,k}(s,a^i)=\mathbb{CPT}_{\pi^{*,-i,k-1}}\big[R^i(s,a^i,a^{-i})+\tilde{\gamma}V^{*,i,k}(s')\big]$. Based on the Boltzmann model (Von Neumann and Morgenstern 2007), we construct $\pi^{*,i,k}$ as

$$\pi^{*,i,k}(s, a^i) = \frac{\exp\big(\beta Q^{*,i,k}(a^i, s)\big)}{\sum_{a'\in A^i} \exp\big(\beta Q^{*,i,k}(a', s)\big)}, \quad (4)$$

where $\beta\geq 0$ defines the level of the agents conforming to the optimal strategy. Without loss of generality, we set $\beta=1$. By iteratively solving (3), the optimal quantal level-k risk-sensitive policy $\pi^{*,i,k}$ for every $i\in\mathcal{P}$ and every $k=1,\dots,k_{\max}$ can be obtained.

4.2 Policy Convergence
Note that the CPT measure in (3) is non-convex and nonlinear, thus the conditions for the convergence of value iteration algorithm for solving (3) needs to be-established.

Theorem 1. *Denote* $\langle s, a^i, a^{-i}\rangle := c_{s,a^i}^{a^{-i}}$ *and normalize* $\tilde{\rho}^i(c_{s,a^i}^{a^{-i}}) := \tilde{\rho}^i(\mathbb{P}(a^{-i}|s,a^i))$ *by*

$$\rho^i(c_{s,a^i}^{a^{-i}}) = \tilde{\rho}^i(c_{s,a^i}^{a^{-i}}) / \sum_{a^{-i'}} \tilde{\rho}^i(c_{s,a^i}^{a^{-i'}}), \quad (5)$$

where $\tilde{\rho}$ refers to $\tilde{\rho}^+$ defined in (1) since we consider only positive rewards. For an arbitrary agent $i\in\mathcal{P}$, if the one-step reward R^i is lower-bounded by R_{min} with $R_{min} \geq 1$, then $\forall s \in \mathcal{S}$ and all intelligence levels with $k=1,2,\cdots$, the dynamic programming problem in (3) can be solved by the following value iteration algorithm (Algorithm 1):

$$V_{m+1}^{i,k}(s) = \max_{a^i \in A^i} \sum_{a^{-i} \in A^{-i}} \rho^i(c_{s,a^i}^{a^{-i}}) u^i\big(R^i(s, a^i, a^{-i}) +$$
$$\tilde{\gamma} V_m^{i,k}(s')\big), \quad s' = \mathcal{T}(s, a^i, a^{-i}). \quad (6)$$

Algorithm 1: Risk-sensitive quanntal level-k policies

Input: Markov Game $\mathcal{G}$, $k_{\max}$, and the anchoring policy π^0.
Output: $\{\pi^{*,i,k}\}$, $i \in \mathcal{P}$ and $k = 1, \dots, k_{\max}$.
for $k = 1 : k_{max}$ **do**
  for $i \in \mathcal{P}$ **do**
    Initialize $V^{i,k}(s), \forall s \in \mathcal{S}$;
    while $V^{i,k}$ *not converged* **do**
      for $s \in \mathcal{S}$ **do**
        $V^{i,k}(s) \leftarrow \mathcal{B}V^{i,k}(s)$;
      end for
    end while
    for $(s, a^i) \in \mathcal{S} \times A^i$ **do**
      Compute $\pi^{*,i,k}(s, a^i)$ based on (4);
    end for
  end for
end for
Return $\{\pi^{*,i,k}\}$, $i \in \mathcal{P}$ and $k \in \mathbb{K}$.

Moreover, as $m \to \infty$, $V_{m+1}^{i,k}$ converges to the optimal value function $V^{*,i,k}(s)$.

Proof. Detailed proof is given in Section B of the supplementary material. Here, we show only the skeleton. As shown in Section 4.1, the iterative format of level-k policies has reduced (3) to a single-agent policy optimization problem with known $\pi^{*,-i,k-1}$ from previous iterations. Hence, we only need to show that the CPT operator defined by $\mathcal{B}V_m^{i,k} = V_{m+1}^{i,k}$ is a contraction when $R_{\min}\geq 1$ for any $k\geq 1$ (Lemma 2 in Section B of the supplementary material). ∎

5 The Inverse Reward Learning Problem
We now consider the inverse learning problem in BRSMGs. Given demonstrated trajectories of two interacting agents who are running the quantal level-k risk-sensitive policies, our goal is to infer agents' rewards, risk-sensitive parameters, and levels of intelligence.

5.1 Formulation of the Inverse Learning Problem
We assume that the one-step rewards for both agents can be linearly parameterized by a group of selected features: $\forall i\in\mathcal{P}, R^i(s,a^i,a^{-i})=(\omega^i)^{\mathsf{T}}\Phi^i(s,a^i,a^{-i})$, where $\Phi^i(s,a^i,a^{-i}):\mathcal{S}\times A^i\times A^{-i}\to\mathbb{R}^d$ is a known feature function that maps a game state s, an action of agent i, and an action of agent $-i$ to a d-dimensional feature vector, and $\omega^i\in\mathbb{R}^d$ is a d-dimensional reward parameter vector. Under such circumstances, we define $\bar{\omega}=(\bar{\gamma},\bar{\omega}^{\mathrm{r}},\bar{k})$, where $\bar{\gamma}=(\gamma^i,\gamma^{-i})$, $\bar{\omega}^{\mathrm{r}}=(\omega^i,\omega^{-i})$, and $\bar{k} = (k^i,k^{-i})$, respectively, represent the parameters in the weighting functions in (1b), the reward parameter vectors, and the levels of intelligence of both agents. Thus, given a set of demonstrated trajectories from the two players in a BRSMG denoted by $\mathcal{D}=\{\xi_1,\cdots,\xi_M\}$ with $\xi=\{(s_0,\bar{a}_0),\dots,(s_{N-1},\bar{a}_{N-1})\}$, $s_t\in\mathcal{S}$, and $\bar{a}_t\in\mathcal{A}$ $(t=0,\dots,N-1)$, the inverse problem aims to retrieve the underlying reward parameters, the risk-sensitive parameter, and the levels of intelligence of the agents, i.e., $\bar{\omega}$, from $\mathcal{D}$. Based on the principle of Maximum Entropy as in (Ziebart et al. 2008), the problem is equivalent to solving the following optimization problem:

$$\max_{\bar{\omega}} \sum_{\xi \in \mathcal{D}} \log \mathbb{P}\left(\xi|\bar{\omega}\right) = \max_{\bar{\omega}} \sum_{\xi \in \mathcal{D}} \log \prod_{t=0}^{N-1} \mathbb{P}(\bar{a}_t|s_t,\bar{\omega}) \quad (7)$$

where $\mathbb{P}(\bar{a}_t|s_t,\bar{\omega})$ is the joint likelihood of agents' actions conditioned on states and parameters. Note that with discrete and finite choices over k, $\log \mathbb{P}(\bar{a}_t|s_t,\bar{\omega})$ can be expressed as

$$\log \mathbb{P}(\bar{a}_t|s_t,\bar{\omega}) = \log \sum_{(k^i,k^{-i})\in\mathbb{K}\times\mathbb{K}} \pi_{(\bar{\gamma},\bar{\omega}^r)}^{*,i,k^i}(s_t,a_t^i)$$
$$\cdot \pi_{(\bar{\gamma},\bar{\omega}^r)}^{*,-i,k^{-i}}(s_t,a_t^{-i})\mathbb{P}(k_t^i)\mathbb{P}(k_t^{-i}), \quad (8)$$

where $\pi_{(\bar{\gamma},\bar{\omega}^r)}^{*,i,k^i}$ and $\pi_{(\bar{\gamma},\bar{\omega}^r)}^{*,-i,k^{-i}}$, respectively, represent the risk-sensitive quantal level-k policies of agent i and agent $-i$ parameterized by $(\bar{\gamma},\bar{\omega}^r)$. The probabilities $\mathbb{P}(k_t^{\pm i})$, $k \in \mathbb{K}$ are the probabilities of the levels of intelligence for agents i and $-i$ at time t.

Inverse learning with hidden state inference. In general, (7) can be solved via gradient ascent. However, via (8), we can see that directly calculating the gradient over k imposes computational challenges for the optimization problem due to the unknown model in $\mathbb{P}(k_t^{\pm i})$. Therefore, we use an approximated solution by substituting $\mathbb{P}(k_t^{\pm i})$ by $\mathbb{P}(k_t^{\pm i}|\xi_{t-1},\bar{\gamma},\bar{\omega}^r)$ which is an estimate of $\mathbb{P}(k_t^{\pm i})$ based on previous trajectories and other model parameters. Intuitively, the intelligence level k for both agents are treated as hidden variables so that we can infer their conditional distributions during the gradient ascent process over $\bar{\omega}=(\bar{\gamma},\bar{\omega}^r)$. Under such an approximation, the gradient of the log-likelihood of a joint trajectory can be analytically obtained in the following sections.

5.2 The Gradient Information

With the proposed approximation above and re-define $\bar{\omega}$ as $(\bar{\gamma},\bar{\omega}^r)$, the log-likelihood of a joint trajectory ξ can be re-written as:

$$\log \mathbb{P}(\xi|\bar{\omega}) = \sum_{t=0}^{N-1} \log \sum_{(k^i,k^{-i})\in\mathbb{K}\times\mathbb{K}} \pi_{\bar{\omega}}^{*,i,k^i}(s_t,a_t^i)$$
$$\cdot \pi_{\bar{\omega}}^{*,-i,k^{-i}}(s_t,a_t^{-i})\mathbb{P}(k^i|\xi_{t-1},\bar{\omega})\mathbb{P}(k^{-i}|\xi_{t-1},\bar{\omega}), \quad (9)$$

where $\mathbb{P}(k|\xi_{t-1},\bar{\omega})$, $k \in \mathbb{K}$, is the posterior belief in an agent's intelligence level inferred based on the joint trajectory history ξ_{t-1} upon time $t-1$. Note that initially, we set $\mathbb{P}(k|\xi_{-1},\bar{\omega})$ as an uniform distribution on $\mathbb{K}$. Then, $\mathbb{P}(k|\xi_{t-1},\bar{\omega})$ can be updated recursively from $t=0$ using Bayesian inference:

$$\mathbb{P}(k|\xi_t,\bar{\omega}) = \frac{\pi_{\bar{\omega}}^{*,i,k}(s_t,a_t^i)\mathbb{P}(k|\xi_{t-1},\bar{\omega})}{\sum_{k'\in\mathbb{K}}\pi_{\bar{\omega}}^{*,i,k'}(s_t,a_t^i,)\mathbb{P}(k'|\xi_{t-1},\bar{\omega})}. \quad (10)$$

From (9) and (10), we can see that the gradient $\partial \log \mathbb{P}(\xi|\bar{\omega})/\partial\bar{\omega}$ depends on two items (details are in Section C of the supplementary material): 1) the gradients of both agents' policies under arbitrary intelligence level $k\in\mathbb{K}$ with respect to $\bar{\omega}$, i.e., $\partial \pi_{\bar{\omega}}^{*,i,k}/\partial\bar{\omega}$, and 2) the gradient of the posterior belief of k with respect to $\bar{\omega}$, i.e., $\partial \log \mathbb{P}(k|\xi_{t-1},\bar{\omega})/\partial\bar{\omega}$.

Gradients of policies. Recall (4), $\partial \pi_{\bar{\omega}}^{*,i,k}/\partial\bar{\omega}$, $\forall i\in\mathcal{P}$ and $k\in\mathbb{K}$, requires the gradient of the corresponding optimal Q function with respect to $\bar{\omega}$, i.e., $\partial Q_{\bar{\omega}}^{*,i,k}/\partial\bar{\omega}$ (detailed derivation is shown in Section C.1 of the supplementary material).

Due to the $\max$ operator in (3), direct differentiation is not feasible. Hence, we use a smooth approximation for the $\max$ function, that is, $\max(x_1,\cdots,x_{n_x})\approx\left(\sum_{i=1}^{n_x}(x_i)^\kappa\right)^{\frac{1}{\kappa}}$ with all $x_i>0$. The parameter $\kappa>0$ controls the approximation error, and when $\kappa\to\infty$, the approximation becomes exact. Therefore, (3) can be re-written as

$$V_{\bar{\omega}}^{*,i,k}(s) = \max_{a^i\in A^i}Q_{\bar{\omega}}^{*,i,k}(s,a^i)$$
$$\approx \left(\sum_{a^i\in A^i}\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^\kappa\right)^{\frac{1}{\kappa}}. \quad (11)$$

Taking derivative of both sides of (11) with respect to $\bar{\omega}$ yields (note that $(\cdot)_{\bar{\omega}}' := \frac{\partial(\cdot)_{\bar{\omega}}}{\partial\bar{\omega}}$):

$$V_{\bar{\omega}}^{',*,i,k}(s)\approx\frac{1}{\kappa}\left(\sum_{a^i\in A^i}\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^\kappa\right)^{\frac{1-\kappa}{\kappa}} \quad (12a)$$
$$\cdot\sum_{a^i\in A^i}\left[\kappa\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^{\kappa-1}\cdot Q_{\bar{\omega}}^{',*,i,k}(s,a^i)\right],$$

$$Q_{\bar{\omega}}^{',*,i,k}(s,a^i) = \sum_{a^{-i}\in A^{-i}}\left(\frac{\partial\rho_{\bar{\omega}}^i}{\partial\bar{\omega}}(c_{s,a^i}^{a^{-i}})u^i\left(R_{\bar{\omega}}^i(s,a^i,a^{-i})\right.\right.$$
$$\left.+\tilde{\gamma}V_{\bar{\omega}}^{*,i,k}(s')\right)+\rho_{\bar{\omega}}^i(c_{s,a^i}^{a^{-i}})\alpha\left(R_{\bar{\omega}}^i(s,a^i,a^{-i})\right. \quad (12b)$$
$$\left.\left.+\tilde{\gamma}V_{\bar{\omega}}^{*,i,k}(s')\right)^{\alpha-1}\left(\frac{\partial R_{\bar{\omega}}^i}{\partial\bar{\omega}}(s,a^i,a^{-i})+\tilde{\gamma}V_{\bar{\omega}}^{',*,i,k}(s')\right)\right).$$

Notice that in (12), $V_{\bar{\omega}}^{',*,i,k}$ is in a recursive format. Hence, we propose below a dynamic programming algorithm to solve for $V_{\bar{\omega}}^{',*,i,k}$ and $Q_{\bar{\omega}}^{',*,i,k}$ at all state and action pairs.

Theorem 2. *If the one-step reward R^i, $i\in\mathcal{P}$, is bounded by $R^i\in[R_{min},R_{max}]$ satisfying $\frac{R_{max}}{R_{min}^{2-\alpha}}\alpha\tilde{\gamma}<1$, then $\partial V_{\bar{\omega}}^{*,i,k}/\partial\bar{\omega}$ can be found via the following value gradient iteration:*

$$V_{\bar{\omega},m+1}^{',i,k}(s) \approx \frac{1}{\kappa}\left(\sum_{a^i\in A^i}\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^\kappa\right)^{\frac{1-\kappa}{\kappa}} \quad (13a)$$
$$\cdot\sum_{a^i\in A^i}\left[\kappa\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^{\kappa-1}\cdot Q_{\bar{\omega},m}^{',i,k}(s,a^i)\right],$$

$$Q_{\bar{\omega},m}^{',i,k}(s,a^i) = \sum_{a^{-i}\in A^{-i}}\left(\frac{\partial\rho_{\bar{\omega}}^i}{\partial\bar{\omega}}(c_{s,a^i}^{a^{-i}})u^i\left(R^i(s,a^i,a^{-i})\right.\right.$$
$$\left.+\tilde{\gamma}V_{\bar{\omega}}^{*,i,k}(s')\right)+\rho_{\bar{\omega}}^i(c_{s,a^i}^{a^{-i}})\alpha\left(R_{\bar{\omega}}^i(s,a^i,a^{-i})\right.$$
$$\left.\left.+\tilde{\gamma}V_{\bar{\omega}}^{*,i,k}(s')\right)^{\alpha-1}\left(\frac{\partial R_{\bar{\omega}}^i}{\partial\bar{\omega}}(s,a^i,a^{-i})+\tilde{\gamma}V_{\bar{\omega},m}^{',i,k}(s')\right)\right). \quad (13b)$$

Moreover, the algorithm converges to $\partial V_{\bar{\omega}}^{*,i,k}/\partial\bar{\omega}$ as $m\to\infty$.

Proof. We first define $\nabla\mathcal{B}V_m^{',i,k} = V_{m+1}^{',i,k}$, and show that the operator $\nabla\mathcal{B}$ is a contraction under the given conditions (derivations of $\partial\rho_{\bar{\omega}}^i/\partial\bar{\omega}$ are shown in Section C.2 of the supplementary material). Then, the statement is proved by induction similar to Theorem 1. More details are given in Section D of the supplementary material. ∎

Gradient of the posterior belief. We summarize the value iteration algorithm that computes the policy gradient in Algorithm 2. The second gradient that we need

Algorithm 2: Gradients of quantal level-k risk-sensitive policies

---

Input: Markov Game model $\mathcal{G}$, highest intelligence level $k_{\max}$, and $\pi^{*,i,k}$, $i \in \mathcal{P}$ and $k = 1, \ldots, k_{\max}$.

Output: $\{\frac{\partial \pi_{\bar{\omega}}^{*,i,k}}{\partial \bar{\omega}}\}$, $i \in \mathcal{P}$ and $k \in \mathbb{K}$.

for $k = 1, \ldots, k_{max}$ **do**

   for $i \in \mathcal{P}$ **do**

      Initialize $V_{\bar{\omega}}^{',i,k}(s)$, $\forall s \in \mathcal{S}$;

      while $V^{',i,k}$ *not converged* **do**

         for $s \in \mathcal{S}$ **do**

            $V^{',i,k}(s) \leftarrow \nabla \mathcal{B} V^{',i,k}(s)$;

         end for

      end while

      for $(s, a^i) \in \mathcal{S} \times A^i$ **do**

         Compute $\frac{\partial \pi_{\bar{\omega}}^{i,k}}{\partial \bar{\omega}}(s, a^i)$ by differentiating Eq. (4) with respect to ω;

      end for

   end for

end for

Return $\{\frac{\partial \pi_{\bar{\omega}}^{i,k}}{\partial \bar{\omega}}\}$, $i \in \mathcal{P}$ and $k \in \mathbb{K}$.

---

to compute is the gradient of the posterior belief in k with respect to $\bar{\omega}$, i.e., $\partial \log \mathbb{P}(k|\xi_{t-1}, \bar{\omega})/\partial \bar{\omega}$. Recalling (10), we have $\partial \log \mathbb{P}(k|\xi_{t-1}, \bar{\omega})/\partial \bar{\omega}$ depending on $\partial \pi_{\bar{\omega}}^{*,i,k}/\partial \bar{\omega}(s_{t-1}, a_{t-1}^i)$ and $\partial \log \mathbb{P}(k|\xi_{t-2}, \bar{\omega})/\partial \bar{\omega}$ for all $k \in \mathbb{K}$. Substituting the gradients of policies obtained through Algorithm 2 in $\partial \log \mathbb{P}(k|\xi_{t-1}, \bar{\omega})/\partial \bar{\omega}$ yields a recursive format from time 0 to time $t-1$, which can be easily computed.

Generalization to other iterative reasoning models. Both Theorem 1 and Theorem 2 naturally extend to other probabilistic iterative reasoning models as long as the optimal policies are iterative and satisfy (3). For instance, the quantal cognitive hierarchy model (Wright and Leyton-Brown 2014) which allows for mixed levels of intelligence can be well applied. Detailed extension and comparison among these models are left to future work.

5.3 The Inverse Learning Algorithm in BRSMG

With the gradient of (9) defined, the gradient ascent algorithm is used to find local optimal parameters in $\bar{\omega}$ that maximize the log-likelihood of the demonstrated joint behaviors of agents in a BRSMG. The algorithm is summarized in Algorithm 3.

Algorithm 3: The inverse learning algorithm

Input: A demonstration set $\mathcal{D}$ and learning rate η

Output: Learned parameters $\bar{\omega}$.

Initialize $\bar{\omega}$.

while *not converged* **do**

   Run Algorithm 1, Algorithm 2

   Compute gradient of the log-likelihood of the demonstration following: $\nabla_{\bar{\omega}} = \sum_{\xi \in \mathcal{D}} \frac{\partial \log (\mathbb{P}(\xi|\bar{\omega}))}{\partial \bar{\omega}}$;

   Update the parameters following: $\bar{\omega} = \bar{\omega} + \eta \nabla_{\bar{\omega}}$;
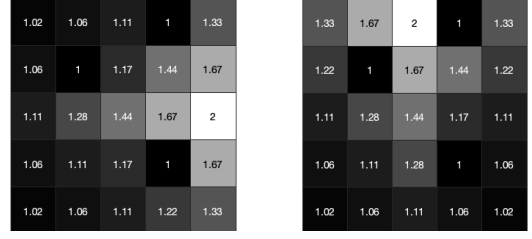
end while

Return: $\bar{\omega}$

Figure 2: The navigation reward maps satisfying $R \geq 1$ (left: the orange agent; right: the blue agent).

6 Experiments

In this section, we utilize a grid-world navigation example to verify the proposed algorithms in both the forward policy design and inverse reward learning problems in a BRSMG. The simulation setup is shown in Fig. 1. Two human agents must exit the room through two different doors while avoiding the obstacles and potential collisions with each other. We assume that the two agents move simultaneously and they can observe the actions and states of each other in the previous time step. Moreover, we let $k_{\max}=2$ in this experiment since psychology studies found that most humans perform at most two layers of strategic thinking (Stahl and Wilson 1995).

6.1 Environment Setup

We define the state as $s=(x^1, y^1, x^2, y^2)$, where x^i and y^i denote the coordinates of the human agent i, $i \in \mathcal{P}$. The two agents share a same action set $A=\{$move left, move right, move up, move down$\}$. In each state, the reward of agent i includes two elements: a navigation reward as shown in Fig. 2 and a safety reward that reflects the penalty for collisions with obstacles or the other agent. We restrict all rewards to be positive, satisfying $R_{\min}=1$. If a collision happens, an agent will collect a fixed reward of 1. If there is no collision, agents receive rewards greater than 1 according to the navigation reward map in Fig. 2.

Recall that a quantal level-0 policy is required to initiate the iterative reasoning process in Algorithm 1. In this work, we use an uncertain follower policy as an exemplary quantal level-0 policy.

Selection of the quanntal level-0 policy. We define the quantal level-0 policy as an uncertain-following policy (i.e., from a quantal level-1 agent's perspective, a quantal level-0 agent is a follower who accommodates the quantal level-1 agent's planned immediate action). Namely, given state s_t and action a^{-i} from the opponent agent (i.e., the leader), the stochastic policy of a level-0 agent i satisfies

$$\pi^{*,i,0}(s_t, a^i|a^{-i}) = \frac{\exp\left(R^i(s_t, a^i, a^{-i})\right)}{\sum_{a' \in A^i} \exp\left(R^i(s_t, a', a^{-i})\right)}, \forall a^i \in A^i.$$
(14)

6.2 Interactions in BRSMG

In this section, we investigate the influence of the risk-sensitive performance measure on agents' policies in a Markov Game by comparing agents' interactive behaviors under risk-neutral and risk-sensitive policies. We set the parameters in the CPT model as $\gamma^{1,2}=0.5$ and $\alpha^{1,2}=0.7$.

Three cases are considered: Case 1 - both agents are quantal level-1 (L1-L1); Case 2 - both agents are quantal level-2 (L2-L2); and Case 3 - one agent is quantal level-1 and the other is quantal level-2 (L1-L2). If both agents exit the environment without collisions and dead-locks, we call it a *success*. We compare the rate of success (RS) of each case under risk-neutral and risk-sensitive policies in 100 simulations with agents starting from different locations.

Impacts of bounded intelligence. First, let us see how a risk-neutral agent behaves under different levels of intelligence. Based on the selected anchoring policy (the level-0 policy) in (14), a risk-neutral quantal level-1 agent will behave quite aggressively since it believes that the other agent is an uncertain-follower. On the contrary, a risk-neutral quantal level-2 agent will perform more conservatively because it believes that the other agent is aggressively executing a quantal level-1 policy. Fig. 3(b) shows an exemplary trajectory of Case 1. We can see that with two level-1 agents, collision happened due to their aggressiveness, i.e., they both assumed that the other would yield. On the other hand, Fig. 3(d) and Fig. 3(f), respectively, show exemplary trajectories of Case 2 and Case 3 with agents starting from the same locations as in the exemplary trajectory in Fig. 3(b). We can see that in both cases, the two agents managed to avoid collisions. In Case 2, both agents behaved more conservatively, and lead to low efficiency (Fig. 3(d)), while in Case 3, both agents behaved as their opponent expected and generated the most efficient and safe trajectories (Fig. 3(f)). To show the statistical results, we conducted 100 simulations for each case with randomized initial states, and the RS is shown in Fig. 3(a) (green). It is shown that similar to what we have observed in the exemplary trajectories, Case 1 lead to the lowest RS, and Case 3 achieved the highest RS. The RS in Case 2 is in the middle because though both agents behaved conservatively, the wrong beliefs over the other's model may still lead to lower RS compared to Case 3.
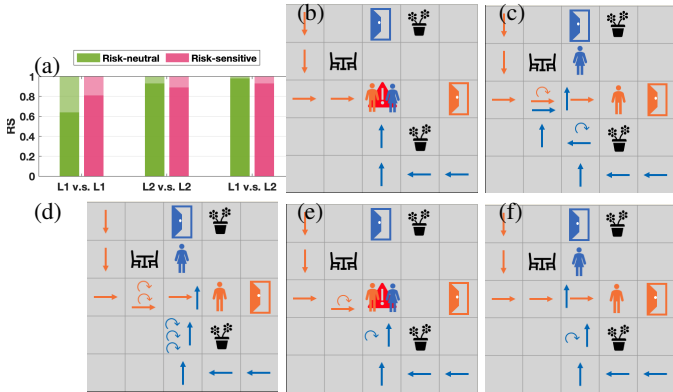

Figure 3: (a) Performance comparison between the bounded risk-neutral policies and the bounded risk-sensitive policies. (b-f) Examples of iterative trajectories (circular arrow denotes the action "stay"); (b) two risk-neutral quantal level-1 agents; (c) two risk-sensitive quantal level-1 agents; (d) two risk-neutral quantal level-2 agents; (e) two risk-sensitive quantal level-2 agents; (f) orange: risk-neutral quantal level-1 agent; blue: risk-neutral quantal level-2 agent.

Impacts of risk sensitivity. In this experiment, we will see how the risk-sensitive CPT model impacts such risk-neutral behaviors. As shown in Fig. 3(a), in Case 1, the risk-sensitive policies help significantly improve the RS of interactions between two quantal level-1 agents, i.e., they performed less aggressively compared to the risk-neutral case. This is because the CPT model makes the quantal level-1 agents underestimate the possibilities of "yielding" from their opponents, and thus lead to more conservative behaviors with higher RS. Such a conclusion can be verified by comparing the exemplary trajectories shown in Fig. 3(b-e). We can see that compared to the risk-neutral case in Fig. 3(b), under the risk-sensitive policy, the blue agent decided to yield to the orange one at the fourth step. At the same time, in Case 2 and Case 3, the risk-sensitivity measure makes the quantal level-2 agents overestimate the possibilities of "yielding" from quantal level-1 agents and generate more aggressive behaviors. An exemplary trajectory is shown in Fig. 3(e). We can see that compared to the risk-neutral quantal level-2 agent in Fig. 3(d), the risk-sensitive quantal level-2 agents waited for less steps and lead to collision. Hence, the RS for both Case 2 and Case 3 are reduced compared to the risk-neutral scenarios, as shown in Fig. 3(a).

6.3 Reward Learning in BRSMG

In this section, we validate Algorithm 3. In the inverse problem, we aim to learn the navigation rewards and the CPT parameter γ of both agents, (i.e., $\bar{\omega}=(\gamma,(\omega^1,\omega^2))$ and $\omega^{1,2} \in \mathbb{R}^{25}$), without prior information on their intelligence levels (we need to infer the intelligence levels simultaneously during the learning based on (7)).

Collecting synthetic expert demonstrations. We first collect some expert demonstrations in the navigation environment via the policies derived in the forward problem in Section 4. Similarly, for generating the demonstrations, we set the parameters of the CPT model as $\gamma^{1,2}=0.5$ and $\alpha^{1,2}=0.7$, and let agents with mixed intelligence levels interact with each other using the risk-sensitive quantal level-k policies. We randomized the initial conditions (initial positions and intelligence levels) of the agents and collected $M=100$ expert demonstrations (i.e., paired navigation trajectories). The approximation parameter κ in Q-value approximation (11) is set to $\kappa = 100$ and the learning rate is set to $\eta = 0.0015$.

Metrics. We evaluate the learning performance via two metrics: the parameter percentage error (PPE), and the policy loss (PL). The PPE of learned parameters $\bar{\omega}^i$ is defined as $|\bar{\omega}^i - \bar{\omega}^{*,i}|/|\bar{\omega}^{*,i}|$ with $\bar{\omega}^{*,i}$ being the ground-truth value. The PL denotes the error between the ground truth quantal level-k policies and the policies obtained using the learned reward functions. It is defined as $\frac{1}{|\mathbb{K} \times \mathcal{S} \times A^i|} \sum_{(k,s,a^i) \in \mathbb{K} \times \mathcal{S} \times A^i} |\pi_{\bar{\omega}}^{*,i,k}(s,a) - \pi_{\bar{\omega}^*}^{*,i,k}(s,a)|$ where $\pi_{\bar{\omega}}^{*,i,k}$ and $\pi_{\bar{\omega}^*}^{*,i,k}$ are, respectively, the quantal level-k policy of agent i under the learned parameter vector $\bar{\omega}$ and the true vector $\bar{\omega}^*$.

Results. Figure 4(a) and Fig. 4(b) show, respectively, the history of averaged PPE over all parameters and the PL during learning. The solid lines represent the means from 25 trials and the shaded areas are the 95% confidence intervals. The PPEs of all parameters are given in Fig. 4(c). We can see
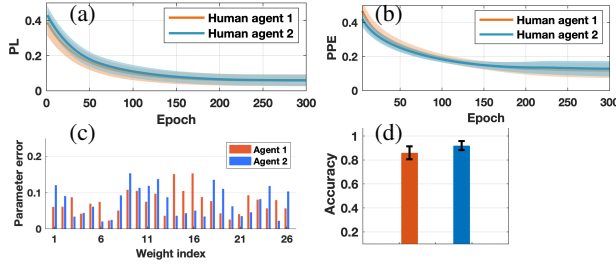
Figure 4: (a-b) Averaged PPE and PL w.r.t. training epochs.(c) PPE of learned parameters. (d) Intelligence level identification accuracy (orange: orange agent; blue: blue agent).

that the proposed inverse learning algorithm can effectively recover both the rewards and the risk-measure parameter γ for both agents, with all PPEs smaller than 15%. In addition, in Fig. 4(d), we show the identification accuracy of the intelligence levels of experts in the data. More specifically, the identified intelligence level of agent i, $i \in \mathcal{P}$, in a demonstration ξ_i is given by $\hat{k}_i = \arg\max_{k \in \mathcal{K}} \mathbb{P}(k|\xi_{N-1})$. We can see that accuracy ratios of 86% and 92% are achieved for the two agents, respectively. Hence, the results show that the proposed inverse reward learning algorithm can effectively recover rewards, risk-parameters and intelligence levels of agents in a BRSMG.

6.4 Performance Comparison with a Baseline Inverse Learning Algorithm

In this section, we compare the performance of the proposed inverse learning algorithm (BRSMG-IRL) against a baseline inverse reward learning algorithm.

The baseline IRL algorithm. The baseline inverse learning algorithm selected in this experiment is a risk-neutral Maximum Entropy IRL (ME-IRL) (Ziebart et al. 2008) without quantal level-k game settings. Instead, it assumes the agents are following Stackelberg strategies similar to (Sun et al. 2019). More specifically, rather than jointly learning rewards for both agents as in the proposed algorithm, in the baseline IRL algorithm, we conduct inverse reward learning separately for each agent. In each IRL formulation, the future trajectories of the opponent agent is assumed to be known, i.e., treating the agent as a leader to the opponent agent.

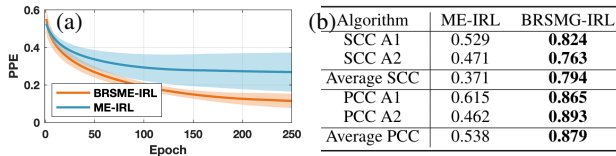| (b) Algorithm	ME-IRL	BRSMG-IRL
SCC A1	0.529	**0.824**
SCC A2	0.471	**0.763**
Average SCC	0.371	**0.794**
PCC A1	0.615	**0.865**
PCC A2	0.462	**0.893**
Average PCC	0.538	**0.879**

Figure 5: Reward learning comparison between our method and a baseline Maximum entropy IRL algorithm. (a) Averaged PPE w.r.t. training epochs. (b) Statistical correlations between the recovered rewards and the ground-truth rewards.

Metrics. In addition to PPE and PL, we also compare the learned rewards with the ground truth rewards using two types of statistical correlations: 1) Pearson's correlation coefficient (PCC) and 2) Spearman's rank correlation coefficient (SCC). PCC characterizes the linear correlation between the

ground truth rewards and the recovered rewards (higher PCC represents higher linear correlations). SCC characterizes the strength and direction of the monotonic relationship between the ground truth rewards and the recovered rewards (higher SCC represents stronger monotonic relationships).

Results. The performance comparison between the proposed approach and the baseline is shown in Fig. 5. We can see that the proposed method can recover more accurate reward values compared to the baseline. This is because the baseline fails to capture the structure biases caused by risk sensitivity and cognitive limitations. Moreover, Fig. 5(b) indicates that the reward values recovered by the proposed method have higher linear correlation and stronger monotonic relationship to the ground-truth rewards.

7 Conclusion

This paper investigated a novel game-theoretic framework: bounded risk-sensitive Markov Game (BRSMG). Drawing on iterative reasoning models and cumulative prospect theory, we embrace that humans have bounded intelligence and maximize risk-sensitive utilities in BRSMGs. Both the forward policy design problem and the inverse reward learning problem under the BRSMG framework have been addressed with theoretical analysis and simulation verification. Simulation results showed that the behaviors of agents demonstrate both risk-averse and risk-seeking phenomena. Moreover, in the inverse reward learning problem, the proposed bounded risk-sensitive inverse learning algorithm outperformed a baseline risk-neutral inverse learning algorithm by effectively recovering not only more accurate rewards but also the intelligence levels and the risk-measure parameters of agents given demonstrations of their interactive behaviors.

In the future, we would like to integrate more human behavior models and risk measures into the BRSMG framework.

References

Amir, R. 2003. Stochastic games in economics and related fields: an overview. In *Stochastic games and applications*, 455–470. Springer.

Ben-Sasson, E.; Kalai, E.; and Kalai, A. 2007. An approach to bounded rationality. In *Advances in Neural Information Processing Systems*, 145–152.

Bu, L.; Babu, R.; De Schutter, B.; et al. 2008. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38(2): 156–172.

Camerer, C. F. 2011. *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Camerer, C. F.; Ho, T.-H.; and Chong, J.-K. 2004. A cognitive hierarchy model of games. *The Quarterly Journal of Economics* 119(3): 861–898.

Chow, Y.; Tamar, A.; Mannor, S.; and Pavone, M. 2015. Risk-sensitive and robust decision-making: a cvar optimization approach. In *Advances in Neural Information Processing Systems*, 1522–1530.

Costa-Gomes, M.; Crawford, V. P.; and Broseta, B. 2001. Cognition and behavior in normal-form games: An experimental study. *Econometrica* 69(5): 1193–1235.

Fisac, J. F.; Bronstein, E.; Stefansson, E.; Sadigh, D.; Sastry, S. S.; and Dragan, A. D. 2019. Hierarchical game-theoretic planning for autonomous vehicles. In *2019 International Conference on Robotics and Automation (ICRA)*, 9590–9596. IEEE.

Goeree, J. K.; and Holt, C. A. 2001. Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review* 91(5): 1402–1422.

Gruver, N.; Song, J.; Kochenderfer, M. J.; and Ermon, S. 2020. Multi-agent Adversarial Inverse Reinforcement Learning with Latent Variables. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 1855–1857.

Halpern, J. Y. 2008. Beyond nash equilibrium: Solution concepts for the 21st century. In *Proceedings of the twenty-seventh ACM symposium on Principles of distributed computing*, 1–10.

Halpern, J. Y.; and Pass, R. 2015. Algorithmic rationality: Game theory with costly computation. *Journal of Economic Theory* 156: 246–268.

Jie, C.; Prashanth, L.; Fu, M.; Marcus, S.; and Szepesvári, C. 2018. Stochastic optimization in a cumulative prospect theory framework. *IEEE Transactions on Automatic Control* 63(9): 2867–2882.

Kahneman, D.; and Tversky, A. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, 99–127. World Scientific.

Kanellopoulos, A.; and Vamvoudakis, K. G. 2019. Non-equilibrium dynamic games and cyber–physical security: A cognitive hierarchy approach. *Systems & Control Letters* 125: 59–66.

Kokolakis, N. T.; Kanellopoulos, A.; and Vamvoudakis, K. G. 2020. Bounded Rational Unmanned Aerial Vehicle Coordination for Adversarial Target Tracking. In *2020 American Control Conference (ACC)*, 2508–2513.

Kwon, M.; Biyik, E.; Talati, A.; Bhasin, K.; Losey, D. P.; and Sadigh, D. 2020. When Humans Aren't Optimal: Robots that Collaborate with Risk-Aware Humans. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 43–52.

Li, N.; Oyler, D. W.; Zhang, M.; Yildiz, Y.; Kolmanovsky, I.; and Girard, A. R. 2018. Game Theoretic Modeling of Driver and Vehicle Interactions for Verification and Validation of Autonomous Vehicle Control Systems. *IEEE Transactions on Control Systems Technology* 26(5): 1782–1797.

Lin, K. 2013. *Stochastic systems with cumulative prospect theory*. Ph.D. thesis.

Lin, K.; and Marcus, S. I. 2013. Dynamic programming with non-convex risk-sensitive measures. In *2013 American Control Conference*, 6778–6783. IEEE.

Majumdar, A.; Singh, S.; Mandlekar, A.; and Pavone, M. 2017. Risk-sensitive Inverse Reinforcement Learning via Coherent Risk Models. In *Robotics: Science and Systems*.

Mazumdar, E.; Ratliff, L. J.; Fiez, T.; and Sastry, S. S. 2017. Gradient-based inverse risk-sensitive reinforcement learning. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, 5796–5801.

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

Pflug, G. C. 2000. Some remarks on the value-at-risk and the conditional value-at-risk. In *Probabilistic constrained optimization*, 272–281. Springer.

Ratliff, L. J.; and Mazumdar, E. 2019. Inverse risk-sensitive reinforcement learning. *IEEE Transactions on Automatic Control* .

Ruszczyński, A. 2010. Risk-averse dynamic programming for Markov decision processes. *Mathematical programming* 125(2): 235–261.

Sadigh, D.; Sastry, S.; Seshia, S. A.; and Dragan, A. D. 2016. Planning for autonomous cars that leverage effects on human actions. In *Robotics: Science and Systems*, volume 2. Ann Arbor, MI, USA.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of go without human knowledge. *Nature* 550(7676): 354–359.

Simaan, M.; and Cruz, J. B. 1973. On the Stackelberg strategy in nonzero-sum games. *Journal of Optimization Theory and Applications* 11(5): 533–555.

Simon, H. A. 1976. From substantive to procedural rationality. In *25 years of economic theory*, 65–86. Springer.

Stahl, D. O.; and Wilson, P. W. 1995. On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior* 10(1): 218–254.

Stahl II, D. O.; and Wilson, P. W. 1994. Experimental evidence on players' models of other players. *Journal of economic behavior & organization* 25(3): 309–327.

Sun, L.; Zhan, W.; Hu, Y.; and Tomizuka, M. 2019. Interpretable modelling of driving behaviors in interactive driving scenarios based on cumulative prospect theory. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 4329–4335. IEEE.

Sun, L.; Zhan, W.; Tomizuka, M.; and Dragan, A. D. 2018. Courteous autonomous cars. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 663–670. IEEE.

Tian, R.; Li, N.; Kolmanovsky, I.; and Girard, A. 2020. Beating humans in a penny-matching game by leveraging cognitive hierarchy theory and Bayesian learning. In *2020 American Control Conference (ACC)*, 4652–4657.

Tian, R.; Li, S.; Li, N.; Kolmanovsky, I.; Girard, A.; and Yildiz, Y. 2018. Adaptive Game-Theoretic Decision Making

for Autonomous Vehicle Control at Roundabouts. In *2018 IEEE Conference on Decision and Control (CDC)*, 321–326.

Tversky, A.; and Kahneman, D. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty* 5(4): 297–323.

Von Neumann, J.; and Morgenstern, O. 2007. *Theory of games and economic behavior (commemorative edition)*. Princeton university press.

Wright, J. R.; and Leyton-Brown, K. 2014. Level-0 meta-models for predicting human behavior in games. In *Proceedings of the fifteenth ACM conference on Economics and computation*, 857–874.

Wright, J. R.; and Leyton-Brown, K. 2017. Predicting human behavior in unrepeated, simultaneous-move games. *Games and Economic Behavior* 106: 16 – 37. ISSN 0899-8256.

Yildiz, Y.; Agogino, A.; and Brat, G. 2014. Predicting pilot behavior in medium-scale scenarios using game theory and reinforcement learning. *Journal of Guidance, Control, and Dynamics* 37(4): 1335–1343.

Yu, L.; Song, J.; and Ermon, S. 2019. Multi-agent adversarial inverse reinforcement learning. In *2Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Ziebart, B. D.; Maas, A. L.; Bagnell, J. A.; and Dey, A. K. 2008. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, 1433–1438. Chicago, IL, USA.