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
University of California, Berkeley
{rantian, litingsun, tomizuka}@berkeley.edu

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 human behaviors such as satisficing with sub-optimal, risk-seeking, and loss-aversion decisions. Drawing on iterative reasoning models and cumulative prospect theory, we propose a new game-theoretic framework, bounded risk-sensitive Markov Game (BRSMG), that captures two aspects of realistic human behaviors: bounded intelligence and risk-sensitivity. General solutions to both the forward policy design problem and the inverse reward learning problem are provided with theoretical analysis and simulation verification. We validate the proposed forward policy design algorithm and the inverse reward learning algorithm in a two-player navigation scenario. The results show that agents demonstrate bounded-intelligence, risk-averse and risk-seeking behaviors in our framework. 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 learning not only more accurate reward values but also the intelligence levels and the risk-measure parameters of agents from demonstrations.

1 Introduction

Markov Game (MG), as an approach to modeling interactions and decision-making processes in multi-agent systems, has been employed in many domains such as economics (Amir 2003), games (Silver et al. 2017), and human-robot/machine interaction (Bu et al. 2008). In classical MGs, agents are commonly assumed to be perfectly rational when computing their policies. For instance, in a two-player MG, 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 such 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 of the game where a Nash equilibrium is reached. Under such assumptions, we can either solve for humans' optimal policies with handcrafted rewards (forward policy design) or learn humans' rewards from demonstrations (inverse reward learning).

However, in real life, humans 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 hunting for the best response, 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 cognitive characteristic as bounded rationality. Among the many developed behavioral models that capture bounded rationality, iterative reasoning models from behavioral game theory (Camerer 2011) are some of the most prominent paradigms. These models do not assume humans perform infinite layers of strategic thinking during interactions but model humans as agents with finite levels of intelligence (bounded rationality). Second, instead of optimizing risk-neutral rewards, humans demonstrate a strong tendency towards risk-sensitive measures when evaluating the outcomes of their actions. They are risk-seeking in terms of gains and risk-averse for losses. Such deviations make it difficult to model realistic human behaviors using classical MGs.

In this work, we aim to establish a new game-theoretic framework (BRSMG) that captures the two aspects of realistic human behaviors discussed above. The incorporation of bounded rationality and risk-sensitivity in classical MGs requires revisiting fundamental concepts in both the forward policy design and the inverse reward learning problem. Standard value iteration and inverse learning algorithms for traditional MGs do not hold any more, and new algorithms should be established to reflect the impact of bounded intelligence and risk sensitivity.

More specifically, in the forward policy design problem, we model humans' bounded intelligence via an instantiation of iterative reasoning models and model the influence of humans' risk sensitivity via cumulative prospect theory (CPT) (Tversky and Kahneman 1992). In the inverse reward learning problem, we develop a bounded risk-sensitive inverse learning algorithm that can recover not only the nominal rewards of agents but also their intelligence levels and risk-measure parameters from demonstrations. *To our best knowledge, our*

*First two authors contributed equally to this work.

work is the first to incorporate both bounded rationality and risk-sensitivity in both the forward problem and the inverse problem of general-sum MGs.

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 processes.
2. The proposed framework makes the first attempt to establish a bridge between inverse reward learning and risk-sensitive iterative reasoning models.
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 risk-sensitive measure parameters. Therefore, our solution provides an interpretable and heterogeneous human behavioral model, which is of critical importance for the development of human-centered robots such as autonomous vehicles.

2 Related Work

Bounded rationality. The influence of bounded rationality in forward policy design problems has been studied in both single-agent and multi-agent settings. One group of studies formulates 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 the QRE model formulate irrational behaviors of humans via sub-optimality, while iterative reasoning models emphasize more on the bounded reasoning depth. Instead of assuming humans perform infinite levels of strategic reasoning, iterative reasoning models only allow for a finite number of strategic reasoning. Iterative reasoning 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; Tian et al. 2020). 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 evaluate humans' decisions. 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 in both single-agent setting (Lin and Marcus 2013; Chow et al. 2015; Mazumdar et al. 2017; Jie et al. 2018; Ratliff and Mazumdar 2019; Kwon et al. 2020) and multi-agents cases (Sun et al. 2019) with a Stackelberg Game assumption.

Inverse reward learning in games. The inverse reward learning problem in games 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) to extend the maximum-entropy inverse reinforcement learning algorithm (Ziebart et al. 2008) in multi-agent settings. (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 compared with QRE, the multi-agent inverse reward learning problem with iterative reasoning models and risk sensitive measure has never been addressed. In this work, we propose the BRSMG framework to fill the gap.

3 Preliminaries

3.1 Classical Markov Game

In this work, we consider two-player Markov Games. We denote a two-player 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 its expected total discounted reward under uncertainties in its 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 $V^{i,\pi^i}(s_t) = \mathbb{E}_{\pi^{-i}}\Big[\sum_{\tau=0}^{\infty} \tilde{\gamma}^\tau R^i(s_{t+\tau}, a^i_{t+\tau}, a^{-i}_{t+\tau})\Big]$ 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 $a^{-i}_{t+\tau}$ and $s_{t+\tau}$, respectively, represent the predicted future action of $-i$ and state of the game at step $t+\tau$. 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 is actually a level-2 agent who 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 a 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*

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

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

$$\tilde{\rho}^{-}\left(\mathbb{P}(X{=}x_i)\right) = \Big[w^{-}\Big(\sum\nolimits_{j=-m}^{i} \mathbb{P}(X{=}x_j)\Big) - w^{-}\Big(\sum\nolimits_{j=-m}^{i-1} \mathbb{P}(X{=}x_j)\Big)\Big]. \tag{1c}$$

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 weighting functions. They describe humans' desire 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.

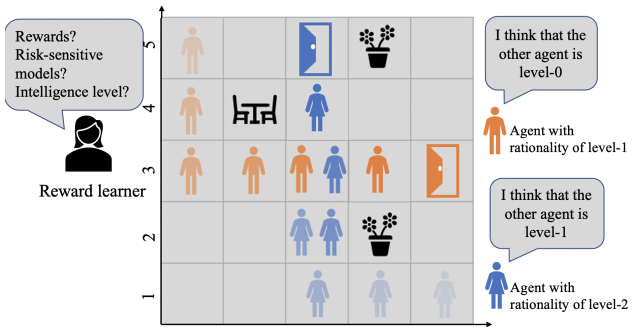

Figure 1: Modeling interactions between humans as a bounded risk-sensitive Markov Game: two human agents aim to exit the room through specified doors without collisions with obstacles and each other. We aim to answer two questions: 1) assuming both humans have bounded intelligence levels and risk-sensitive performance measures, how will their optimal policies differ from those in classical MGs? and 2) how to recover the rewards, intelligence levels, and risk-sensitivity parameters from their demonstrations?

4 Bounded Risk-Sensitive Markov Game

In this section, we investigate agents' policies in a new general-sum two-player MG, i.e., the bounded risk-sensitive MG (BRSMG). In particular, agents in BRSMG are bounded-rational with risk-sensitive performance measures.

4.1 Bounded Risk-Sensitive Policies

According to the quantal level-k model described in Section 3.2, a quantal level-k agent ($k{\in}\mathbb{N}^{+}$) assumes its opponent agent is quantal level-$(k-1)$ agent, predicts its quantal level-$(k-1)$ policy, and quantally best responds to the quantal level-$(k-1)$ policy. Such an iterative reasoning process traces back to the quantal level-0 policy, which is normally a pure responsive policy. Therefore, on the basis of a selected quantal level-0 policy[1], we can sequentially and iteratively solve for the closed-loop quantal level-k policies for every agent and every $k = 1, \ldots, k_{max}$.

If we strictly consider positive rewards and set $x^0{=}0$, we have the CPT value in (1) reduced to a form that includes 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:

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

[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.

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 $a^{-i}_{t+\tau}$ denotes the predicted action of agent $-i$ sampled from $\pi^{*,-i,k-1}$ at time step $t+\tau$.

We define $V^{*,i,k}(s_t)\triangleq J_{\pi^{*,i,k}}(s_t)$ 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):

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

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 re-construct $\pi^{*,i,k}$ as

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

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

In classical MGs, $V^{*,i,k}(s)$ in (3) can be solved via standard value iteration algorithm. 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) need to be-established.

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

$$\rho^i(c^{a^{-i}}_{s,a^i}) = \tilde{\rho}^i(c^{a^{-i}}_{s,a^i})/\sum\nolimits_{a^{-i\prime}}\tilde{\rho}^i(c^{a^{-i\prime}}_{s,a^i}), \tag{5}$$

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):*

$$V^{i,k}_{m+1}(s) = \max_{a^i\in A^i}\sum\nolimits_{a^{-i}\in A^{-i}}\rho^i(c^{a^{-i}}_{s,a^i})u^i\big(R^i(s,a^i,a^{-i})+ \tilde{\gamma}V^{i,k}_m(s')\big),\quad s'=\mathcal{T}(s,a^i,a^{-i}), \tag{6}$$

where u^i *refers to agent* i*'s instance of* u^+ *in* (1)*. Moreover, as* $m\to\infty$*,* $V^{i,k}_{m+1}$ *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 $\mathcal{B}V^{i,k}_m = V^{i,k}_{m+1}$ is a contraction when $R_{\min}\geq 1$ for any $k\geq 1$ (Lemma 2 in Section B of the supplementary material). ■

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 $k\in\mathbb{K}$.

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: $\forall i\in\mathcal{P}, R^i(s,a^i,a^{-i})=(\omega^i)^\intercal\Phi^i(s,a^i,a^{-i})$, 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. 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 parameters, and the levels of intelligence of the agents from $\mathcal{D}$. Based on the principle of Maximum Entropy (Ziebart et al. 2008), the problem is equivalent to solving the following optimization problem:

$$\max_{\bar{\omega}}\sum\nolimits_{\xi\in\mathcal{D}}\log\mathbb{P}(\xi|\bar{\omega}) = \max_{\bar{\omega}}\sum\nolimits_{\xi\in\mathcal{D}}\log\prod\nolimits_{t=0}^{N-1}\mathbb{P}(\bar{a}_t|s_t,\bar{\omega}), \tag{7}$$

where $\mathbb{P}(\bar{a}_t|s_t,\bar{\omega})$ is the joint likelihood of agents' actions conditioned on states and parameters and can be expressed as

$$\log \mathbb{P}(\bar{a}_t|s_t,\bar{\omega}) = \log \pi^{*,i,k^i}_{(\bar{\gamma},\bar{\omega}^r)}(s_t,a_t^i)\pi^{*,-i,k^{-i}}_{(\bar{\gamma},\bar{\omega}^r)}(s_t,a_t^{-i}), \quad (8)$$

where $\pi^{*,i,k^i}_{(\bar{\gamma},\bar{\omega}^r)}$ and $\pi^{*,-i,k^{-i}}_{(\bar{\gamma},\bar{\omega}^r)}$, respectively, represent the risk-sensitive quantal level-k policies of agent i and agent $-i$ induced by parameters $(\bar{\gamma},\bar{\omega}^r)$.

Problem approximation. The optimization (7) can be formulated as a mixed-integer optimization which is infeasible to solve. Therefore, we make the following approximation: we remove $\bar{k}$ from $\bar{\omega}$, and treat $\bar{k}$ as representations of agents' internal states which can be inferred based on agents' demonstrations and most recent estimates of their reward parameters and risk-measure parameters. With that, we evaluate the expected likelihood of $\bar{a}_t$ with respect to the inferred distributions of $\bar{k}$, and solve (7) via gradient ascent.

5.2 The Gradient Information

With the proposed approximation described above, we redefine $\bar{\omega}$ as $(\bar{\gamma},\bar{\omega}^{\mathrm{r}})$, then (8) can be re-written as:

$$\log \mathbb{E}_{\bar{k}|\xi_{t-1},\bar{\omega}}\Big[\mathbb{P}(\bar{a}_t|s_t,\bar{\omega})\Big] = \log \sum\nolimits_{(k^i,k^{-i})\in\mathbb{K}^2} \pi^{*,i,k^i}_{\bar{\omega}}(s_t,a_t^i) \cdot \pi^{*,-i,k^{-i}}_{\bar{\omega}}(s_t,a_t^{-i})\mathbb{P}(k^i|\xi_{t-1},\bar{\omega})\mathbb{P}(k^{-i}|\xi_{t-1},\bar{\omega}), \quad (9)$$

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

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

From (7), (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., $\partial\pi^{*,i,k}_{\bar{\omega}}/\partial\bar{\omega}$ and 2) the gradients of posterior beliefs in agents' intelligence levels with respect to $\bar{\omega}$, i.e., $\partial\log\mathbb{P}(k^i|\xi_{t-1},\bar{\omega})/\partial\bar{\omega}$.

Gradients of policies. Recall (4), $\partial\pi^{*,i,k}_{\bar{\omega}}/\partial\bar{\omega}$, $\forall i\in\mathcal{P}$ and $k\in\mathbb{K}$, requires the gradient of the corresponding optimal Q function with respect to $\bar{\omega}$, i.e., $\partial Q^{*,i,k}_{\bar{\omega}}/\partial\bar{\omega}$ (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, $\max(x_1,\cdots,x_{n_x})\approx\big(\sum_{i=1}^{n_x}(x_i)^\kappa\big)^{\frac{1}{\kappa}}$ 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

$$V^{*,i,k}_{\bar{\omega}}(s) = \max\nolimits_{a^i\in A^i} Q^{*,i,k}_{\bar{\omega}}(s,a^i) \approx \left(\sum\nolimits_{a^i\in A^i}\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^\kappa\right)^{\frac{1}{\kappa}}. \quad (11)$$

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

$$V'^{,*,i,k}_{\bar{\omega}}(s) \approx \frac{1}{\kappa}\left(\sum\nolimits_{a^i\in A^i}\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^\kappa\right)^{\frac{1-\kappa}{\kappa}} \quad (12a)$$
$$\cdot \sum\nolimits_{a^i\in A^i}\Big[\kappa\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa-1}\cdot Q'^{,*,i,k}_{\bar{\omega}}(s,a^i)\Big],$$
$$Q'^{,*,i,k}_{\bar{\omega}}(s,a^i) = \sum\nolimits_{a^{-i}\in A^{-i}}\Bigg(\frac{\partial\rho^i_{\bar{\omega}}}{\partial\bar{\omega}}(c^{a^{-i}}_{s,a^i})u^i\big(R^i_{\bar{\omega}}(s,a^i,a^{-i}) + \tilde{\gamma}V^{*,i,k}_{\bar{\omega}}(s')\big) + \rho^i_{\bar{\omega}}(c^{a^{-i}}_{s,a^i})\alpha\big(R^i_{\bar{\omega}}(s,a^i,a^{-i}) \quad (12b)$$
$$+ \tilde{\gamma}V^{*,i,k}_{\bar{\omega}}(s')\big)^{\alpha-1}\big(\frac{\partial R^i_{\bar{\omega}}}{\partial\bar{\omega}}(s,a^i,a^{-i}) + \tilde{\gamma}V'^{,*,i,k}_{\bar{\omega}}(s')\big)\Bigg).$$

Notice that in (12), $V'^{,*,i,k}_{\bar{\omega}}$ is in a recursive format. Hence, we propose below a dynamic programming algorithm to solve for $V'^{,*,i,k}_{\bar{\omega}}$ and $Q'^{,*,i,k}_{\bar{\omega}}$ 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 $\partial V^{*,i,k}_{\bar{\omega}}/\partial\bar{\omega}$ can be found via the following value gradient iteration:*

$$V'^{,i,k}_{\bar{\omega},m+1}(s) \approx \frac{1}{\kappa}\left(\sum\nolimits_{a^i\in A^i}\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^\kappa\right)^{\frac{1-\kappa}{\kappa}} \quad (13a)$$
$$\cdot \sum\nolimits_{a^i\in A^i}\left[\kappa\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa-1}\cdot Q'^{,i,k}_{\bar{\omega},m}(s,a^i)\right],$$
$$Q'^{,i,k}_{\bar{\omega},m}(s,a^i) = \sum\nolimits_{a^{-i}\in A^{-i}}\Bigg(\frac{\partial\rho^i_{\bar{\omega}}}{\partial\bar{\omega}}(c^{a^{-i}}_{s,a^i})u^i\big(R^i(s,a^i,a^{-i}) + \tilde{\gamma}V^{*,i,k}_{\bar{\omega}}(s')\big) + \rho^i_{\bar{\omega}}(c^{a^{-i}}_{s,a^i})\alpha\big(R^i_{\bar{\omega}}(s,a^i,a^{-i}) \quad (13b)$$
$$+ \tilde{\gamma}V^{*,i,k}_{\bar{\omega}}(s')\big)^{\alpha-1}\big(\frac{\partial R^i_{\bar{\omega}}}{\partial\bar{\omega}}(s,a^i,a^{-i}) + \tilde{\gamma}V'^{,i,k}_{\bar{\omega},m}(s')\big)\Bigg).$$

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

Proof. We first define $\nabla\mathcal{B}V'^{,i,k}_m = V'^{,i,k}_{m+1}$, and show that the operator $\nabla\mathcal{B}$ is a contraction under the given conditions (derivations of $\partial\rho^i_{\bar{\omega}}/\partial\bar{\omega}$ 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 to compute is the gradient of the posterior belief in k with respect to $\bar{\omega}$, i.e., $\partial\log\mathbb{P}(k^i|\xi_{t-1},\bar{\omega})/\partial\bar{\omega}$. Recall (10), we have $\partial\log\mathbb{P}(k^i|\xi_{t-1},\bar{\omega})/\partial\bar{\omega}$ depending on $\partial\pi^{*,i,k}_{\bar{\omega}}/\partial\bar{\omega}(s_{t-1},a^i_{t-1})$ 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

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:$\{\frac{\partial \pi^{*,i,k}_{\bar{\omega}}}{\partial \omega}\}$, $i \in \mathcal{P}$ and $k \in \mathbb{K}$.
for $k = 1, \ldots, k_{max}$ **do**
  for $i \in \mathcal{P}$ **do**
    Initialize $V^{'i,k}_{\bar{\omega}}(s), \forall s \in \mathcal{S}$;
    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 $\frac{\partial \pi^{i,k}_{\bar{\omega}}}{\partial \bar{\omega}}(s, a^i)$ by differentiating Eq. (4) with respect to ω;
    end for
  end for
end for
Return $\{\frac{\partial \pi^{i,k}_{\bar{\omega}}}{\partial \bar{\omega}}\}$, $i \in \mathcal{P}$ and $k \in \mathbb{K}$.

policies are iterative and satisfy (3). For instance, the quantal cognitive hierarchy model (Wright and Leyton-Brown 2014) that 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 (7) defined, the gradient ascent algorithm is used to find local optimal parameters in $\bar{\omega}$ that maximize the log-likelihood of demonstrations in a BRSMG 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: $\nabla_{\bar{\omega}} = \sum_{\xi \in \mathcal{D}} \frac{\partial \log\left(\mathbb{P}(\xi|\bar{\omega})\right)}{\partial \bar{\omega}}$;
  Update the parameters following: $\bar{\omega} = \bar{\omega} + \eta \nabla_{\bar{\omega}}$;
end while
Return: $\bar{\omega}$

6 Experiments

In this section, we utilize a grid-world navigation example to verify the proposed algorithms in both the forward policy design problem and the inverse reward learning problem 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).

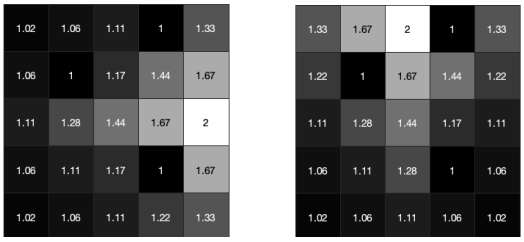

Figure 2: The navigation reward maps satisfying $R \geq 1$ (left: the orange agent; right: the blue agent).

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 $A=\{\text{move left}, \text{move right}, \text{move up}, \text{move down}, \text{stay}\}$. 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. We restrict all rewards to be positive, satisfying $R_{\min}=1$ and $\frac{R_{\max}}{R_{\min}^{2-\alpha}} \alpha \tilde{\gamma} < 1$. 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.

Selection of the quanntal level-0 policy. 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-following policy as an exemplary quantal level-0 policy: 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

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

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 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 opponents 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 belief over the other's model may still lead to lower RS compared to Case 3.

Impacts of risk sensitivity. In this experiment, we will see how the risk-sensitive CPT model impacts 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. This is because the CPT model makes the quantal level-1 agents underestimate the possibilities of "yielding" from their opponents, leading 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 CPT model makes the quantal level-2 agents overestimate the possibilities of "yielding" from quantal level-1 agents, leading to more aggressive behaviors. An exemplary trajectory is shown in Fig. 3(e). We can see that compared to the risk-neutral quantal level-2 agents in Fig. 3(d), the risk-sensitive quantal level-2 agents waited for less steps and collide with each other. Hence, the RS for both Case 2 and Case 3 are reduced compared to the risk-neutral scenarios, as shown in Fig. 3(a).

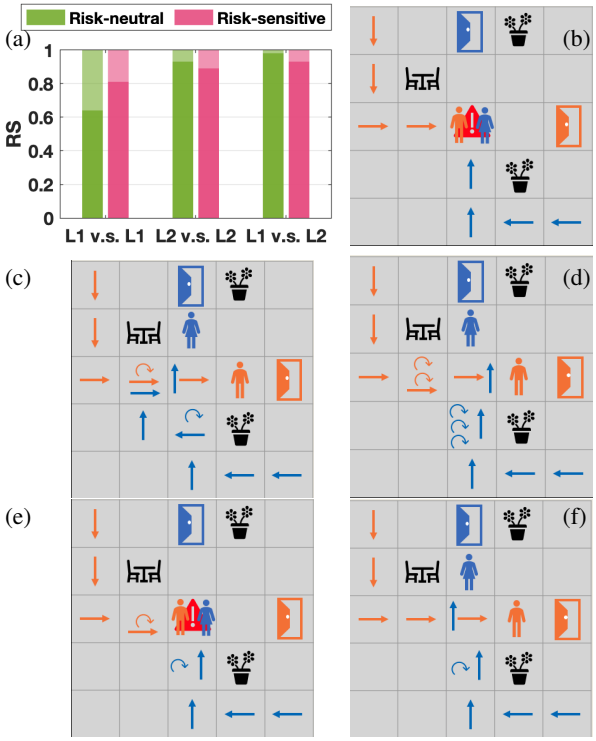

Figure 3: (a) Performance comparison between the bounded risk-neutral policies and the bounded risk-sensitive policies. (b-f) Examples of interactive trajectories (circular arrow denotes the action "stay"); (b) two risk-neutral quantal level-1 agents; (c) two risk-sensitive quantal level-1 agents; (d)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

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).

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 $\frac{1}{|\mathbb{K}\times\mathcal{S}\times A^i|}\sum_{(k,s,a^i)\in\mathbb{K}\times\mathcal{S}\times A^i}|\pi^{*,i,k}_{\bar{\omega}}(s,a)-\pi^{*,i,k}_{\bar{\omega}^*}(s,a)|$ where $\pi^{*,i,k}_{\bar{\omega}}$ and $\pi^{*,i,k}_{\bar{\omega}^*}$ are, respectively, the quantal level-k policy of agent i under the learned parameter vector $\bar{\omega}$ and the true vector $\bar{\omega}^*$.

Results. Fig. 4(a) and Fig. 4(b) show, respectively, the his-

tories of PPE and PL during learning. The solid lines represent the means from 25 trials, and the shaded areas are the 95% confidence intervals. The average errors of each learned parameter are given in Fig. 4(c). We can see that the proposed inverse learning algorithm can effectively recover both agents' rewards and risk-measure parameter γ. In addition, in Fig. 5(a), we show the identification accuracy of the intelligence levels of agents in the data. More specifically, the identified intelligence level of agent i, $i \in \mathcal{P}$, in a demonstration ξ 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. Overall, the results show that the proposed inverse reward learning algorithm can effectively recover rewards, risk-parameters, and intelligence levels of agents in a BRSMG.

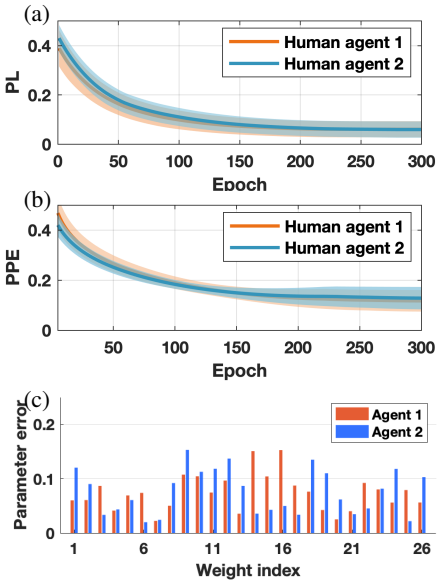

Figure 4: (a-b) Averaged PL and PPE w.r.t. training epochs. (c) Average errors of each learned parameter.

6.4 Performance Comparison with a Baseline

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

The baseline IRL algorithm. The baseline IRL algorithm is a risk-neutral Maximum Entropy IRL algorithm (ME-IRL) without consideration to bounded intelligence, similar to the approach in (Sadigh et al. 2016; Sun et al. 2018; Sun, Zhan, and Tomizuka 2018). Rather than jointly learning both agents' rewards, the baseline runs Maximum Entropy IRL from each agent's perspective separately. In each ego agent' IRL formulation, the interaction is formulated as an open-loop leader-follower game in which the opponent's ground truth trajectory is assumed to be known, making the ego agent a follower to its opponent during learning.

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).

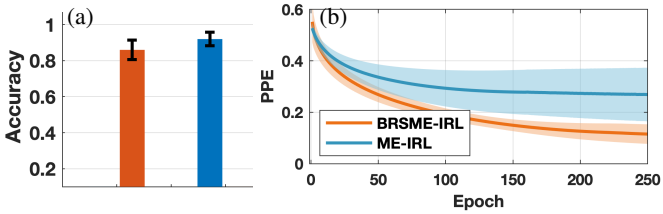

Figure 5: (a): Intelligence level identification accuracy (orange: orange agent; blue: blue agent. (b): Reward learning comparison between our method and a baseline Maximum entropy IRL algorithm.

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**

Table 1: Statistical correlations between the learned reward functions and the ground-truth rewards

Results. The performance comparison between the proposed approach and the baseline is shown in the right plot of 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 structural biases caused by agents' risk sensitivity and bounded intelligence. Moreover, Table 1 indicates that the reward values recovered by the proposed method have a higher linear correlation and stronger monotonic relationship to the ground-truth reward values.

7 Conclusion

Drawing on iterative reasoning models and cumulative prospect theory, we proposed a new game-theoretic framework (BRSMG) that captures two aspects of realistic human behaviors: bounded intelligence and risk-sensitivity. We provided general solutions to both the forward policy design problem and the inverse reward learning problem with theoretical analysis and simulation verification. Our future work will focus on using the proposed framework for practical applications such as learning human driver reward functions from naturalistic driving data.

Acknowledgements

We thank Ruichao Jiang for helpful discussion and feedback.

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.

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, University of Maryland, College Park* .

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

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.

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.; and Tomizuka, M. 2018. Probabilistic Prediction of Interactive Driving Behavior via Hierarchical

Inverse Reinforcement Learning. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2111–2117.

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, N.; Kolmanovsky, I.; Yildiz, Y.; and Girard, A. R. 2020. Game-theoretic modeling of traffic in unsignalized intersection network for autonomous vehicle control verification and validation. *IEEE Transactions on Intelligent Transportation Systems* .

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.

A Cumulative Prospect Theory

The cumulative prospect theory (CPT) is a non-expected utility theory that describes the risk-sensitivity in humans' decision-making processes. In this section, we illustrate the probability weighting function and the utility function, specifically when they are using the following functional forms:

$$w^+(p) = \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{1/\gamma}}, \tag{1}$$

$$w^-(p) = \frac{p^\delta}{(p^\delta + (1-p)^\delta)^{1/\delta}}, \tag{2}$$

$$u(x) = \begin{cases} (x)^\alpha, & \text{if } x \geq 0, \\ \lambda(-x)^\beta, & \text{otherwise.} \end{cases} \tag{3}$$

In Fig. 6 (a), we show an example of the probability weighting function w^+, which describes the characteristics of humans to deflate high probabilities and inflate low probabilities. In Fig. 6 (b), we show an example of the utility function u with $x^0 = 0$ as the reference point.

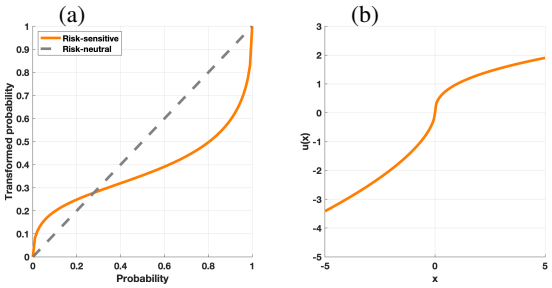

Figure 6: (a) Probability weighting function w^+ with $\gamma = 0.7$ in (1). (b) Utility function with $\alpha = 0.4$, $\beta = 0.6$, λ = -1.3 in (3).

B Policy Convergence

In this section, we show the proof of Theorem 1. To begin with, we show two lemmas that facilitate the proof.

Lemma 1. *If $a \geq 1$, $b \geq 1$, and $\alpha \in (0, 1]$, then $|a^\alpha - b^\alpha| \leq |a - b|$.*

Proof. First, it is clear that the above argument holds when $a = b$. Then, without loss of generality, we assume that $a > b$. We define a differentiable function $f(x) : \mathbb{R} \to \mathbb{R}$ and $f(x) = x^\alpha$. Then, following the mean value theorem, we can have $f(a) - f(b) = (a-b)f'(c)$, where $c \in (b, a)$. Note that $f'(c) = \alpha c^{\alpha-1} \leq 1$ since $\alpha \in (0, 1]$ and $c > 1$. Then we have $f(a) - f(b) \leq (a - b)$, and thus $a^\alpha - b^\alpha \leq a - b$ holds. Similarly, we can have $b^\alpha - a^\alpha \leq b - a$ if $a < b$. ■

Lemma 2. *Assume that $\sum_{a^{-i}\in A^{-i}} \rho^i(c^{a^{-i}}_{s,a^i}) \leq 1$. Then, the CPT Bellman operator $\mathcal{B}V^{i,k}_m(s) = \max_{a^i\in A^i} \sum_{a^{-i}\in A^{-i}} \rho^i(c^{a^{-i}}_{s,a^i}) u^i\big(R^i(s, a^i, a^{-i}) + \tilde{\gamma} V^{i,k}_m(s')\big)$ defined in (6) in Section 4.2 of the submitted manuscript is a $\tilde{\gamma}$-contraction mapping when R_{min} satisfies $R_{min} \geq 1$. That is, for any two value function estimates $V^{i,k}_1$ and $V^{i,k}_2$, we have*

$$\max_{s\in\mathcal{S}} \left|\mathcal{B}V^{i,k}_1(s) - \mathcal{B}V^{i,k}_2(s)\right| \leq \tilde{\gamma} \max_{s\in\mathcal{S}} \left|V^{i,k}_1(s) - V^{i,k}_2(s)\right|. \tag{4}$$

Proof. Define $r^i_{1,2}(c^{a^{-i}}_{s,a^i}) = R^i(s, a^i, a^{-i}) + \tilde{\gamma} V^{i,k}_{1,2}(s')$, then, we can write

$$\begin{aligned}
&\left|\mathcal{B}V^{i,k}_1(s) - \mathcal{B}V^{i,k}_2(s)\right| \\
&= \left| \max_{a^i\in A^i} \sum_{a^{-i}\in A^{-i}} \rho^i(c^{a^{-i}}_{s,a^i}) u^i\big(r^i(c^{a^{-i}}_{s,a^i})\big) \right. \\
&\qquad \left. - \max_{a^i\in A^i} \sum_{a^{-i}\in A^{-i}} \rho^i(c^{a^{-i}}_{s,a^i}) u^i\big(r^i(c^{a^{-i}}_{s,a^i})\big) \right| \\
&\leq \max_{a^i\in A^i} \left| \sum_{a^{-i}\in A^{-i}} \rho^i(c^{a^{-i}}_{s,a^i}) u^i(r^i(c^{a^{-i}}_{s,a^i})) - \right. \\
&\qquad \left. \sum_{a^{-i}\in A^{-i}} \rho^i(c^{a^{-i}}_{s,a^i}) u^i\big(r^i(c^{a^{-i}}_{s,a^i})\big) \right| \qquad (5a) \\
&\leq \max_{a^i\in A^i} \sum_{a^{-i}\in A^{-i}} \rho^i(c^{a^{-i}}_{s,a^i}) \left| u^i\big(r^i_1(c^{a^{-i}}_{s,a^i})\big) - u^i\big(r^i_2(c^{a^{-i}}_{s,a^i})\big) \right| \qquad (5b) \\
&\leq \max_{a^i\in A^i} \sum_{a^{-i}\in A^{-i}} \rho^i(c^{a^{-i}}_{s,a^i}) \left| r^i_1(c^{a^{-i}}_{s,a^i}) - r^i_2(c^{a^{-i}}_{s,a^i}) \right| \qquad (5c) \\
&\leq \max_{a^i\in A^i} \sum_{a^{-i}\in A^{-i}} \rho^i(c^{a^{-i}}_{s,a^i}) \left| \tilde{\gamma} V^{i,k}_1(s') - \tilde{\gamma} V^{i,k}_2(s') \right| \qquad (5d) \\
&= \max_{a^i\in A^i} \tilde{\gamma} \sum_{a^{-i}\in A^{-i}} \rho^i(c^{a^{-i}}_{s,a^i}) \left| V^{i,k}_1(s') - V^{i,k}_2(s') \right|. \qquad (5e)
\end{aligned}$$

Note that the inequality (5)(c) holds based on the definition of u^i defined in Section 3.3 of the submitted manuscript, namely, $u^i(x) = x^\alpha, x \geq 0, \alpha \in (0, 1]$, as shown in (3). Therefore, we have $r^i_{(1,2)}(c^{a^{-i}}_{s,a^i}) = R^i(s, a^i, a^{-i,n}) + \tilde{\gamma} V^{i,k}_{(1,2)}(s') > R_{\min} \geq 1$. With Lemma 1, we have (5)(c). Hence,

$$\begin{aligned}
&\max_{s\in\mathcal{S}} \left|\mathcal{B}V^{i,k}_1(s) - \mathcal{B}V^{i,k}_2(s)\right| \\
&\leq \max_{s\in\mathcal{S}} \max_{a^i\in A^i} \tilde{\gamma} \sum_{a^{-i}\in A^{-i}} \rho^i(c^{a^{-i}}_{s,a^i}) \left| V^{i,k}_1(s') - V^{i,k}_2(s') \right| \\
&\leq \max_{s\in\mathcal{S}} \max_{a^i\in A^i} \tilde{\gamma} \sum_{a^{-i}\in A^{-i}} \rho^i(c^{a^{-i}}_{s,a^i}) \max_{s''} \left| V^{i,k}_1(s'') - V^{i,k}_2(s'') \right| \\
&= \tilde{\gamma} \max_{s''} \left| V^{i,k}_1(s'') - V^{i,k}_2(s'') \right| \max_{s\in\mathcal{S}} \max_{a^i\in A^i} \sum_{a^{-i}\in A^{-i}} \rho^i(c^{a^{-i}}_{s,a^i}) \\
&\leq \tilde{\gamma} \max_{s\in\mathcal{S}} \left| V^{i,k}_1(s) - V^{i,k}_2(s) \right|, \qquad (6a)
\end{aligned}$$

where the inequality (6a) holds since $\sum_{a^{-i}\in A^{-i}} \rho^i(c_{s,a^i}^{a^{-i}}) \leq 1$. Proceeding in this way, we conclude that the CPT operator $\mathcal{B}$ is a $\tilde{\gamma}$-contraction mapping. ■

Now, we restate Theorem 1 in the submitted manuscript and show its proof.

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*

$$\rho^i(c_{s,a^i}^{a^{-i}}) = \begin{cases} \tilde{\rho}^i(c_{s,a^i}^{a^{-i}})/\max_{a^i}\sum_{a^{-i}}\tilde{\rho}^i(c_{s,a^i}^{a^{-i}}), & \text{if } k=1, \\ \tilde{\rho}^i(c_{s,a^i}^{a^{-i}})/\sum_{a^{-i}}\tilde{\rho}^i(c_{s,a^i}^{a^{-i}}), & \text{otherwise.} \end{cases} \tag{7}$$

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) of the submitted manuscript can be solved by the following value iteration algorithm:

$$V_{m+1}^{i,k}(s) = \max_{a^i\in A^i}\sum\nolimits_{a^{-i}\in A^{-i}}\rho^i(c_{s,a^i}^{a^{-i}})u^i\big(R^i(s,a^i,a^{-i}) + \tilde{\gamma}V_m^{i,k}(s'))\big), \quad s'=\mathcal{T}(s,a^i,a^{-i}). \tag{8}$$

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

Proof. We prove the theorem by induction. When $k=1$, $\pi^{*,-i,k-1}=\pi^{-i,0}$, which is defined in Definition 2 in the submitted manuscript. Hence, the dynamic programming problem defined in (3) in the submitted manuscript reduces to a single-agent policy optimization problem since the anchoring policy is known and (3) can be expressed as $V^{*,i,1}(s) = \mathcal{B}V^{*,i,1}(s)$. According to Lemma 2, we have

$$\begin{aligned} &\max_{s\in\mathcal{S}}\left|V_{m+1}^{i,1}(s) - V^{*,i,1}(s)\right| \\ &= \max_{s\in\mathcal{S}}\left|\mathcal{B}V_m^{i,1}(s) - \mathcal{B}V^{*,i,1}(s)\right| \\ &\leq \tilde{\gamma}\max_{s\in\mathcal{S}}\left|V_m^{i,1}(s) - V^{*,i,1}(s)\right| \\ &= \tilde{\gamma}\max_{s\in\mathcal{S}}\left|\mathcal{B}V_{m-1}^{i,1}(s) - \mathcal{B}V^{*,i,1}(s)\right| \\ &\leq \tilde{\gamma}^2\max_{s\in\mathcal{S}}\left|V_{m-1}^{i,1}(s) - V^{*,i,1}(s)\right| \\ &\quad\vdots \\ &\leq \tilde{\gamma}^m\max_{s\in\mathcal{S}}\left|V_1^{i,1}(s) - V^{*,i,k}(s)\right|, \end{aligned} \tag{9}$$

and it is clear that $V_m^{i,1}\to V^{*,i,1}$ as $m\to\infty$. Hence, when $k=1$, the algorithm in (8) can solve for the optimal CPT value and the policy $\pi^{*,i,1}$ can be obtained for all $i\in\mathcal{P}$. Note that $\pi^{*,i,1}$ depends on i's intelligence level.

Next, we will show that for any $k'\in\mathbb{N}^+$ and $k'>1$, assuming the convergence of $V^{*,i,k'-1}$ is proved and the policy $\pi^{*,i,k'-1}$ is obtained for all $i\in\mathcal{P}$, then, similar to (9), we have $V_m^{i,k'}\to V^{*,i,k'}$ as $m\to\infty$.

Again, with the above assumption on $V^{*,i,k'-1}$ and $\pi^{*,i,k'-1}$, we can see that the dynamic programming problem defined in (3) in the submitted manuscript has been reduced to a single-agent optimal policy optimization problem since the opponent's policy $\pi^{*,-i,k'-1}$ is already obtained from previous iteration and thus only depends on agent $-i$'s intelligence level. Moreover, (7) assures that $\sum_{a^{-i}\in\mathcal{A}^{-i}}\rho^i(c_{s,a^i}^{a^{-i}})=1$ for $k'>1$, satisfying the condition in Lemma 2. Hence, via the conclusion from Lemma 2 and (9), we can see that $V^{*,i,k'}$ can be solved by the value iteration algorithm in (8). Then the policy $\pi^{*,i,k'}$ can also be obtained correspondingly for all $i\in\mathcal{P}$. Hence, we have proved that argument in Theorem 1 holds. ■

C Supporting Derivations for the Inverse Learning Algorithm

In this section, we show the detailed derivations that facilitate the computation of the gradient of the objective function (8) in the submitted manuscript.

C.1 Gradient of the log-likelihood of a demonstration

Here, we show the derivation of the gradient of the log-likelihood of a demonstration. Recall (8) in the submitted manuscript, we can write:

$$\frac{\partial\log\big(\mathbb{P}(\xi|\bar{\omega})\big)}{\partial\bar{\omega}} = \sum_{t=0}^{N-1}\frac{1}{\mathfrak{P}_t}\frac{\partial\mathfrak{P}_t}{\partial\bar{\omega}}, \tag{10a}$$

$$\mathfrak{P}_t := \sum_{(k^i,k^{-i})\in\mathbb{K}\times\mathbb{K}}\pi_{\bar{\omega}}^{*,i,k}(s_t,a_t^i)\pi_{\bar{\omega}}^{*,-i,k^{-i}}(s_t,a_t^{-i}) \mathbb{P}(k^i|\xi_{t-1},\bar{\omega})\mathbb{P}(k^{-i}|\xi_{t-1},\bar{\omega}), \tag{10b}$$

$$\begin{aligned} \frac{\partial\mathfrak{P}_t}{\partial\bar{\omega}} = \sum_{(k^i,k^{-i})\in\mathbb{K}\times\mathbb{K}}\Big(&\frac{\partial\pi_{\bar{\omega}}^{*,i,k}}{\partial\bar{\omega}}(s_t,a_t^i)\pi_{\bar{\omega}}^{*,-i,k^{-i}}(s_t,a_t^{-i}) \\ &\cdot\mathbb{P}(k^i|\xi_{t-1},\bar{\omega})\mathbb{P}(k^{-i}|\xi_{t-1},\bar{\omega}) \\ &+\pi_{\bar{\omega}}^{*,i,k}(s_t,a_t^i)\frac{\partial\pi_{\bar{\omega}}^{*,-i,k^{-i}}}{\partial\bar{\omega}}(s_t,a_t^{-i})\mathbb{P}(k^i|\xi_{t-1},\bar{\omega})\mathbb{P}(k^{-i}|\xi_{t-1},\bar{\omega}) \\ &+\pi_{\bar{\omega}}^{*,i,k}(s_t,a_t^i)\pi_{\bar{\omega}}^{*,-i,k^{-i}}(s_t,a_t^{-i})\frac{\partial\mathbb{P}(k^i|\xi_{t-1},\bar{\omega})}{\partial\bar{\omega}}\mathbb{P}(k^{-i}|\xi_{t-1},\bar{\omega}) \\ &+\pi_{\bar{\omega}}^{*,i,k}(s_t,a_t^i)\pi_{\bar{\omega}}^{*,-i,k^{-i}}(s_t,a_t^{-i})\mathbb{P}(k^i|\xi_{t-1},\bar{\omega})\frac{\partial\mathbb{P}(k^{-i}|\xi_{t-1},\bar{\omega})}{\partial\bar{\omega}}\Big). \end{aligned} \tag{10c}$$

From the above expression, we can know that the gradient of the log-likelihood of a demonstration depends on gradient of the agent's policies and the gradient of the posterior belief in agents' intelligence levels. The gradient of the policies can be computed by differentiating (4) in the submitted manuscript, using the value gradient obtained through Algorithm 2 (details are provided in C.2 of this supplementary material). With the policy gradient, the gradient of the posterior belief can be

computed in a recursive fashion as described in Sec 5.2 of the submitted manuscript.

C.2 Supporting derivations for the gradient of policies

In this subsection, we show the derivation of $\frac{\partial \rho^i_{\bar{\omega}}}{\partial \bar{\omega}}(c^{a^{-i}}_{s,a^i})$, the gradient of the probability of an event in CPT model with respect to the parameters $\bar{\omega}$, that is required in (12) in the submitted manuscript to compute the gradient of risk-sensitive quantal level-k policies.

Recall (5) in the submitted manuscript, we can compute $\frac{\partial \rho^i_{\bar{\omega}}}{\partial \bar{\omega}}(c^{a^{-i}}_{s,a^i})$ as follows:

$$\frac{\partial \rho^i_{\bar{\omega}}}{\partial \bar{\omega}}(c^{a^{-i}}_{s,a^i}) = \tag{11}$$
$$\begin{cases} \frac{\frac{\partial \tilde{\rho}^i_{\bar{\omega}}}{\partial \bar{\omega}}(c^{a^{-i}}_{s,a^i}) \max_{a^i} \sum_{a^{-i}} \tilde{\rho}^i(c^{a^{-i}}_{s,a^i}) - \tilde{\rho}^i_{\bar{\omega}}(c^{a^{-i}}_{s,a^i}) \frac{\partial \max_{a^i} \sum_{a^{-i}} \tilde{\rho}^i(c^{a^{-i}}_{s,a^i})}{\partial \omega}}{\left(\max_{a^i} \sum_{a^{-i}} \tilde{\rho}^i(c^{a^{-i}}_{s,a^i})\right)^2}, \\ \qquad\qquad \text{if } k = 1, \\ \frac{\frac{\partial \tilde{\rho}^i_{\bar{\omega}}}{\partial \bar{\omega}}(c^{a^{-i}}_{s,a^i}) \sum_{a^{-i}} \tilde{\rho}^i(c^{a^{-i}}_{s,a^i}) - \tilde{\rho}^i_{\bar{\omega}}(c^{a^{-i}}_{s,a^i}) \frac{\partial \sum_{a^{-i}} \tilde{\rho}^i(c^{a^{-i}}_{s,a^i})}{\partial \omega}}{\left(\sum_{a^{-i}} \tilde{\rho}^i(c^{a^{-i}}_{s,a^i})\right)^2}, \text{ if } k > 1. \end{cases}$$

It can be observed that (11) only depends on $\frac{\partial \tilde{\rho}^i_{\bar{\omega}}}{\partial \bar{\omega}}(c^{a^{-i}}_{s,a^i})$, and the treatment for the $\max$ operator follows the smooth approximation method used in (10) in Section 5.2 of the submitted manuscript. Next, we will show how to compute $\frac{\partial \tilde{\rho}^i_{\bar{\omega}}}{\partial \bar{\omega}}(c^{a^{-i}}_{s,a^i})$.

Note that based on the CPT model defined in (1) in the submitted manuscript, $\tilde{\rho}^i_{\bar{\omega}}(c^{a^{-i}}_{s,a^i})$ is a transform of the probability that agent $-i$ takes the action a^{-i} given current state s (and the action a^i from agent i if $k = 1$, i.e., $\pi^{*,-i,0}(s, a^{-i}, a^i)$, since the anchoring policy depends on the actions from both agents). Without loss of generality, we assume that all $N_A = |A^{-i}|$ utilities induced by agent $-i$'s possible actions are ordered in increasing order, i.e., $0 \le r^i(c^{a^{-i}_1}_{s,a^i}) \le \cdots \le r^i(c^{a^{-i}_{N_A}}_{s,a^i})$, where $r^i(c^{a^{-i}}_{s,a^i}) = u^i\big(R^i(s, a^i, a^{-i}) + \tilde{\gamma} V^{*,i,k}(s')\big)\big)$. Then recall (1b) in the submitted manuscript (since all rewards are positive), for any $g \in \{1, \ldots, N_A\}$, we define

$$p^1(c^{a^{-i}_g}_{s,a^i}) = \begin{cases} \sum_{j=g}^{N_A} \pi^{*,-i,k-1}_{\bar{\omega}}(s, a^{-i}_j, a^i), & k = 1 \\ \sum_{j=g}^{N_A} \pi^{*,-i,k-1}_{\bar{\omega}}(s, a^{-i}_j), & k > 1 \end{cases}, \tag{12a}$$

$$p^2(c^{a^{-i}_g}_{s,a^i}) = \begin{cases} \sum_{j=g+1}^{N_A} \pi^{*,-i,k-1}_{\bar{\omega}}(s, a^{-i}_j, a^{-i}_j), & k = 1 \\ \sum_{j=g+1}^{N_A} \pi^{*,-i,k-1}_{\bar{\omega}}(s, a^{-i}_j), & k > 1 \end{cases}, \tag{12b}$$

then we have $\tilde{\rho}^i_{\bar{\omega}}(c^{a^{-i}_g}_{s,a^i}) = w^{i,+}(p^1) - w^{i,+}(p^2)$. Note that both $w^{i,+}$, p^1 and p^2 depend on the parameter γ^i since $\gamma^i \in \bar{\omega}$, but only p^1 and p^2 depend on the parameter $\bar{\omega}^{-\gamma^i}$ (note that $\bar{\omega}^{-\gamma^i} \triangleq \bar{\omega} \setminus \{\gamma^i\}$), thus we compute $\frac{\partial \tilde{\rho}^i_{\bar{\omega}}}{\partial \gamma^i}$ and $\frac{\partial \tilde{\rho}^{i,n}_{\bar{\omega}}}{\partial \bar{\omega}^{-\gamma^i}}$ separately:

$$\frac{\partial \tilde{\rho}^i_{\bar{\omega}}}{\partial \gamma^i}(c^{a^{-i}_g}_{s,a^i}) = \Phi^1 - \Phi^2, \tag{13a}$$

$$\Phi^j = w^{i,+}(p^j)\Bigg(\log(p^j) + \frac{\gamma^i}{p^j}\frac{\partial p^j}{\partial \gamma^i}(c^{a^{-i}_g}_{s,a^i}) \tag{13b}$$
$$+ \frac{\log\left((p^j)^{\gamma^i} + (1-p^j)^{\gamma^i}\right)}{(\gamma^i)^2} - \frac{1}{\gamma^i\left((p^j)^{\gamma^i} + (1-p^j)^{\gamma^i}\right)} \cdot$$
$$\cdot \Big((p^j)^{\gamma^i}\big(\log(p^j) + \frac{\gamma^i}{p^j}\frac{\partial p^j}{\partial \gamma^i}(c^{a^{-i}_g}_{s,a^i})\big)$$
$$+ (1-p^j)^{\gamma^i}\big(\log(1-p^j) - \frac{\gamma^i}{1-p^j}\frac{\partial p^j}{\partial \gamma^i}(c^{a^{-i}_g}_{s,a^i})\big)\Big)\Bigg), j = 1, 2,$$

$$\frac{\partial \tilde{\rho}^{i,n}_{\bar{\omega}}}{\partial \bar{\omega}^{-\gamma^i}}(c^{a^{-i}_g}_{s,a^i}) \tag{13c}$$
$$= w'^{(+)}_p(p^1)\frac{\partial p^1}{\partial \bar{\omega}^{-\gamma^i}}(c^{a^{-i}_g}_{s,a^i}) - w'^{(+)}_p(p^2)\frac{\partial p^2}{\partial \bar{\omega}^{-\gamma^i}}(c^{a^{-i}_g}_{s,a^i}),$$

where

$$\frac{\partial p^{(1,2)}}{\partial \bar{\omega}}(c^{a^{-i}_g}_{s,a^i}) = \begin{cases} \sum_{j=(g,g+1)}^{N_A} \frac{\partial \pi^{*,-i,0}_{\bar{\omega}}}{\partial \bar{\omega}}(s, a^{-i}_g, a^i), & k = 1 \\ \sum_{j=(g,g+1)}^{N_A} \frac{\partial \pi^{*,-i,k-1}_{\bar{\omega}}}{\partial \bar{\omega}}(s, a^{-i}_g), & k > 1 \end{cases}, \tag{14}$$

and $w'^{(+)}_p$ is the partial derivative with respect to the variables p and can be computed straightforwardly based on the functional form $w^+(p)$ defined in Section 3.3 of the submitted manuscript (or (1) in this supplementary material). $\frac{\partial \pi^{*,-i,0}_{\bar{\omega}}}{\partial \bar{\omega}}$ can be computed straightforwardly based on the definition of the anchoring policy (Definition 2 in the submitted manuscript). We note that when $\frac{\partial \rho^i_{\bar{\omega}}}{\partial \bar{\omega}}(c^{a^{-i}}_{s,a^i})$ is computed in deriving the gradient of agent i's risk-sensitive quantal level-k policy $\frac{\partial \pi^{*,i,k}_{\bar{\omega}}}{\partial \bar{\omega}}$ for $k > 1$, the item $\frac{\partial \pi^{*,-i,k-1}_{\bar{\omega}}}{\partial \bar{\omega}}$ in (14) is already known, since we compute $\frac{\partial \pi^{*,i,k}_{\bar{\omega}}}{\partial \bar{\omega}}$ iteratively and sequentially for $k = 1, 2, \ldots, k_{\max}$ and for $i \in \mathcal{P}$ as shown in Algorithm 2 in the submitted manuscript.

D Value Gradient Convergence

In this section, we show the proof of Theorem 2. To begin with, we first restate Theorem 2 as follows:

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 $\partial V^{*,i,k}_{\bar{\omega}} / \partial \bar{\omega}$ can be found via the following value gradient iteration:*

$$V'^{,i,k}_{\bar{\omega},m+1}(s) \approx \frac{1}{\kappa}\Bigg(\sum\nolimits_{a^i \in A^i}\Big(Q^{*,i,k}_{\bar{\omega}}(s, a^i)\Big)^{\kappa}\Bigg)^{\frac{1-\kappa}{\kappa}}$$
$$\cdot \sum\nolimits_{a^i \in A^i}\Bigg[\kappa\Big(Q^{*,i,k}_{\bar{\omega}}(s, a^i)\Big)^{\kappa-1} \cdot Q'^{,i,k}_{\bar{\omega},m}(s, a^i)\Bigg], \tag{15a}$$

$$Q'^{,i,k}_{\bar{\omega},m}(s, a^i) = \sum\nolimits_{a^{-i} \in A^{-i}}\Bigg(\frac{\partial \rho^i_{\bar{\omega}}}{\partial \bar{\omega}}(c^{a^{-i}}_{s,a^i}) u^i\big(R^i(s, a^i, a^{-i})$$
$$+ \tilde{\gamma} V^{*,i,k}_{\bar{\omega}}(s')\big) + \rho^i_{\bar{\omega}}(c^{a^{-i}}_{s,a^i})\alpha\big(R^i_{\bar{\omega}}(s, a^i, a^{-i})$$
$$+ \tilde{\gamma} V^{*,i,k}_{\bar{\omega}}(s')\big)^{\alpha-1}\big(\frac{\partial R^i_{\bar{\omega}}}{\partial \bar{\omega}}(s, a^i, a^{-i}) + \tilde{\gamma} V'^{,i,k}_{\bar{\omega},m}(s')\big)\Bigg). \tag{15b}$$

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

To prove Theorem 2, we begin with several lemmas that facilitate the proof.

Lemma 3. *When* $u^+(x) = x^\alpha$, $\mathbb{CPT}(\epsilon x) = u^+(\epsilon)\mathbb{CPT}(x), \forall\epsilon>0, \forall x\geq 0$.

Proof. The discrete CPT model is shown in (1) in the submitted manuscript, of which the continuous version is as follows: $\mathbb{CPT}(x)=\int_0^\infty w^+\left(\mathbb{P}\left(u^+(X-x^0)>y\right)\right)dy$, where the u^- term is omitted since we only consider positive rewards. Then we can write:

$$\begin{aligned}\mathbb{CPT}(\epsilon x) &= \int_0^\infty w^+\left(\mathbb{P}\left(u^+(\epsilon x)>y\right)\right)dy\\ &= \int_0^\infty w^+\left(\mathbb{P}\left(u^+(x)>\frac{y}{u^+(\epsilon)}\right)\right)dy. \quad (16)\end{aligned}$$

We let $z := \frac{y}{u^+(\epsilon)}$, then we have $dy = u^+(\epsilon)dz$, and

$$\begin{aligned}\mathbb{CPT}(\epsilon x) &= u^+(\epsilon)\int_0^\infty w^+\left(\mathbb{P}\left(u^+(x)>z\right)\right)dz\\ &= u^+(\epsilon)\mathbb{CPT}(x). \quad (17)\end{aligned}$$

■

Lemma 4. *For an arbitrary agent* $i\in\mathcal{P}$*, if* i*'s one-step reward is lower-bounded by* R_{min} *and upper-bounded by* R_{max}*, then* $\forall k\in\mathbb{N}^+$*, we have* $V_{max}^{i,k}\leq\frac{R_{max}}{R_{min}}V_{min}^{i,k}$.

Proof. We define $\theta=\frac{R_{\max}}{R_{\min}}$, then according to (2) in the submitted manuscript, $V_{\max}^{i,k}$ can only be achieved if agent i collects the maximum one-step reward at every step. Similarly, $V_{\min}^{i,k}$ can only be achieved if agent i collects the minimum one-step reward at every step. Hence, we have

$$\begin{aligned}&V_{\max}^{i,k} \quad (18)\\ &=\mathbb{CPT}_{\pi^*,-i,k-1}\left[R_{\max}+\tilde{\gamma}\mathbb{CPT}_{\pi^*,-i,k-1}\left[R_{\max}+\dots\right]\right]\\ &=\mathbb{CPT}_{\pi^*,-i,k-1}\left[\theta R_{\min}+\tilde{\gamma}\mathbb{CPT}_{\pi^*,-i,k-1}\left[\theta R_{\min}+\dots\right]\right].\end{aligned}$$

Since $\mathbb{CPT}_{\pi^*,-i,k-1}\left[\theta R_{\min}\right] = u^i(\theta)\mathbb{CPT}_{\pi^*,-i,k-1}\left[R_{\min}\right] \leq \theta\mathbb{CPT}_{\pi^*,-i,k-1}\left[R_{\min}\right]$ based on Lemma 3 and the fact that $u^i(\theta)=u^+(\theta)=\theta^\alpha\leq\theta$, then we can have

$$\begin{aligned}&V_{\max}^{i,k}\\ &=\mathbb{CPT}_{\pi^*,-i,k-1}\left[\theta R_{\min}+\tilde{\gamma}\mathbb{CPT}_{\pi^*,-i,k-1}\left[\theta R_{\min}+\dots\right]\right]\\ &\leq\mathbb{CPT}_{\pi^*,-i,k-1}\left[\theta R_{\min}+\theta\tilde{\gamma}\mathbb{CPT}_{\pi^*,-i,k-1}\left[R_{\min}+\dots\right]\right]\\ &\leq\theta\mathbb{CPT}_{\pi^*,-i,k-1}\left[R_{\min}+\tilde{\gamma}\mathbb{CPT}_{\pi^*,-i,k-1}\left[R_{\min}+\dots\right]\right]\\ &=\frac{R_{\max}}{R_{\min}}V_{\min}^{i,k}. \quad (19)\end{aligned}$$

■

Lemma 5. *Recall* (15)*(a) in Theorem 2 in this supporting material, we define an operator* $\nabla\mathcal{B}V_m^{',i,k}=V_{m+1}^{',i,k}$, $\forall i,\in\mathcal{P}$, $\forall k\in\mathbb{N}^+$*. Then, the operator* $\nabla\mathcal{B}$ *is a* $\bar{\gamma}$*-contraction mapping if the one-step reward* R^i *is bounded by* $R^i\in[R_{min},R_{max}]$ *satisfying* $\bar{\gamma}=\frac{R_{max}}{R_{min}^{2-\alpha}}\alpha\tilde{\gamma}<1$*, that is, for any value function gradient estimates* $V_{\bar{\omega},1}^{',i,k}$ *and* $V_{\bar{\omega},2}^{',i,k}$*, we have*

$$\max_{s\in\mathcal{S}}\left|\nabla\mathcal{B}V_{\bar{\omega},1}^{',i,k}(s)-\nabla\mathcal{B}V_{\bar{\omega},2}^{',i,k}(s)\right|\leq\bar{\gamma}\max_{s\in\mathcal{S}}\left|V_{\bar{\omega},1}^{',i,k}(s)-V_{\bar{\omega},2}^{',i,k}(s)\right|. \quad (20)$$

Proof. Recall (15)(a) in this supporting material, we can write

$$\begin{aligned}V_{\bar{\omega},m+1}^{',i,k}(s)&\approx\frac{1}{\kappa}\left(\sum\nolimits_{a^i\in A^i}\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^\kappa\right)^{\frac{1-\kappa}{\kappa}}\\ &\cdot\sum\nolimits_{a^i\in A^i}\left[\kappa\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^{\kappa-1}\cdot Q_{\bar{\omega},m}^{',i,k}(s,a^i)\right]\\ &=V_{\bar{\omega}}^{*,i,k}\frac{1}{\kappa}\left(\sum\nolimits_{a^i\in A^i}\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^\kappa\right)^{-1}\\ &\cdot\sum\nolimits_{a^i\in A^i}\left[\kappa\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^{\kappa-1}\cdot Q_{\bar{\omega},m}^{',i,k}(s,a^i)\right]. \quad (21)\end{aligned}$$

Then, we have the following:

$$\begin{aligned}&\left|\nabla\mathcal{B}V_{\bar{\omega},1}^{',i,k}(s)-\nabla\mathcal{B}V_{\bar{\omega},2}^{',i,k}(s)\right|\\ &=\left|V_{\bar{\omega}}^{*,i,k}\frac{1}{\kappa}\left(\sum\nolimits_{a^i\in A^i}\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^\kappa\right)^{-1}\right.\\ &\left.\cdot\sum_{a^i\in A^i}\left[\kappa\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^{\kappa-1}\left(Q_{\bar{\omega},1}^{',i,k}(s,a^i)-Q_{\bar{\omega},2}^{',i,k}(s,a^i)\right)\right]\right|\\ &=\left|\sum_{a^i\in A^i}\left[V_{\bar{\omega}}^{*,i,k}\frac{1}{\kappa}\left(\sum\nolimits_{a^i\in A^i}\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^\kappa\right)^{-1}\right.\right.\\ &\left.\left.\cdot\kappa\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^{\kappa-1}\left(Q_{\bar{\omega},1}^{',i,k}(s,a^i)-Q_{\bar{\omega},2}^{',i,k}(s,a^i)\right)\right]\right|\\ &=\left|\sum_{a^i\in A^i}\left[\frac{V_{\bar{\omega}}^{*,i,k}(s)\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^{\kappa-1}}{\sum_{a^i\in A^i}\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^\kappa}\right.\right. \quad (22a)\\ &\left.\left.\cdot\left(Q_{\bar{\omega},1}^{',i,k}(s,a^i)-Q_{\bar{\omega},2}^{',i,k}(s,a^i)\right)\right]\right|\\ &\leq\sum_{a^i\in A^i}\left[\frac{V_{\bar{\omega}}^{*,i,k}(s)\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^{\kappa-1}}{\sum_{a^i\in A^i}\left(Q_{\bar{\omega}}^{*,i,k}(s,a^i)\right)^\kappa}\right.\\ &\left.\cdot\left|Q_{\bar{\omega},1}^{',i,k}(s,a^i)-Q_{\bar{\omega},2}^{',i,k}(s,a^i)\right|\right]. \quad (22b)\end{aligned}$$

Recall (15)(b) in this supplementary material, we can have the following

$$
\begin{aligned}
&\left|Q^{'}_{\bar{\omega},1}{}^{,i,k}(s,a^i)-Q^{'}_{\bar{\omega},2}{}^{,i,k}(s,a^i)\right| \\
&=\left|\sum_{a^{-i}\in A^{-i}}\left(\rho^i_{\bar{\omega}}(c^{a^{-i}}_{s,a^i})\alpha\big(R^i_{\bar{\omega}}(s,a^i,a^{-i})\right.\right. \\
&\left.\left.+\tilde{\gamma}V^{*,i,\mathrm{k}}_{\bar{\omega}}(s')\big)^{\alpha-1}\tilde{\gamma}\Big(V^{'}_{\bar{\omega},1}{}^{,i,k}(s')-V^{'}_{\bar{\omega},2}{}^{,i,k}(s')\Big)\right)\right| \\
&\le \tilde{\gamma}\alpha(R_{\min}+\tilde{\gamma}V^{*,i,k}_{\min})^{\alpha-1} \\
&\cdot\sum_{a^{-i}\in A^{-i}}\left(\rho^i_{\bar{\omega}}(c^{a^{-i}}_{s,a^i})\left|V^{'}_{\bar{\omega},1}{}^{,i,k}(s')-V^{'}_{\bar{\omega},2}{}^{,i,k}(s')\right|\right) \qquad (23a)\\
&\le \max_{s''\in\mathcal{S}}\tilde{\gamma}\alpha R_{\min^{\alpha-1}}\left|V^{'}_{\bar{\omega},1}{}^{,i,k}(s'')-V^{'}_{\bar{\omega},2}{}^{,i,k}(s'')\right| \\
&\cdot\sum_{a^{-i}\in A^{-i}}\rho^i_{\bar{\omega}}(c^{a^{-i}}_{s,a^i}) \qquad (23b)\\
&\le \max_{s''\in\mathcal{S}}\tilde{\gamma}\alpha R^{\alpha-1}_{\min}\left|V^{'}_{\bar{\omega},1}{}^{,i,k}(s'')-V^{'}_{\bar{\omega},2}{}^{,i,k}(s'')\right|, \qquad (23c)
\end{aligned}
$$

where the inequality (23)(a) holds since $\alpha\in(0,1]$, the inequality (23)(c) holds since $\sum_{a^{-i}\in A^{-i}}\rho^i_{\bar{\omega}}(c^{a^{-i}}_{s,a^i})\le 1$, which is governed by (5) in the submitted manuscript. We substitute (23) into (22), then we have

$$
\begin{aligned}
&\left|\nabla\mathcal{B}V^{'}_{\bar{\omega},1}{}^{,i,k}(s)-\nabla\mathcal{B}V^{'}_{\bar{\omega},2}{}^{,i,k}(s)\right| \\
&\le\sum_{a^i\in A^i}\left[\frac{V^{*,i,k}_{\bar{\omega}}(s)\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa-1}}{\sum_{a^i\in A^i}\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa}}\right. \\
&\left.\cdot\max_{s''\in\mathcal{S}}\tilde{\gamma}\alpha R^{\alpha-1}_{\min}\left|V^{'}_{\bar{\omega},1}{}^{,i,k}(s'')-V^{'}_{\bar{\omega},2}{}^{,i,k}(s'')\right|\right] \\
&=\max_{s''\in\mathcal{S}}\tilde{\gamma}\alpha R^{\alpha-1}_{\min}\left|V^{'}_{\bar{\omega},1}{}^{,i,k}(s'')-V^{'}_{\bar{\omega},2}{}^{,i,k}(s'')\right| \\
&\cdot\sum_{a^i\in A^i}\frac{V^{*,i,k}_{\omega}(s)\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa-1}}{\sum_{a^i\in A^i}\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa}}. \qquad (24)
\end{aligned}
$$

Also note that

$$
\begin{aligned}
&\sum_{a^i\in A^i}\frac{V^{*,i,k}_{\omega}(s)\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa-1}}{\sum_{a^i\in A^i}\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa}} \\
&\le\sum_{a^i\in A^i}\frac{V^{*,i,k}_{\omega,\max}\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa-1}}{\sum_{a^i\in A^i}\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa}} \qquad (25a)\\
&\le\frac{\sum_{a^i\in A^i}\frac{R_{\max}}{R_{\min}}V^{*,i,k}_{\omega,\min}\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa-1}}{\sum_{a^i\in A^i}\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa}} \qquad (25b)\\
&=\frac{R_{\max}}{R_{\min}}\frac{\sum_{a^i\in A^i}V^{*,i,k}_{\omega,\min}\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa-1}}{\sum_{a^i\in A^i}\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa}} \qquad (25c)\\
&\le\frac{R_{\max}}{R_{\min}}, \qquad (25d)
\end{aligned}
$$

where the inequality (25) (b) holds based on Lemma 4, and the inequality (25)(d) holds since

$$
\begin{aligned}
&\sum_{a^i\in A^i}V^{*,i,k}_{\omega,\min}\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa-1}-\sum_{a^i\in A^i}\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa} \\
&=\sum_{a^i\in A^i}\Big(Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big)^{\kappa-1}\Big(V^{*,i,k}_{\omega,\min}-Q^{*,i,k}_{\bar{\omega}}(s,a^i)\Big) \\
&\le 0. \qquad (26)
\end{aligned}
$$

Now, we substitute (25) into (24), and then we can write

$$
\begin{aligned}
&\max_{s\in\mathcal{S}}\left|\nabla\mathcal{B}V^{'}_{\bar{\omega},1}{}^{,i,k}(s)-\nabla\mathcal{B}V^{'}_{\bar{\omega},2}{}^{,i,k}(s)\right| \\
&\le\max_{s''\in\mathcal{S}}\tilde{\gamma}\alpha R^{\alpha-1}_{\min}\left|V^{'}_{\bar{\omega},1}{}^{,i,k}(s'')-V^{'}_{\bar{\omega},2}{}^{,i,k}(s'')\right|\max_{s\in\mathcal{S}}\frac{R_{\max}}{R_{\min}} \\
&=\frac{R_{\max}}{R^{2-\alpha}_{\min}}\alpha\tilde{\gamma}\max_{s\in\mathcal{S}}\left|V^{'}_{\bar{\omega},1}{}^{,i,k}(s)-V^{'}_{\bar{\omega},2}{}^{,i,k}(s)\right|. \qquad (27)
\end{aligned}
$$

Proceeding in this way, we conclude that the operator $\nabla\mathcal{B}$ is a $\bar{\gamma}$-contraction mapping, where $\bar{\gamma}=\frac{R_{\max}}{R^{2-\alpha}_{\min}}\alpha\tilde{\gamma}<1$. ■

Now, we show the proof of Theorem 2.

Proof. We first define $\nabla\mathcal{B}V^{'}_{m}{}^{,i,k}=V^{'}_{m+1}{}^{,i,k}$, and then Lemma 5 shows that the operator $\nabla\mathcal{B}$ is a contraction under the given conditions. Then, the statement is proved by induction in a similar way as for Theorem 1, and thus is omitted. ■