TRACTABLE MULTI-AGENT REINFORCEMENT LEARNING THROUGH BEHAVIORAL ECONOMICS

Eric Mazumdar*, **Kishan Panaganti***, **& Laixi Shi***
Department of Computing and Mathematical Sciences
California Insitute of Technology
Pasadena, CA, USA
{mazumdar,kpb,laixis}@caltech.edu

ABSTRACT

A significant roadblock to the development of principled multi-agent reinforcement learning (MARL) algorithms is the fact that desired solution concepts like Nash equilibria may be intractable to compute. We show how one can overcome this obstacle by introducing concepts from behavioral economics into MARL. To do so, we imbue agents with two key features of human decision-making: risk aversion and bounded rationality. We show that introducing these two properties into games gives rise to a class of equilibria—risk-averse quantal response equilibria (RQE)—which are tractable to compute in *all* n-player matrix and finite-horizon Markov games. In particular, we show that they emerge as the endpoint of no-regret learning in suitably adjusted versions of the games. Crucially, the class of computationally tractable RQE is independent of the underlying game structure and only depends on agents' degrees of risk-aversion and bounded rationality. To validate the expressivity of this class of solution concepts we show that it captures peoples' patterns of play in a number of 2-player matrix games previously studied in experimental economics. Furthermore, we give a first analysis of the sample complexity of computing these equilibria in finite-horizon Markov games when one has access to a generative model. We validate our findings on a simple multi-agent reinforcement learning benchmark. Our results open the doors for to the development of new decentralized multi-agent reinforcement learning algorithms.

1 INTRODUCTION

Machine learning algorithms are increasingly being deployed in dynamic environments in which they interact with other agents like people or other algorithms. Often, these agents have their own goals that may not be aligned with those of the algorithm—making the interactions *strategic*. These interactions are naturally modeled as *games* between rational agents and their ubiquity has driven a surge of interest in learning in games (Cesa-Bianchi and Lugosi, 2006) and multi-agent reinforcement learning (Zhang et al., 2021a) in recent years. Indeed, real-world applications of these problems range from the decentralized control of the smart grid (Mohsenian-Rad et al., 2010), autonomous driving (Kannan et al., 2017), and financial trading (Wellman and Wurman, 1998) to problems of aligning large language models (Munos et al., 2024) and agentic AI (Verma et al., 2024).

When viewed through the lens of game theory, many of these problems can be cast as problems of *equilibrium computation* under varying information structures, where the equilibrium represents a stable outcome for rational agents. The most common equilibrium concept is that of a Nash equilibrium (NE) (Nash, 1950): a solution under which no rational agent has an incentive to unilaterally seek to improve their outcome. Despite its popularity as a solution concept, computing a NE outside of highly structured games is known to be computationally intractable even for two-player matrix games (Daskalakis, 2013). Coupled with a host of negative results on their computation using gradient-based algorithms (Mertikopoulos et al., 2018; Mazumdar et al., 2020), converging to Nash is increasingly viewed as an unreasonable goal for decentralized reinforcement learning algorithms.

*Alphabetical order.

While relaxations of NE like (coarse) correlated equilibria (CCE) are known to be more tractable to compute—and thus a more attainable goal for learning algorithms— they also have their limitations. Indeed, while CCE arise out of the use of no-regret learning algorithms (Cesa-Bianchi and Lugosi, 2006), the set of CCE can be large (exacerbating the problem of equilibrium selection that arises with NE) and may have support on strictly dominated strategies (Viossat and Zapechelnyuk, 2013), which means that they cannot necessarily be rationalized by individual agents in isolation (Dekel and Fudenberg, 1990). Furthermore, a dynamic versions of CCE—stationary Markov CCE—can also be intractable to compute in general-sum Markov games (Daskalakis et al., 2023b).

Beyond these hardness results, solution concepts like NE and CCE also fail to be predictive of what strategies people play in games (McKelvey and Palfrey, 1995; Erev and Roth, 1998), with people being observed to be imperfect optimizers (Goeree and Holt, 1999; Capra et al., 2002) and risk-averse (Goeree et al., 2003) when confronted with game theoretic scenarios. This aligns with celebrated work in behavioral economics and mathematical psychology which has repeatedly shown that dominant features of human decision-making are a failure to perfectly optimize (Luce, 1959) and risk-aversion (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992).

The first observation is often referred to as *bounded rationality* which posits that individuals are naturally prone to making mistakes and often fail to be perfectly optimal (Luce, 1959). This is often captured in games through the idea of a *quantal response equilibrium* (QRE) (McKelvey and Palfrey, 1995). The second observation can be attributed to the fact that players typically face uncertainty and risk in their decisions. These arise from environmental uncertainties like unknown future events, noise, or even the mere presence of other players. This can lead people to prefer risk-averse strategies, i.e., strategies which give more certain outcomes at the cost of lower expected returns (Gollier, 2001). Interestingly, there is experimental evidence that neither of these properties alone can account for people's patterns of play observed in controlled experiments (Goeree et al., 2003; Goeree and Offerman, 2002), and that models of decision-making that incorporate *both* of these features have the best predictive power (Goeree et al., 2003).

Contributions: Motivated by these findings, in this paper we study games in which players are risk-averse and have bounded rationality and study the computational properties of the natural equilibrium concept: a risk-averse quantal response equilibrium (RQE). At first glance, the introduction of these features of human decision-making into games introduces non-linearities that break existing game-theoretic structures. Indeed we show how neither of these features alone give rise to a single class of equilibria that can efficiently computed across all games. However, by relying on dual formulations of risk we show how—for a large range of degrees of risk-aversion and bounded rationality—a class of RQE are computationally tractable in *all* n-player matrix and finite horizon Markov games. Importantly, these conditions are *independent* of the underlying game and only depend on the class of risk metrics and quantal responses under consideration. Thus, RQE not only capture important features of human decision-making but are also more amenable to computation than QRE or NE in matrix and finite-horizon Markov games. This opens the door for the development of theoretically principled decentralized algorithms for multi-agent reinforcement learning (MARL) centered around RQE as opposed to CCE or NE. We refer readers to the discussion of related works in Appendix A for a more in-depth discussion of different equilibria concepts in the context for our work.

To emphasize the practical relevance of this theoretical result we show how the regime for which RQE are computationally tractable captures real-world data from behavioral economics on people's observed pattern of play in 13 different games. This is illustrated in Fig. 1, where the blue region represents the set of RQE that are computationally tractable which is a function of agents' degree of risk-aversion (τ_1/τ_2), and level of bounded rationality (ϵ_1/ϵ_2).

Altogether, our results show that imbuing artificial agents with features of human decision-making from behavioral economics gives rise to an expressive class of equilibria in games. Crucially, this new solution concept appears to overcome many of the computational limitations of existing concepts and thus appears to be a promising foundation for the development of principled MARL algorithms.

Notations: We use Δ_n to denote probability simplex of size n. In addition, we denote $[N] = \{1, 2, \cdots, N\}$ for any positive integer $N > 0$. We denote $x = \big[x(s,a)\big]_{(s,a)\in\mathcal{S}\times\mathcal{A}} \in \mathbb{R}^{SA}$ (resp. $x = \big[x(s)\big]_{s\in\mathcal{S}} \in \mathbb{R}^{S}$) as any vector that constitutes certain values for each state-action pair (resp. state).

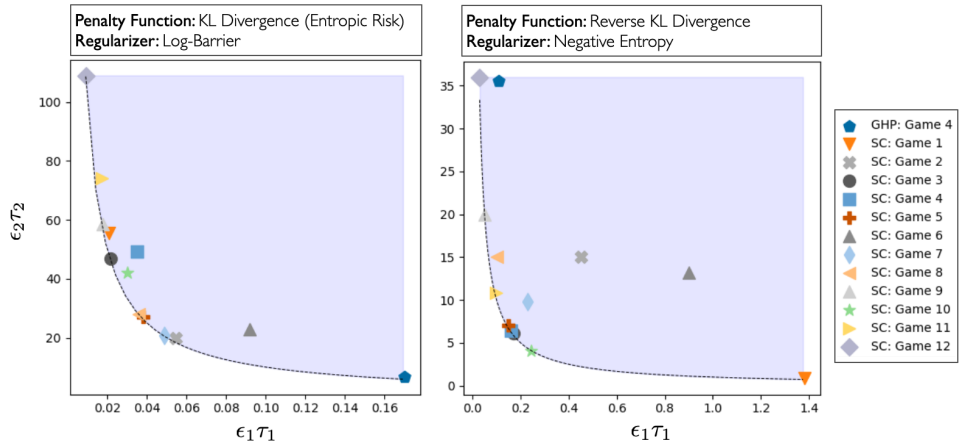

Figure 1: The shaded blue region depicts the regime of risk-aversion and bounded rationality preferences that allow for computationally tractable RQE in all 2-player games as shown in Theorem 3. The markers *GHP: Game 4* (Goeree et al., 2003) and *SC: Game 1-12* (Selten and Chmura, 2008) represent the necessary parameter values required to recreate the average strategy played by people in various 2-player games in observational data up to 1% accuracy.

2 Risk-aversion and Bounded Rationality in Games

To begin, we focus on n-player general-sum finite-action games. In these games, each player i has access to a (finite) action set $\mathcal{A}_i$ with $A_i = |\mathcal{A}_i|$ pure strategies. For each tuple of joint strategies $a = (a_1, ..., a_n) \in \mathcal{A} := \prod_{i=1}^n \mathcal{A}_i$ each player has an associated reward or utility $R_i(a)$. As is common in the study of these games, we consider the case where players play over mixed strategies and seek to maximize their expected utility. Player i's expected utility in this case can be written as a function $U_i : \mathcal{P} \to \mathbb{R}$ which can be written as:

$$U_i(\pi_1, ..., \pi_n) = \mathbb{E}_{a\sim\pi}[R_i(a)] \tag{1}$$

where $\pi = (\pi_1, ..., \pi_n) \in \mathcal{P} := \prod_{i=1}^n \Delta_{A_i}$ is the joint strategy of all the players. For ease of exposition we often write the utility as $U_i(\pi_i, \pi_{-i})$ where $\pi_{-i} = (\pi_1, ..., \pi_{i-1}, \pi_{i+1}, ..., \pi_n)$ represents the joint strategies of all players *other* than i. A natural outcome in this class of games is the notion of a (mixed) Nash equilibrium.

Definition 1 (Nash Equilibrium). *A (mixed) Nash equilibrium of a n-player general-sum finite-action game is a joint strategy $\pi^* = (\pi_1^*, ..., \pi_n^*) \in \mathcal{P}$ such that no player has any incentive to unilaterally deviate—i.e., for all $i = 1, ..., n$: $U_i(\pi_i^*, \pi_{-i}^*) \geq U_i(\pi_i, \pi_{-i}^*) \quad \forall \pi_i \in \Delta_{A_i}$.*

As introduced, n-player general-sum finite-action games are well known to admit at least one Nash equilibrium (NE) in mixed strategies (Nash, 1951). While NE are known to always the exist in mixed strategies in the games we consider, they have also been shown to be intractable to compute (Daskalakis, 2013). Moreover, NE may not realistically predict human decision-making behavior. Indeed a preponderance of empirical evidence suggests that people do *not* play their Nash strategies and in fact are potentially bounded rational and risk-averse in their decision-making—inducing new forms of equilibria (see e.g., Luce (1959); Tversky and Kahneman (1992); Goeree et al. (2003); McKelvey and Palfrey (1995)). To formally introduce these features of human decision-making into the problem, we construct generalizations of the expected utility game that allow us to model agents as either imperfect optimizers and risk-averse.

2.1 Bounded Rationality in Games

In contrast to risk-aversion, which has been under-explored in game theory and multi-agent reinforcement learning, bounded rationality is more common. Many works studying the computational benefits of incorporating it into games (Sokota et al., 2023; Mertikopoulos and Sandholm, 2016; Cen et al., 2021; Leonardos et al., 2021; Evans and Ganesh, 2024; Jacob et al., 2022), with the most common form of bounded rationality found in the literature on learning in games being that of a quantal response. These capture bounded rationality by either assuming that the players are rational in a stochastically perturbed version of the game or equivalently that players' strategies are constrained to the set of *quantal response functions* (McKelvey and Palfrey, 1995; 1998).

Definition 2. *(Quantal Response Function) A* quantal response function *is a continuous function* $\sigma : \mathbb{R}^n \to \Delta_n$ *such that for any* $x \in \mathbb{R}^n$*: If* $x_i < x_j$*,* $\sigma_i(x) > \sigma_j(x)$*, where* $x_k, \sigma_k(x)$ *represent the* k*-th components of* x *and* σ*, respectively.*

These functions restrict players' strategies to subsets of the simplex—effectively smoothing their best responses and preventing them from being complete utility maximizers. One way to restrict a player's strategy to quantal responses is by regularizing their objective with a suitable strongly convex regularizer ν_i (see, e.g., Föllmer and Schied (2002, Proposition 7), or Sokota et al. (2023); Mertikopoulos and Sandholm (2016)). This yields adjusted utility functions of the form:

$$U_i^{\epsilon_i}(\pi_1, ..., \pi_n) = \mathbb{E}_{a\sim\pi}[R_i(a)] - \epsilon_i \nu(\pi_i), \tag{2}$$

where $\epsilon_i \geq 0$ captures player i's degree of bounded rationality. As ϵ_i increases their strategy space becomes more constrained, yielding more boundedly rational strategies. When all players achieve a Nash equilibrium in the space of their quantal responses—the class of strategies that can be represented by a fixed class of quantal response functions (cf. Definition 2)—the resulting equilibrium is known as a quantal response equilibrium (QRE).

Despite the many works that focus on computing QRE, to the best of our knowledge there are no classes of QRE that are universally computable across all games. Indeed most works focus on zero-sum or approximately zero-sum games (Sokota et al., 2023; Mertikopoulos and Sandholm, 2016; Leonardos et al., 2021). In more general classes of games the class of QRE or equivalently the level of bounded rationality needed for computational tractability depends on the underlying game structure (e.g., the size of player's action spaces and the magnitude of their rewards) which may not be known a priori (Sun et al., 2024). Furthermore, we note that more work in behavioral economics has consistently highlighted that bounded rationality on its own is not enough to capture the nuances of human decision-making, even for simple games such as matching pennies (Goeree et al., 2003; Tversky and Kahneman, 1992). These findings motivate us to introduce risk-aversion into games.

2.2 Risk-aversion in Games

To allow agents to have risk preferences we make use of a general class of convex risk metrics from mathematical finance and operations research (Föllmer and Schied, 2002). In this framing we move into a regime where agents seek to *minimize* a measure of risk.

Definition 3 (Convex Risk Measures)**.** *Let* $\mathcal{X}$ *be the set of functions mapping from a space of outcomes* Ω *to* $\mathbb{R}$*. A convex measure of risk is a mapping* $\rho : \mathcal{X} \to \mathbb{R}$ *satisfying:*
1. Monotonicity: *If* $X \leq Y$ *almost surely, then* $\rho(X) \geq \rho(Y)$.
2. Translation Invariance: *If* $m \in \mathbb{R}$ *then* $\rho(X + m) = \rho(X) - m$.
3. Convexity: *For all* $\lambda \in (0, 1)$*,* $\rho(\lambda X + (1 - \lambda)Y) \leq \lambda\rho(X) + (1 - \lambda)\rho(Y)$.

Typically, and as we assume in the remainder of the paper, the set $\mathcal{X}$ is the set of measurable functions defined on a probability space $(\Omega, \mathcal{F}, P)$. Under this assumption, the convex measures of risk allow us to generalize expectations to trade off high variance for and higher utilities. For ease of exposition, we write $\rho_\pi(X)$ to highlight the distribution that the convex risk measure is taken with respect to π.

Given these definitions, we present two ways of incorporating risk-aversion into games.

Remark 1. A crucial feature of our formulation of risk-aversion is that players are *not* risk-averse to the randomness introduced by their own strategy and only to the uncertainty caused by the environment and their opponents. This is a common approach taken (often implicitly) in the literature on single-agent risk-sensitive and robust decision-making (Shen et al., 2014). It also appears to be necessary to ensure the existence of equilibria introduced in Theorem 2 shortly; otherwise as studied in Fiat and Papadimitriou (2010), equilibria may not exist.

To see why our framing of risk-aversion is natural we introduce the first form of risk that we consider: *action-dependent* risk-aversion.

Action-dependent Risk Aversion: In this framing, a player first evaluates the risk associated with each of their pure strategies a_i: $\rho_{i,a_{-i}\sim\pi_{-i}}(R(a_i, a_{-i}))$. The player then minimizes their expected risk. We capture this by transforming the player's utilities in (1) into costs f_i^{act} of the form:

$$f_i^{act}(\pi_i, \pi_{-i}) = \mathbb{E}_{\pi_i}\left[\rho_{i,\pi_{-i}}(R_i(a))\right] = \textstyle\sum_{a_i\in\mathcal{A}_i} \pi_i(a_i)\rho_{i,\pi_{-i}}\left(R_i(a_i, a_{-i})\right), \tag{3}$$

where $\rho_{i,\pi_{-i}}$ is used to capture agent i's risk preference which depends on the product distribution of opponents strategies π_{-i}.

Under action-dependent risk-aversion a player uses mixed strategies only when two pure strategies yield the same level of risk. Thus introducing additional risk-aversion to π_i does not change outcomes.

Aggregate Risk Aversion: The second form of risk-aversion we consider, *aggregate* risk-aversion is more conservative, but allows for a simpler analysis. To capture aggregate risk aversion, we transform the player's utilities in (1) into costs f_i^{agg} which take the form

$$f_i^{agg}(\pi_i, \pi_{-i}) = \rho_{i,\pi_{-i}}(\mathbb{E}_{\pi_i}[R_i(a)]) = \rho_{i,\pi_{-i}}\left(\sum_{a_i \in \mathcal{A}_i} \pi_i(a_i) R_i(a_i, a_{-i})\right). \tag{4}$$

Aggregate risk-aversion captures the fact that risks may be correlated across pure strategies, and is related (due to convexity of ρ_i) by Jensen's inequality to action-dependent risk-aversion such that $f_i^{agg}(\pi) < f_i^{act}(\pi)$. Interestingly, recent work in behavioral economics has shown that this may be a better model of risk-aversion in people (Oprea and Robalino, 2024).

We note that if $\rho_i(X) = \mathbb{E}[-X]$ for all players $i = 1, ..., n$ (which satisfies the requirements in Definition 3), then the new formulations reduce to the original expected utility objective. Thus, the class of risk-averse games can be seen as generalizations of the classic expected utility games (1).

Remark 2. For brevity, as both generalizations have similar implications, we provide details and results of action-dependent risk aversion in our supplementary material and focus on aggregate risk aversion for the remainder of the paper. Thus we denote $f_i = f_i^{agg}$ for the remainder of the paper and analyze $f_i = f_i^{act}$ in the supplementary material.

While at first glance the modified game looks significantly more complex than the previous expected utility maximization setup (cf. (1)), we can rely on a particularly powerful property of convex measures of risk to simplify and expose some structure in this class of problems.

Theorem 1 (Dual Representation Theorem for Convex Risk Measures (Föllmer and Schied, 2002))**.** *Suppose that the set $\mathcal{X}$ is the set of functions mapping from a finite set Ω to $\mathbb{R}$. Then a mapping $\rho : \mathcal{X} \to \mathbb{R}$ is a convex risk measure (cf. Definition 3) if and only if there exists a penalty function $D : \Delta_\Omega \to (-\infty, \infty]$ such that: $\rho(X) = \sup_{p \in \Delta_\Omega} E_p[-X] - D(p)$, where Δ_Ω is the set of all probability measures on Ω. Furthermore, the function $D(p)$ can be taken to be convex, lower-semi-continuous, and satisfy $D(p) > -\rho(0)$ for all $p \in \Delta_\Omega$.*

Throughout, we make a simplifying assumption that D is continuous in both arguments, which is satisfied by various widely used risk measures Table 1. We provide more details on penalty functions in Appendix B. Given this result, we derive the aggregate risk-averse game (4) in the following form:

$$f_i(\pi_i, \pi_{-i}) = \sup_{p_i \in \mathcal{P}_{-i}} -\pi_i^T R_i p_i - \frac{1}{\tau_i} D_i(p_i, \pi_{-i}) \;\; \forall i = 1, .., n, \tag{5}$$

where $\mathcal{P}_{-i} = \mathcal{P}/\Delta_{A_i} \subset \mathbb{R}^{A_{-i}}$, $A_{-i} = \prod_{j \neq i} A_j$, and $R_i \in \mathbb{R}^{A_i \times A_{-i}}$ is player i's payoff matrix.

We note that we differentiate the penalty functions D_i to allow agents to have different risk preferences in different risk metrics. The parameter τ_i captures a player's degree of risk-aversion. In this form, one can see that in a risk-averse game, the players imagine that intermediate adversaries seek to maximize their cost but are penalized from deviating too far from the opponents' realized strategies. As $\tau_i \to \infty$ they become increasingly risk-averse—and in the extreme, treat their opponents and environment as adversarial.

Since a NE of the risk-averse game can be qualitatively different from that of the original game, we now define a risk-averse Nash equilibrium (RNE). For brevity, we do not introduce aggregate and action-dependent RNE separately, since similar results hold for either formulation.

Definition 4. *(Risk-Averse Nash Equilibrium) A risk-averse Nash equilibrium (RNE) of is joint strategy $\pi^* \in \mathcal{P}$ such that no player has any incentive to unilaterally deviate in the risk adjusted game—i.e., for all $i = 1, ..., n$: $f_i(\pi_i^*, \pi_{-i}^*) \leq f_i(\pi_i, \pi_{-i}^*) \;\; \forall \;\; \pi_i \in \Delta_{A_i}$*

Note that since players would like to *minimize* risk, the direction of the inequality has changed. The convexity and continuity of the penalty function guarantees that the risk-averse games admit at least one RNE. A general statement and proof are provided in Theorem 7.

Theorem 2. *All aggregate risk-averse games (5) admit at least one RNE.*

The existence of a RNE is a consequence of the fact that under our risk formulations, players are not risk averse to the randomness of their own by playing a mixed strategy. This is in stark contrast to previous works which consider games in which players are risk-averse to all randomness (including their own), and in which a RNE may not exist (Fiat and Papadimitriou, 2010). Thus, our specific risk formulations are crucial to our later results.

Finally, we note that the additional convexity induced by the introduction of risk aversion guarantee is not enough to ensure the computational tractability of the risk-averse Nash equilibrium (see e.g., (McMahan et al., 2024)). This is not surprising given the fact that computing NE in general convex games can be intractable in general.

3 Risk-Averse Quantal Response Equilibria in Matrix Games

Since neither bounded rationality nor risk-aversion alone guarantees computational tractability of their associated solution concepts, we consider the equilibrium concept that arises out of their combination and show that it can often be efficiently computed. Interestingly, this echoes work in behavioral economics in which the combination of risk aversion and bounded rationality has also been found to be a better predictor of human play (Goeree et al., 2003) than either of the properties alone.

Towards this goal, in the final form of the risk-adjusted game we consider, players' costs are given by:

$$f_i^{\epsilon_i}(\pi_i, \pi_{-i}) = f_i(\pi_i, \pi_{-i}) + \epsilon_i \nu_i(\pi_i) \;\; \forall i = 1, .., n, \tag{6}$$

where ν_i is a strictly convex regularizer which gives rise to a set of quantal responses. The natural outcome of this game is what we term as a risk-averse quantal-response equilibrium (RQE).

Definition 5 (Risk-Averse Quantal Response Equilibrium (RQE))**.** *A risk-averse quantal response equilibrium of a n-player general-sum finite-action game is a joint strategy $\pi^* \in \mathcal{P}$ such that for each player $i = 1, ..., n$:*

$$f_i^{\epsilon_i}(\pi_i^*, \pi_{-i}^*) \leq f_i^{\epsilon_i}(\pi_i, \pi_{-i}^*) \;\; \forall \;\; \pi_i \in \Delta_{A_i}.$$

For any set of convex regularizers ν_i, it is easy to observe that the game remains a convex game and thus RQE (as defined) will always exist in all matrix games.

3.1 Conditions for the Computational Tractability of RQE

Given our definition of RQE, we now demonstrate that if players' risk preferences (i.e., their degree of risk aversion) and families of quantal response functions (i.e., their bounded rationality parameters) satisfy a simple relationship, the game admits a RQE that can be efficiently computed using arbitrary no-regret learning algorithms (e.g., gradient-play or mirror descent) in a decentralized manner. Importantly this result is *independent* of the underlying payoffs $\{R_i\}_{i=1}^n$.

To derive our results on the computational tractability of RQE under aggregate risk-aversion, we first introduce a related $2n$-player game in which we associate to each original player i an adversary whose strategy is denoted as $p_i \in \mathcal{P}_{-i}$. Let $p = (p_1, ..., p_n) \in \bar{\mathcal{P}} = \prod_{i=1}^n \mathcal{P}_{-i}$. In this new game, each original player's loss function takes the form:

$$J_i(\pi_i, \pi_{-i}, p) = -\pi_i^T R_i p_i - \frac{1}{\tau_i} D_i(p_i, \pi_{-i}) + \epsilon_i \nu_i(\pi_i). \tag{7}$$

For each player p_i we associate them to a new loss function which we denote:

$$\bar{J}_i(\pi, p_i, p_{-i}) = \pi_i^T R_i p_i + \frac{1}{\tau_i} D_i(p_i, \pi_{-i}) - \sum_{j \neq i} \xi_{i,j} \nu_j(\pi_j). \tag{8}$$

In this $2n$ player game, the i-th player's adversary, whose strategy is p_i seeks to minimize their loss $\bar{J}_i$ which is strategically the same as maximizing J_i.

Given the definition of the $2n$-player game we show how no-regret learning algorithms can be used to compute a RQE. To prove this, we first define coarse correlated equilibria (CCE).

Definition 6. *A coarse correlated equilibrium of the $2n$-player game is a probability measure σ on $\mathcal{P} \times \bar{\mathcal{P}}$ such that for all $i = 1, ..., n$:*

$$\mathbb{E}_{(\pi,p)\sim\sigma}[J_i(\pi, p)] \leq \mathbb{E}_{(\pi_{-i},p)\sim\sigma}[J_i(\pi_i', \pi_{-i}, p)] \quad \forall \pi_i' \in \Delta_{A_i}$$
$$\mathbb{E}_{(\pi,p)\sim\sigma}[\bar{J}_i(\pi, p)] \leq \mathbb{E}_{(\pi,p_{-i})\sim\sigma}[\bar{J}_i(\pi, p_i', p_{-i})] \quad \forall p_i' \in \mathcal{P}_{-i}.$$

CCE are the natural outcome of no-regret learning algorithms, and we show that CCEs of the $2n$-player game coincide with RQE of the original n-player matrix game. This is a phenomenon known as equilibrium collapse which is well known to happen in zero-sum games and certain generalizations of zero-sum games (Cai et al., 2016; Kalogiannis and Panageas, 2023). For brevity we present our results on 2-player matrix games now and extend the results to n-player setting in Appendix C.3.

To simplify our results we define $H_1(p_1, \pi_2) = \frac{1}{\tau_1} D_1(p_1, \pi_2) - \xi_1 \nu_2(\pi_2)$ and $H_2(p_2, \pi_1) = \frac{1}{\tau_2} D_2(p_2, \pi_1) - \xi_2 \nu_1(\pi_1)$. Let $\xi_1^* > 0$ and $\xi_2^* > 0$ be the smallest values of ξ_1 and ξ_2 such that $H_1(p_1, \pi_2)$ and $H_2(p_2, \pi_1)$ are concave in π_2 and π_1 for all p_1 and p_2 respectively. These parameters capture the player's relative degrees of risk aversion, and as we show, for certain instantiations of risk measures and quantal response functions $\xi_i^* = \frac{1}{\tau_i}$ for $i = 1, 2$. The following theorem gives conditions on ξ_1^*, ξ_2^* and ϵ_1, ϵ_2, under which RQEs can be computed through no-regret learning. The proof is postponed to Appendix C.2.

Theorem 3. *Assume the penalty functions that give rise to the players' risk preferences $D_1(\cdot, \cdot)$ and $D_2(\cdot, \cdot)$ are jointly convex in both their arguments. If σ is a CCE of the four player game with $\xi_{1,2} = \xi_1^*$ and $\xi_{2,1} = \xi_2^*$, and $\epsilon_1 \epsilon_2 \geq \xi_1^* \xi_2^*$, then $\hat{\pi}_1 = \mathbb{E}_\sigma[\pi_1]$ and $\hat{\pi}_2 = \mathbb{E}_\sigma[\pi_2]$ constitute an RQE of the risk-averse game.*

This theorem gives a range of risk aversion and bounded rationality parameters under which an RQE can be computationally tractable using no-regret learning. This range is independent of the structure in the underlying game—making the class of tractably computable equilibria universal to all games. In the case where $\xi_i^* = \frac{1}{\tau_i}$, we observe that for any $\epsilon_1, \epsilon_2 > 0$, as $\tau_i \to \infty$, the game becomes solvable. This captures the fact that for any degree of bounded rationality one can compute a players' security strategy using no-regret learning—recovering existing results on the computational tractability of min-max strategies.

To validate this result, we show that the class of computationally tractable RQE is sufficiently rich to capture the aggregate strategies played by people in matrix games studied in laboratory experiments in behavioral economics to showcase how human behavior can differ from Nash equilibrium strategies (Selfridge, 1989; Goeree et al., 2003; Selten and Chmura, 2008). In particular, we showcase in Fig. 1 that for some choice of parameters (ϵ_j and τ_j's) satisfying conditions in Theorem 3, we recover human behaviors in various matching pennies games investigated in Tables 2 and 3. We provide more details about the games and experiments in Appendix F.1.

4 Risk-Averse Quantal Response Equilibria in Markov Games

In this section, we extend our results to finite-horizon Markov games, or stochastic games (Shapley, 1953) which allow us to model dynamic games played out over Markov decision processes (MDP).

A Markov game can be seen as a sequential matrix game with stochastic dynamics. In this paper, we consider general-sum finite-horizon Markov games involving n players, represented as $\mathcal{MG} = \{H, \mathcal{S}, \{\mathcal{A}_i\}_{i\in[n]}, \{R_{i,h}, P_{i,h}\}_{i\in[n],h\in[H]}\}$. Here, H is the time horizon length of the Markov game and $\mathcal{S} = \{1, 2, \cdots, S\}$ represents the state space of the underlying MDP with size S. We also adopt the same notation as in Section 2, and let $\{\mathcal{A}_i\}_{i\in[n]}$ represent the action spaces of each player, each with cardinality $|\mathcal{A}_i| = A_i$. Similarly, the joint action space is given by $\mathcal{A}$, a joint action profile is $a \in \mathcal{A}$, and $\mathcal{P}$ is the product policy space. Similarly, $R_{i,h}$ represents the utility or reward function of the i-th player at time step h for any $(i, h) \in [n] \times [H]$ and $R_{i,h}(s, \boldsymbol{a})$ represents the immediate reward (or utility) received by the i-th player given state-action pair $(s, \boldsymbol{a})$. For simplicity, we assume $R_{i,h}$ is deterministic. Lastly, the dynamics are captured by transition kernels $P_{i,h} : \mathcal{S} \times \mathcal{A} \mapsto \Delta_S$ where $P_{i,h}(s' \mid s, \boldsymbol{a})$ which capture probability transitioning from current state s to the next state s' conditioned on the joint action $\boldsymbol{a}$.

Markov policies and value functions. In the classical setup of finite-horizon Markov games, players are assumed to play over the space of Markov policies, where at any state s at any time

step $h \in [H]$, the action selection rule depends only on the current state s, and is independent of previous trajectories and other players. Specifically, the i-th player or agent executes actions according to a policy $\pi_i = \{\pi_{i,h} : \mathcal{S} \mapsto \Delta_{A_i}\}_{1 \le h \le H}$, with the probability of selecting action a in state s at time step h given by $\pi_{i,h}(a \,|\, s)$. Since we are operating in a finite horizon regime, it is natural to assume the policies are time-dependent. For simplicity, we define the joint policy of all agents as a product policy defined as $\pi = (\pi_1, \dots, \pi_n) : \mathcal{S} \times [H] \mapsto \mathcal{P}$ and the joint policy space as Π. As such, the joint action profile $\boldsymbol{a}$ of all agents is drawn from distribution specified by $\pi_h(\cdot \,|\, s) = (\pi_{1,h}, \pi_{2,h} \dots, \pi_{n,h})(\cdot \,|\, s) \in \mathcal{P}$ conditioned on state s at time step h. For any given π, we employ $\pi_{-i} : \mathcal{S} \times [H] \mapsto \mathcal{P}_{-i}$ to represent the policies of all agents excluding the i-th agent.

4.1 Risk-Averse Markov games

Given these definitions, we now generalize Markov games by allowing agents to be risk-averse to both the uncertainties arising from other agents' strategies and from the stochastic dynamics. This results in a new formulation of risk-averse Markov games (RAMGs). To address the two sources of risk-aversion we allow agents to have different risk preferences for their opponents and for the environment. To do so we fix two penalty functions $\{D_{\mathsf{pol},i}(\cdot,\cdot)\}_{i\in[n]}$ and $\{D_{\mathsf{env},i}(\cdot,\cdot)\}_{i\in[n]}$, which in turn give rise to two risk metrics. This allows us to define the following two functions which separately capture the risk associated with other players' actions and the environment respectively:

$$\begin{aligned} \forall i \in [n]: \quad & f^{\pi}_{\mathsf{pol},i}(Q_i) = \sup_{p_i \in \mathcal{P}_i} -\pi_i^T Q_i p_i - D_{\mathsf{pol},i}(p_i, \pi_{-i}), \\ \forall i \in [n]: \quad & f_{\mathsf{env},i}(R, P, V) = \inf_{\widetilde{P} \in \Delta_S} \left[R + \widetilde{P} V + D_{\mathsf{env},i}(\widetilde{P}, P) \right], \end{aligned} \tag{9}$$

for any joint policy π, where $Q_i \in \mathbb{R}^{A_i \times \sum_{j \neq i} A_j}$, $R \in \mathbb{R}^A$, $V \in \mathbb{R}^S$ will be used to represent cumulative payoff matrices and $P \in \Delta_S$, represents the dynamics of the MDP. Note that for simplicity, we choose the same $\{D_{\mathsf{pol},i}\}_{i\in[n]}$ and $\{D_{\mathsf{env},i}\}_{i\in[n]}$ for all time steps $h \in [H]$, though this can be considered to be time dependent as well.

Given the definition of these two penalty functions $f^{\pi}_{\mathsf{pol},i}(\cdot)$ and $f_{\mathsf{env},i}(\cdot)$, we can now define an agent's cumulative risk in a RAMG. For a risk-averse agent i their goal is now to minimize their own long-term risk, which can be captured by a risk-averse value function $\{V_{i,h}(\pi)\}_{h\in[H]}$, where $V_{i,h}(\pi) : \mathcal{S} \mapsto \mathbb{R}$ for any joint strategy $\pi \in \Pi$. This can be defined recursively as follows: for a given joint policy π, for all $(h, s) \in [H] \times \mathcal{S}$,

$$V_{i,h}(\pi; s) = f^{\pi}_{\mathsf{pol},i}\left(Q_{i,h}(\pi; s, :)\right) \tag{10}$$

where $Q_{i,h}(\pi; s, \boldsymbol{a}) = f_{\mathsf{env},i}\left(R_{i,h}(s, \boldsymbol{a}), P_{h,s,\boldsymbol{a}}, V_{i,h+1}(\pi)\right)$ and $P_{h,s,\boldsymbol{a}} \coloneqq P_h(\cdot \,|\, s, \boldsymbol{a}) \in \mathbb{R}^{1\times S}$.

To parse this definition, we observe that $Q_{i,h}(\pi; s, :) \in \mathbb{R}^{A_i \times \sum_{j\neq i} A_j}$ represents a payoff matrix for the i-th player at state s and time step h, where the value at the a_i-th row and the a_{-i}-th column is specified by $Q_{i,h}(\pi; s, \boldsymbol{a})$ with $\boldsymbol{a} = (a_i, a_{-i})$ which captures an agents' cumulative risk if joint action $\boldsymbol{a}$ is executed at state s and time h, and the policy π is followed subsequently. This can also be captured concisely through a recursive definition:

$$V_{i,h}(\pi; s) = f^{\pi}_{\mathsf{pol},i} \circ f_{\mathsf{env},i}\big(R_{i,h}(s, :), P_{h,s,:}, \big[f^{\pi}_{\mathsf{pol},i} \circ f_{\mathsf{env},i}\big(R_{i,h+1}, P_{h+1}, \cdots\big)\big]\big).$$

We remark that this definition reduces to the classical setup of multi-agent reinforcement learning (Zhang et al., 2021a) when agents are risk-neutral and to the well studied setup of risk-sensitive reinforcement learning (Shen et al., 2014) when there is only one agent.

As before, we focus on the computation of risk-averse quantal response equilibria in Markov games. To do, we constrain players to quantal responses by regularizing their value functions as follows:

$$\forall (i, h, s) \in [n] \times [H] \times \mathcal{S}: \quad V^{\epsilon_i}_{i,h}(\pi; s) = V_{i,h}(\pi; s) + \epsilon_i \nu_i(\pi_i), \tag{11}$$

where $\{\epsilon_i \geq 0\}_{i\in[n]}$ capture players' degrees of bounded rationality. To define Markov RQE we denote $\pi_{i,-h} \coloneqq \{\pi_{i,t}\}_{t=1,2,3,\cdots,h-1,h+1,\cdots,H}$ as agent i's policy over all times other than h.

Definition 7 (Markov RQE)**.** *A product policy $\pi = \pi_1 \times \cdots \times \pi_n \in \Pi$ is said to be a risk-averse quantal response equilibrium of RAMG if:*

$$\forall (i, s, h) \in [n] \times \mathcal{S} \times [H]: \quad V^{\epsilon_i}_{i,h}(\pi; s) \leq \min_{\pi'_h : \mathcal{S} \mapsto \Delta_{\mathcal{A}_i}} V^{\epsilon_i}_{i,h}\big((\pi'_h, \pi_{i,-h}) \times \pi_{-i}; s\big), \tag{12}$$

where $\nu_i(\cdot)$ is a strictly convex function over the simplex.

4.2 COMPUTING AND APPROXIMATING MARKOV RQE

Given our setup we now investigate how to compute Markov RQE, both with full information over the dynamics and rewards and without. For the full-information regime we propose a modified form of dynamic programming, in which we work backwards from time $h = H$ and then recursively compute the RQE until step $h = 0$. Towards this goal, at each time step h, for all $(i, h, s, \boldsymbol{a}) \in [n] \times [H] \times \mathcal{S} \times \mathcal{A}$, we construct the payoff matrices for the underlying for all agents at any state s at time step h as

$$\forall (i, s, h) \in [n] \times \mathcal{S} \times [H] : \ \widehat{Q}_{i,h}(s, \boldsymbol{a}) = R_{i,h}(s, \boldsymbol{a}) + \inf_{\widetilde{P} \in \Delta_S} \widetilde{P}\widehat{V}_{i,h+1} + D_{\mathsf{env},i}(\widetilde{P}, P_{h,s,\boldsymbol{a}}). \quad (13)$$

These payoff matrices capture both the future payoffs as well as the risk associated with the stochastic transitions in the MDP associated with a joint action $\boldsymbol{a}$. Together, they define a matrix game at state s and time h, for which we can compute a RQE using the results presented in Section 2. We denote the routine for computing a RQE of a matrix game as $\mathsf{RQE}(\cdot)$. Proceeding recursively, we compute the RQE at time $t + 1$ and use the resulting policies to define the underlying payoff matrices at time t. We summarize the algorithm for computing Markov RQE in Algorithm 1 in the appendix. The following theorem guarantees that the procedure will output a Markov RQE.

Theorem 4. *For any RAMG $\mathcal{MG}$, assume the penalty functions that give rise to the players' risk preferences $\{D_{\mathsf{pol},i}(\cdot, \cdot)\}$ are jointly convex in both of their arguments. If, for all $(s, h) \in \mathcal{S} \times [H]$, $\epsilon_i \geq \sum_{j \neq i} \xi^*_{j,i}$ for all $(i, j) \in [n] \times [n]$, where $\xi^*_{i,j}$ is defined as in Theorem 9 for given $D_{\mathsf{pol},i}(\cdot, \cdot)$ and ν_j, then the output policy $\widehat{\pi}$ from Algorithm 1 is a Markov RQE of $\mathcal{MG}$.*

A consequence of this result is that a class of Markov RQE are computationally tractable to compute via no-regret learning in finite-horizon Markov games. We remark that this is in stark contrast to both Markov Nash equilibria and Markov quantal response equilibria which cannot efficiently be computed in general-sum Markov games (Daskalakis et al., 2023b).

Computing Markov RQE in unknown Markov Games. The previous results focused on the case when the dynamics and rewards of the RAMG are known. In practice, especially in multi-agent reinforcement learning, the environments can be complex and unknown and must be *learned* by interacting with (i.e., sampling from) the environments. In this subsection, we focus on such scenarios and provide the theoretical guarantees for Algorithm 1.

We focus on a fundamental problem setup which assumes access to a generative model or a simulator (Kearns and Singh, 1998; Agarwal et al., 2020). This allows us to collect N independent samples for each state-action pair generated based on the true environment $\{R_{i,h}, P_h\}_{i \in [n], h \in [H]}$:

$$\forall (s, \boldsymbol{a}, j) \in \mathcal{S} \times \mathcal{A} \times [n] : \quad s_{i,s,\boldsymbol{a},h} \overset{i.i.d}{\sim} P_h(\cdot \mid s, \boldsymbol{a}), \quad r^i_{j,s,\boldsymbol{a},h} = R_{j,h}(s, \boldsymbol{a}), \quad i \in [N]. \quad (14)$$

The total sample size is, therefore, $N_{\mathsf{all}} \coloneqq NS \prod_{i \in [n]} A_i$.

For problems we propose a model-based approach to the computation of Markov RQE which first constructs an empirical reward function and nominal transition kernel based on the collected samples and then applies Algorithm 1 to learn a Markov RQE. First, the empirical reward function and transition kernel $\widehat{P} \in \mathbb{R}^{S \prod_{i=1}^n A_i \times S}$ are constructed from the empirical frequency of state transitions,

$$\forall (s, \boldsymbol{a}, j, h) : \widehat{R}_{j,h}(s, \boldsymbol{a}) = r^1_{j,s,\boldsymbol{a},h} \quad \text{and} \quad \widehat{P}_h(s' \mid s, \boldsymbol{a}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{s_{i,s,\boldsymbol{a}} = s'\}. \quad (15)$$

Then with such empirical reward and transition, we can apply the oracle summarized in Algorithm 1 to compute Markov RQE using model-based MARL. The following theorem provides the first finite-sample guarantees for the computation or Markov RQE.

Theorem 5. *For any RAMG $\mathcal{MG}$, we consider penalty functions $\{D_{\mathsf{env},i}(\cdot, \cdot)\}_{i \in [n]}$ are L-Lipschitz w.r.t the ℓ_1 norm of the second argument with any fixed first argument. Applying Algorithm 1 with the estimated reward $\{\widehat{R}_{j,h}\}$ and transition kernels $\{\widehat{P}_h\}_{h \in [H]}$ as input, the output solution $\widehat{\pi}$ is an δ-RQE of $\mathcal{MG}$. Namely, we have* $\max_{(i,s,h) \in [n] \times \mathcal{S} \times [H]} \Big\{ V^{\epsilon_i}_{i,h}(\widehat{\pi}; s) - \min_{\pi'_h : \mathcal{S} \mapsto \Delta_{\mathcal{A}_i}} V^{\epsilon_i}_{i,h}((\pi'_h, \widehat{\pi}_{i,-h}) \times$

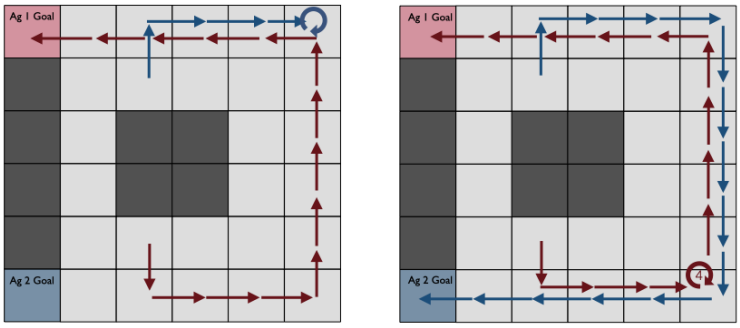

Figure 2: *Cliff Walk Description and Results*: The *Cliff Walk* grid-world is depicted here with color codes of grids: black is the cliff, blue and pink are agents/players 1 and 2's goals respectively. The agents' proximity to one another increases the randomness in the dynamics making falling into one of the cliff states more likely. Agent 2's policy in the left figure showcases more risk-aversion and less bounded rationality than on the right.

$\widehat{\pi}_{-i}; s)\Big\} \leq \delta$ *as long as the total number of samples satisfies*

$$N_{\mathsf{all}} = NS \prod_{i\in[n]} A_i \geq 8S \prod_{i\in[n]} A_i HL \sqrt{\frac{S}{N} \log\left(2SH \prod_{i\in[n]} A_i/\delta\right)}.$$

The proof is postponed to Appendix E. We remark that L can be some constant for various penalty function, such as $L = 1$ when $\{D_{\mathsf{env},i}\}_{i\in[n]}$ are defined as any ℓ_p norm including total variation (TV). We also note that our result suffers from what is known as the curse of mutliagency (Bai et al., 2020) through the dependence on $\prod_{i\in[n]} A_i$.

4.3 Experiments and Evaluation

We consider a grid-world problem to test our algorithm (Algorithm 1) and showcase the effects of risk-aversion and bounded rationality in games.

Cliff Walk Environment description: A grid consists of some tiles representing a cliff where they will remain stuck for all time and goal states for agents as well as goal states of agents. The cliff is the black grid with rewards -2. Agents/players are rewarded 0 for taking each step and 1 for reaching their respective goals. Agents actions are {up,down,left,right} and they are followed with probability $p_{\mathrm{d}} = 0.9$ with random movements happening otherwise. To introduce multi-agent effects we reduce p_{d} to 0.5 when the agents are at least a grid cell apart— making the likelihood of falling into the cliff higher. The episode horizon $H = 200$ and the joint state space is the tuple of players' positions.

Results Discussion: We present two results in Fig. 2. For both results, agent 1 starts at the 5th row and 3rd column of the cliff-walking grid-world, and agent 2 starts at the 2nd row and 3rd column grid position. In both figures, the red and blue paths depict the maximum likelihood paths taken by agents 1 and 2 respectively. Agent 2's policy in the left figure showcases more risk-aversion and less bounded rationality resulting in them preferring to hide far from obstacles than run the risk of falling off the path. On the right, agent 2 successfully reaches their goal. The equilibrium strategy of agent 1 changes in both scenarios: a more risk-seeking agent 2 forces agent 1—in an effort to minimize risk—to wait until the path is clear to attempt the journey to its goal.

5 Conclusion

By incorporating risk aversion and bounded rationality into agents' decision-making processes, we introduced a new class of equilibria for matrix games and finite-horizon Markov games: RQE. RQE were shown to align well with observed human behavior and we provided theoretical results showing that classes of RQE are tractably computable in all finite horizon Markov games. We also provided sample complexity results in the generative modeling setting for multi-agent reinforcement learning. Altogether, our results open the doors to the principled development of new MARL algorithms.

ACKNOWLEDGMENT

The work of LS is supported in part by the Resnick Institute and Computing, Data, and Society Postdoctoral Fellowship at California Institute of Technology. KP acknowledges support from the 'PIMCO Postdoctoral Fellow in Data Science' fellowship at the California Institute of Technology. EM acknowledges support from NSF Award 2240110.

REFERENCES

A. Agarwal, S. Kakade, and L. F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020. 9

A. Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155:1105–1123, 2012. 19, 20

P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008. 28

R. J. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1):67–96, 1974. 16

R. J. Aumann. Correlated equilibrium as an expression of bayesian rationality. *Econometrica*, 55(1): 1–18, 1987. 16

Y. Bai, C. Jin, and T. Yu. Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems*, 33:2159–2170, 2020. 10

T. Basar. Nash equilibria of risk-sensitive nonlinear stochastic differential games. *Journal of Optimization Theory and Applications*, 100(3):479–498, 1999. 17, 18

A. Beck. *First-order methods in optimization*. SIAM, 2017. 30

J. Blanchet, M. Lu, T. Zhang, and H. Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *Advances in Neural Information Processing Systems*, 36, 2024. 18

V. Borkar. Risk-sensitive control, single controller games and linear programming. *Journal of Dynamics and Games*, 11, 01 2023. 17, 19

J. N. Brown and R. Rosenthal. Testing the minimax hypothesis: A re-examination of o'neill's game experiment. *Econometrica*, 58(5):1065–81, 1990. 17

Y. Cai, O. Candogan, C. Daskalakis, and C. Papadimitriou. Zero-sum polymatrix games: A generalization of minmax. *Mathematics of Operations Research*, 41(2):648–655, 2016. 7, 16

C. M. Capra, J. K. Goeree, R. Gomez, and C. A. Holt. Learning and noisy equilibrium behavior in an experimental study of imperfect price competition. *International Economic Review*, 43(3): 613–636, 2002. 2

S. Cen, Y. Wei, and Y. Chi. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34:27952–27964, 2021. 3, 17

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006. ISBN 978-0-511-54692-1. 1, 2, 16

C. Daskalakis. On the complexity of approximating a nash equilibrium. *ACM Transactions on Algorithms (TALG)*, 9(3):1–35, 2013. 1, 3, 16

C. Daskalakis, N. Golowich, N. Haghtalab, and A. Shetty. Smooth nash equilibria: Algorithms and complexity. *arXiv preprint arXiv:2309.12226*, 2023a. 16

C. Daskalakis, N. Golowich, and K. Zhang. The complexity of markov equilibrium in stochastic games. In G. Neu and L. Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4180–4234. PMLR, 12–15 Jul 2023b. 2, 9, 16

E. Dekel and D. Fudenberg. Rational behavior with payoff uncertainty. *Journal of Economic Theory*, 52(2):243–267, 1990. 2

I. Erev and A. Roth. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, 88(4):848–81, 1998. 2, 17

H. Eriksson, D. Basu, M. Alibeigi, and C. Dimitrakakis. Risk-sensitive bayesian games for multi-agent reinforcement learning under policy uncertainty. *arXiv preprint arXiv:2203.10045*, 2022. 17

B. P. Evans and S. Ganesh. Learning and calibrating heterogeneous bounded rational market behaviour with multi-agent reinforcement learning. *arXiv preprint arXiv:2402.00787*, 2024. 3, 17

E. Even-Dar, Y. Mansour, and U. Nadav. On the convergence of regret minimization dynamics in concave games. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 523–532, 2009. 16

A. Fiat and C. Papadimitriou. When the players are not expectation maximizers. In *Algorithmic Game Theory: Third International Symposium, SAGT 2010, Athens, Greece, October 18-20, 2010. Proceedings 3*, pages 1–14. Springer, 2010. 4, 6, 16, 18

H. Föllmer and A. Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4):429–447, 2002. 4, 5, 19, 20

S. Ganesh, N. Vadori, M. Xu, H. Zheng, P. Reddy, and M. Veloso. Reinforcement learning for market making in a multi-agent dealer market. *Advances in Neural Information Processing Systems*, 2019. 17

Y. Gao, K. Y. C. Lui, and P. Hernandez-Leal. Robust risk-sensitive reinforcement learning agents for trading markets. *arXiv preprint arXiv:2107.08083*, 2021. 17

I. M. Gemp and S. Mahadevan. Online monotone games. *ArXiv*, abs/1710.07328, 2017. 23

J. K. Goeree and C. A. Holt. Stochastic game theory: For playing games, not just for doing theory. *Proceedings of the National Academy of Sciences*, 96(19):10564–10567, 1999. 2

J. K. Goeree, C. A. Holt, and T. R. Palfrey. Risk averse behavior in generalized matching pennies games. *Games and Economic Behavior*, 45(1):97–113, 2003. 2, 3, 4, 6, 7, 17, 30

K. Goeree and T. Offerman. Efficiency in auctions with private and common values: An experimental study. *American Economic Review*, 92(3):625–643, June 2002. 2, 17

C. Gollier. *The economics of risk and time*. MIT press, 2001. 2

N. Golowich, S. Pattathil, and C. Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. *Advances in neural information processing systems*, 33:20766–20778, 2020. 16

S. He, S. Han, S. Su, S. Han, S. Zou, and F. Miao. Robust multi-agent reinforcement learning with state uncertainty. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. 18

T. Ho, C. Camerer, and J.-K. Chong. A cognitive hierarchy model games. *The Quarterly Journal of Economics*, 119:861–898, 02 2004. 17

G. N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005. 17

A. P. Jacob, D. J. Wu, G. Farina, A. Lerer, H. Hu, A. Bakhtin, J. Andreas, and N. Brown. Modeling strong and human-like gameplay with kl-regularized search. In *International Conference on Machine Learning*, pages 9695–9728. PMLR, 2022. 3, 17

D. Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Transactions on Automatic Control*, 18(2): 124–131, 1973. 17

D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. 2

F. Kalogiannis and I. Panageas. Zero-sum polymatrix markov games: Equilibrium collapse and efficient computation of nash equilibria, 2023. 7, 16

S. Kannan, M. Kearns, J. Morgenstern, M. Pai, A. Roth, R. Vohra, and Z. S. Wu. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 369–386, 2017. 1

M. Kearns and S. Singh. Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in neural information processing systems*, 11, 1998. 9

S. Leonardos, G. Piliouras, and K. Spendlove. Exploration-exploitation in multi-agent competition: Convergence with bounded rationality. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26318–26331. Curran Associates, Inc., 2021. 3, 4, 17

R. D. Luce. *Individual Choice Behavior: A Theoretical analysis*. Wiley, New York, NY, USA, 1959. 2, 3, 17

E. Mazumdar, L. J. Ratliff, and S. S. Sastry. On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131, 2020. 1

R. D. McKelvey and T. R. Palfrey. An experimental study of the centipede game. *Econometrica*, 60 (4):803–836, 1992. 17

R. D. McKelvey and T. R. Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995. 2, 3, 17

R. D. McKelvey and T. R. Palfrey. Quantal response equilibria for extensive form games. *Experimental economics*, 1:9–41, 1998. 3, 17

J. McMahan, G. Artiglio, and Q. Xie. Roping in uncertainty: Robustness and regularization in markov games. In *Forty-first International Conference on Machine Learning*, 2024. 6, 18, 20

P. Mertikopoulos and W. H. Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016. 3, 4, 17, 20

P. Mertikopoulos, C. Papadimitriou, and G. Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, page 2703–2717, USA, 2018. Society for Industrial and Applied Mathematics. 1

A.-H. Mohsenian-Rad, V. W. Wong, J. Jatskevich, R. Schober, and A. Leon-Garcia. Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid. *IEEE transactions on Smart Grid*, 1(3):320–331, 2010. 1

J. Moon, T. E. Duncan, and T. Basar. Risk-sensitive zero-sum differential games. *IEEE Transactions on Automatic Control*, 64(4):1503–1518, 2019. 17, 18

H. Moulin and J. P. Vial. Strategically zero-sum games: The class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory*, 7(3):201–221, 1978. 16

R. Munos, M. Valko, D. Calandriello, M. G. Azar, M. Rowland, Z. D. Guo, Y. Tang, M. Geist, T. Mesnard, C. Fiegel, A. Michi, M. Selvi, S. Girgin, N. Momchev, O. Bachem, D. J. Mankowitz, D. Precup, and B. Piot. Nash learning from human feedback. In *Forty-first International Conference on Machine Learning*, 2024. 1

J. Nash. Non-cooperative games. *ANNALS OF MATHEMATICS*, 54(2), 1951. 3

J. F. Nash. Non-cooperative games. *Princeton University*, 1950. 1

B. O'Neill. Nonmetric test of the minimax theory of two-person zerosum games. *Proceedings of the National Academy of Sciences of the United States of America*, 84(7):2106–2109, 1987. 17

R. Oprea and N. Robalino. Are humans more averse to aggregate risk? testing an evolutionary economic theory. 2024. 5

K. Panaganti and D. Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 9582–9602, 2022. 17, 28

K. Panaganti, Z. Xu, D. Kalathil, and M. Ghavamzadeh. Robust reinforcement learning using offline data. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 17

W. Qiu, X. Wang, R. Yu, R. Wang, X. He, B. An, S. Obraztsova, and Z. Rabinovich. Rmix: Learning risk-sensitive policies for cooperative reinforcement learning agents. *Advances in Neural Information Processing Systems*, 34:23049–23062, 2021. 17

J. B. Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica*, 33(3):520–534, 1965. 20, 21

T. Roughgarden. Intrinsic robustness of the price of anarchy. *Journal of the ACM (JACM)*, 62(5): 1–42, 2015. 16

O. G. Selfridge. Adaptive strategies of learning a study of two-person zero-sum competition. In *Proceedings of the sixth international workshop on Machine learning*, pages 412–415. Elsevier, 1989. 7

R. Selten and T. Chmura. Stationary concepts for experimental 2x2-games. *American Economic Review*, 98(3):938–966, 2008. 3, 7, 29, 30

L. S. Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953. 7

S. Shen, C. Ma, C. Li, W. Liu, Y. Fu, S. Mei, X. Liu, and C. Wang. Riskq: risk-sensitive multi-agent reinforcement learning value factorization. *Advances in Neural Information Processing Systems*, 36:34791–34825, 2023. 17

Y. Shen, W. Stannat, and K. Obermayer. Risk-sensitive markov control processes. *SIAM Journal on Control and Optimization*, 51(5):3652–3672, 2013. 17

Y. Shen, M. J. Tobia, T. Sommer, and K. Obermayer. Risk-sensitive reinforcement learning. *Neural computation*, 26(7):1298–1328, 2014. 4, 8, 17, 18, 19

L. Shi, G. Li, Y. Wei, Y. Chen, M. Geist, and Y. Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 2023. 28

L. Shi, E. Mazumdar, Y. Chi, and A. Wierman. Sample-efficient robust multi-agent reinforcement learning in the face of environmental uncertainty. *arXiv preprint arXiv:2404.18909*, 2024. 18

O. Slumbers, D. H. Mguni, S. B. Blumberg, S. M. Mcaleer, Y. Yang, and J. Wang. A game-theoretic framework for managing risk in multi-agent systems. In *International Conference on Machine Learning*, pages 32059–32087. PMLR, 2023. 17

S. Sokota, R. D'Orazio, J. Z. Kolter, N. Loizou, M. Lanctot, I. Mitliagkas, N. Brown, and C. Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 4, 17, 20

Y. Sun, T. Liu, P. R. Kumar, and S. Shahrampour. Linear convergence of independent natural policy gradient in games with entropy regularization. *IEEE Control Systems Letters*, 8:1217–1222, 2024. 4

A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5:297–323, 1992. 2, 3, 4

M. Verma, S. Bhambri, and S. Kambhampati. On the brittle foundations of react prompting for agentic large language models. *arXiv preprint arXiv:2405.13966*, 2024. 1

Y. Viossat and A. Zapechelnyuk. No-regret dynamics and fictitious play. *Journal of Economic Theory*, 148(2):825–842, 2013. 2, 16

Z. Wang, Y. Shen, M. M. Zavlanos, and K. H. Johansson. Learning of nash equilibria in risk-averse games. *arXiv preprint arXiv:2403.10399*, 2024. 17

M. P. Wellman and P. R. Wurman. Market-aware agents for a multiagent world. *Robotics and Autonomous Systems*, 24(3-4):115–125, 1998. 1

Z. Xu*, K. Panaganti*, and D. Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Conference on Artificial Intelligence and Statistics, 2023. 17, 28

A. Yekkehkhany, T. Murray, and R. Nagi. Risk-averse equilibrium for games. *arXiv preprint arXiv:2002.08414*, 2020. 17

K. Zhang, T. Sun, Y. Tao, S. Genc, S. Mallya, and T. Basar. Robust multi-agent reinforcement learning with model uncertainty. *Advances in neural information processing systems*, 33:10571–10583, 2020. 18

K. Zhang, Z. Yang, and T. Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021a. 1, 8

R. Zhang, Y. Hu, and N. Li. Soft robust MDPs and risk-sensitive MDPs: Equivalence, policy gradient, and sample complexity. In *The Twelfth International Conference on Learning Representations*, 2024. 17

S. Zhang, B. Liu, and S. Whiteson. Mean-variance policy iteration for risk-averse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10905–10913, 2021b. 17

Algorithm 1: Computation method of RQE for risk-averse Markov games (RAMGs).

Input: reward function $\{R_{i,h}\}_{i\in[n]\times h\in[H]}$, transition kernel $\{P_h\}_{h\in[H]}$.
Initialization: $\widehat{Q}_{i,h}(s,a)=0$, $\widehat{V}_{i,h}(s)=0$ for all $(s,\boldsymbol{a},h)\in\mathcal{S}\times\mathcal{A}\times[H+1]$.
for $h=H,H-1,\cdots,1$ **do**
  for $i=1,2,\cdots,n$ and $s\in\mathcal{S},\boldsymbol{a}\in\mathcal{A}$ **do**
    Set $\widehat{Q}_{i,h}(s,\boldsymbol{a})$ according to (13).
  end for
  for $s\in\mathcal{S}$ **do**
    Get $\pi_h(s)=\{\pi_{i,h}(s)\}_{1\le i\le n}\leftarrow \mathsf{RQE}\Big(\big\{f_{\mathsf{pol},i}^{\pi,\epsilon_i}(\widehat{Q}_{i,h}(s,:)\big\}_{1\le i\le n}\Big)$.
    Set $\widehat{V}_{i,h}(s)=f_{\mathsf{pol},i}^{\pi_h,\epsilon_i}\big(\widehat{Q}_{i,h}(s,:)\big)$.
  end for
end for **Output:** $\widehat{\pi}=\{\pi_h\}_{1\le h\le H}$.

A Related Works

We put our work into context, and discuss related works here.

Computational tractability of game theoretic solution concepts. This work proposes a new solution concept for game theoretic settings that is computationally tractable, yet retains many of the desirable properties of classical equilibrium concepts. This general question emerged from the finding that computing a Nash equilibrium—perhaps the most natural solution concept for a game between rational self-interested agents—is PPAD-hard (Daskalakis, 2013), even for two-player general-sum matrix games. Despite this negative result, a large amount of subsequent work has focused on understanding the classes of games in which one can compute, approximate, or learn Nash equilibria efficiently. This is often done by assuming additional structure on the players' utilities and their relationships to one another, with large classes of games being zero-sum or competitive games, zero-sum polymatrix games (Cai et al., 2016; Kalogiannis and Panageas, 2023), monotone games (Golowich et al., 2020), smooth games (Roughgarden, 2015), or socially concave games (Even-Dar et al., 2009).

In games without such structure however, the natural targets for computation and learning became correlated (Moulin and Vial, 1978) and coarse correlated equilibria (Aumann, 1974; 1987)(CE and CCE respectively), both of which can be shown to emerge as the endpoint of no-regret learning and are thus considered to be computationally tractable targets for the design of learning algorithms. Despite this desirable property, the two concepts have significant drawbacks. Indeed both CE and CCE require some form of coordination between players to implement, introduce a highly nontrivial equilibrium selection problem (Cesa-Bianchi and Lugosi, 2006), and may have support on dominated strategies (Viossat and Zapechelnyuk, 2013). Furthermore, in the dynamic game context of Markov games, stationary CE and CCE are also computationally intractable to compute (Daskalakis et al., 2023b).

More recently, a new equilibrium concept—a smoothed Nash equilibrium— has been proposed as an alternative to these other equilibrium concepts (Daskalakis et al., 2023a) and motivated by similar considerations of individual and independent rationalizability and computational tractability. By applying ideas from smoothed analysis to the problem of computing Nash equilibria the authors show that one can efficiently find approximate classes of smoothed Nash equilibria—though to the best of our knowledge this cannot be done in a decentralized way.

Our approach is orthogonal and is rooted in giving MARL agents a foundation rooted in behavioral economics by imbuing them with a realistic feature of human decision-making: risk-aversion. The question of computational tractability of risk-averse Nash equilibria has been analyzed in (Fiat and Papadimitriou, 2010). The work shows that if agents are risk-averse with respect to all the randomness in the game (including their own) then a risk-averse Nash equilibrium may not even exist in mixed strategies, and even understanding if such equilibria exist can be NP-complete. Our formulation overcomes this by incorporating risk-aversion in a different way. Indeed, we show that when agents are risk-averse *only to the randomness introduced to their opponents (and the environment)* then the

risk-averse Nash equilibria will *always* exist. We note that such formulations of risk-aversion are common in the literature on risk-sensitive control (Shen et al., 2013; Borkar, 2023) and risk-sensitive reinforcement learning (Shen et al., 2014) where agents are implicitly presumed to be risk-averse only to the randomness that is outside their control (i.e., the environment). Furthermore we show that introducing bounded rationality into the game allows a class of risk-averse quantal response equilibria (RQE) to be computationally tractable in *all* finite action and finite-horizon Markov games.

Predictive power of equilibrium concepts. Another driving force in moving beyond the Nash and correlated equilibrium concepts stems from their lack of predictive power in experimental settings (see e.g., (Brown and Rosenthal, 1990; O'Neill, 1987; McKelvey and Palfrey, 1992; Erev and Roth, 1998; McKelvey and Palfrey, 1995)). To address this, a line of work originating in economics seeks to understand the natural solution concepts in game where players have behaviorally plausible restrictions to their strategy spaces, and to study whether such equilibria were better predictors of human play than Nash or (coarse) correlated equilibria (Goeree et al., 2003; Goeree and Offerman, 2002; Ho et al., 2004). The most common restriction is that players have *bounded rationality*,—i.e., they may fail to perfectly optimize—a model with roots in mathematical psychology (Luce, 1959). Under this restriction, a natural equilibrium concept that emerged was that of a quantal response equilibrium (QRE) which induces bounded rationality by either assuming that the players are rational in a stochastically perturbed version of the game or equivalently that they optimize a regularized version of their utility (McKelvey and Palfrey, 1995; 1998; Mertikopoulos and Sandholm, 2016). Beyond their use as a better model for human decision-making in games, QRE have also increasingly been adopted as a solution concept in multi-agent reinforcement learning and learning in games (Sokota et al., 2023; Mertikopoulos and Sandholm, 2016; Cen et al., 2021; Leonardos et al., 2021; Evans and Ganesh, 2024; Jacob et al., 2022) due to their links with KL and entropy regularized reinforcement learning. Despite these developments QRE are not computable in all games. Indeed the class of QRE or equivalently the level of bounded rationality needed for computational tractability depends on the underlying game structure which may not be known a priori. In contrast we show that the addition of risk aversion allows for the *same* class of quantal response equilibria to be computationally tractable to compute in all finite action games and finite-horizon Markov games. Furthermore we show that this class of risk-averse QRE is nontrivial and can capture human data better than risk-neutral QRE—a finding which is in line with findings in behaviorial economics (Goeree et al., 2003; Goeree and Offerman, 2002).

Risk-averse and robust multi-agent reinforcement learning. Our work builds on and provides an additional justification for risk-sensitive (multi-agent) reinforcement learning. This line of work has roots going back to seminal work by Jacobson on risk-sensitive control (Jacobson, 1973), and more recently in risk-sensitive reinforcement learning (Shen et al., 2014). In these works, the aim is to find a controller or policy for a system that accounts for stochasticity or uncertainty in the environment or system in a more nuanced way than risk-neutral approaches like optimal control or reinforcement learning (Borkar, 2023). Due to classic duality results (see e.g., (Panaganti et al., 2022; Zhang et al., 2024)) this line of work is closely related to the literature on robust control and distributionally robust reinforcement learning (Iyengar, 2005; Panaganti and Kalathil, 2022; Xu* et al., 2023) which seeks to find solutions that are robust to worst case environmental disturbances.

Our work rigorously extends these formulations to the multi-agent regime though it is not the first to consider risk-aversion in MARL. Indeed, risk-sensitive MARL has been the focus of several recent works (e.g., (Yekkehkhany et al., 2020; Wang et al., 2024; Gao et al., 2021; Slumbers et al., 2023)). Several provide rigorous definitions of risk-averse equilibria and some guarantees on their computation by assuming structure on the risk-averse game. Oftentimes this is done by assuming that the risk-averse game is itself zero-sum (Yekkehkhany et al., 2020), monotone (Wang et al., 2024), or that it satisfies other strong conditions (Gao et al., 2021). Other works are more empirical in nature (Ganesh et al., 2019; Eriksson et al., 2022; Zhang et al., 2021b; Qiu et al., 2021; Shen et al., 2023; Slumbers et al., 2023), showing the promise of risk-averse algorithms for MARL.

Another line of work originates in the literature on robust and risk-sensitive control where the focus has largely been on understanding the solution to stochastic dynamic games (Borkar, 2023; Basar, 1999; Moon et al., 2019). In these contexts, the focus has often remained on exponential forms of risk and on characterizing properties of the Nash equilibrium solutions—but questions of computational tractability are mostly side-stepped or the problems are analyzed under conditions on the risk-adjusted

game like weak coupling (Basar, 1999) or zero-sum structures (Moon et al., 2019) which may not arise from risk-averse problems.

One last closely related line of work is the emerging literature on robust multi-agent reinforcement learning (Zhang et al., 2020; He et al., 2023; Shi et al., 2024; Blanchet et al., 2024). Once again due to duality arguments, these works can be seen as tackling a similar problem to the risk-averse MARL problem. The focus of these previous works, however, is on robustness in the face of only environmental uncertainties (and not opponent strategies), and questions of existence and computational tractability are either assumed away or the focus is on extensions of correlated equilibrium concepts. A recent related work in this literature analyzed the computational tractability of robust Nash equilibria in Markov games, but only provided strong guarantees on the zero-sum regime, showing that computing such equilibria in general is PPAD-hard (McMahan et al., 2024).

To the best of our knowledge, no previous work in either of these literatures highlights the broad benefits afforded by risk-aversion in MARL in terms of computational tractability of equilibria. In our work we show that risk-aversion (and by extension distributional robustness) to all external randomness, when combined with bounded rationality yields a computationally tractable class of individually rationalizable equilibria in *all* finite-horizon n-player Markov games. Furthermore we show that these equilibria can be computed using no-regret learning algorithms.

B Details of Risk-averse and Bounded rational Matrix Games

In the next two subsections, we provide detailed technical results pertaining to modeling agents as both risk-averse and imperfect optimizers—embedding agents with human decision-making capabilities. These technical results serve as a more rigorous and detailed understanding of our main methodology choices presented in Section 2.

B.1 Risk-Aversion in Matrix Games

In the case of matrix games, we consider the case where agents are risk averse with respect to the mixing or randomness introduced into the game by their opponents. In Markov games, we will also consider the case where agents are risk-averse with respect to the underlying dynamics. To emphasize, we assume players are not risk-averse to their own randomness. We note that is a common approach taken in the literature on risk-sensitive and robust decision-making (Shen et al., 2014) and it is necessary since if agents are risk-averse to their own randomness then an equilibrium may cease to exist (Fiat and Papadimitriou, 2010). We refer to related works for more discussion.

We now recall our generalization of the original game that differ in how risk is incorporated into the problem: *aggregate risk aversion* (4). We also formulate the *action-dependent risk aversion* generalization of the original game here in (17).

Aggregate Risk Aversion Game: The player's utilities take the form

$$f_i(\pi_i, \pi_{-i}) = \rho_{i,\pi_{-i}}(\mathbb{E}_{\pi_i}[R_i(a)]) = \rho_{i,\pi_{-i}}\left(\sum_{a_i \in \mathcal{A}_i} \pi_i(a_i) R_i(a_i, a_{-i})\right) \tag{16}$$

where $\rho_{i,\pi_{-i}}$ is used to denote the potentially different risk preference of agent i which depends on the product distribution of opponents strategies π_{-i}.

Action-dependent Risk Aversion Game: The player's utilities take the form

$$f_i(\pi_i, \pi_{-i}) = \mathbb{E}_{\pi_i}\left[\rho_{i,\pi_{-i}}(R_i(a))\right] = \sum_{a_i \in \mathcal{A}_i} \pi_i(a_i) \rho_{i,\pi_{-i}}\left(R_i(a_i, a_{-i})\right) \tag{17}$$

where again $\rho_{i,\pi_{-i}}$ is used to denote the potentially different risk preference of agent i which depends on the product distribution of opponents strategies π_{-i}.

We remark that in both of these formulations, if $\rho_i(X) = \mathbb{E}[-X]$ for all players $i = 1, ..., n$ (which satisfies the requirements in Definition 3) then the new formulation reduces to the original expected utility objective. Thus, both can be seen as generalizations of the classic setup described in the previous section.

Risk-measure	Penalty function $D(p, q)$
Entropic Risk (Ahmadi-Javid, 2012)	Kullback-Leibler (KL): $KL(p, q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$
(Föllmer and Schied, 2002)	Reverse KL (RKL): $\sum_i q_i \log\left(\frac{q_i}{p_i}\right)$
ϕ-Entropic Risk (Ahmadi-Javid, 2012)	ϕ-Divergence: $\sum_i p_i \, \phi\left(\frac{p_i}{q_i}\right)$
Utility-based shortfall (Föllmer and Schied, 2002)	Utility-based Shortfall (u): $\inf_{\lambda>0} \frac{1}{\lambda}[r + \sum_i p_i u^\star(\lambda \frac{q_i}{p_i})]$

Table 1: We list several widely-used convex risk-measures with its penalty function $D(p, q)$. Here, p, q are distributions of the same finite dimension, ϕ and u are differentiable convex functions on $\mathbb{R}$ satisfying $\phi(1) = 0$ and $\phi(t) = +\infty$ for $t < 0$. The utility u is equipped with risk tolerance level of r, and its penalty function depends on its convex conjugate $u^\star$.

We also remark that in both of these formulations, due to monotonicity and linearity of expectation both can be seen as generalizations of risk-sensitive decision-making investigated in the literature on risk-sensitive control (Borkar, 2023) and risk-sensitive RL (Shen et al., 2014). To see this reduction, one can simply take the other agents to be part of an unknown environment.

We now rely on a particularly powerful property of convex measures of risk to simplify and expose some structure for these two risk-adjusted games. In particular, we make use of the following *dual* representation theorem for convex risk measures.

Theorem 6 (Dual Representation Theorem for Convex Risk Measures (Föllmer and Schied, 2002))**.** *Suppose that the set $\mathcal{X}$ is the set of functions mapping from a finite set Ω to $\mathbb{R}$. Then a mapping $\rho : \mathcal{X} \to \mathbb{R}$ is a convex risk measure (cf. Definition 3) if and only if there exists a penalty function $D : \Delta_\Omega \to (-\infty, \infty]$ such that: $\rho(X) = \sup_{p\in\Delta_\Omega} E_p[-X] - D(p)$, where Δ_Ω is the set of all probability measures on Ω. Furthermore, the function $D(p)$ can be taken to be convex, lower-semi-continuous, and satisfy $D(p) > -\rho(0)$ for all $p \in \Delta_\Omega$.*

When the set $\mathcal{X}$ is again the set of measurable functions defined on a probability space, one can choose the penalty function D to represent a notion of of distance from the probability law or distribution of the random variable π. In such cases, the dual representation theorem takes the form:

$$\rho_\pi(X) = \sup_{p\in\Delta_\Omega} E_p[-X] - D(p, \pi),$$

where $D(p, \pi)$ is convex in p for a fixed π. We also make a simplifying assumption that D is continuous in both its arguments, which is satisfied by various widely-used risk measures. This general form allows us to draw connections with a large class of risk and robustness metrics that are based around ϕ-divergences. We provide examples of common risk measures in Table 1.

Given this reformulation tool, we recover the alternate aggregate risk-averse game (5) in the following form:

$$f_i(\pi_i, \pi_{-i}) = \sup_{p_i\in\mathcal{P}_{-i}} -\pi_i^T R_i p_i - D_i(p_i, \pi_{-i}) \tag{18}$$

where $\mathcal{P}_{-i} = \mathcal{P}/\Delta_{A_i} \subset \mathbb{R}^{A_{-i}}$, $A_{-i} = \sum_{j\neq i} A_j$, and $R_i \in \mathbb{R}^{A_i \times A_{-i}}$ is player i's payoff matrix. Similarly, the action-dependent risk-averse game also takes the form:

$$f_i(\pi_i, \pi_{-i}) = \sum_{j\in\mathcal{A}_i} \pi_i(j) \left(\sup_{p_{i,j}\in\mathcal{P}_{-i}} -\langle R_{i,j}, p_{i,j}\rangle - D_i(p_{i,j}, \pi_{-i}) \right), \tag{19}$$

where $R_{i,j}$ corresponds to the jth row of R_i.

As noted before, different penalty functions D_i allow different agents to have different risk preferences. In this form, one can see that in a risk-averse game, the players imagine that intermediate adversaries seek to maximize their cost but are penalized from deviating too far from the opponents' realized strategies. Thus, agents in risk-averse games introduce a certain amount of worst-case thinking into their strategies.

We have thus introduced a risk-averse Nash equilibrium (RNE) concept (see Definition 4). The convexity and continuity of the penalty function guarantees that the risk-averse games admit at least one RNE.

Theorem 7. *There always exists a RNE for all aggregate and action-dependent risk-averse games presented in* (5) *and* (19) *respectively.*

Proof. To begin, we show that $f_i(\pi_i, \pi_{-i})$ is convex in π_i for all $\pi_i \in \mathcal{P}_{-i}$ in (5). This follows from the fact that $D_i(\cdot, \cdot)$ is assumed to be convex and continuous in its first argument. Invoking Danskin's theorem guarantees us that f_i is convex in π_i for all fixed values of π_{-i}. Since the probability simplex is compact and convex, the game satisfies the conditions of a convex game (Rosen, 1965) and thus a Nash equilibrium must exist. The action-dependent risk aversion regime follows from the linearity of f_i in π_i in (19) and a similar invocation of the existence of Nash in convex games (Rosen, 1965). □

We note that even though risk preferences already help convexify player's objectives, the addition of risk can serve to weaken existing structures in the player's cost function. To illustrate this, we show that even in two-player zero-sum games where $R_1 = R = -R_2^T$, the risk-averse game loses any zero-sum structure and may cease to even be strictly competitive in the sense that one player's gain is the other's loss. Thus, additional convexity induced by the introduction of risk aversion guarantee is not enough to ensure the computational tractability of the Nash equilibrium (see e.g., (McMahan et al., 2024)).

Example 1. *Consider a 2-player zero-sum game where $R_1 = R = -R_2^T$ where players have aggregate risk aversion in the entropic risk metric with different degrees of risk aversion τ_1 and τ_2. Under these conditions, the players loss functions take the following form:*

$$f_1(\pi_1, \pi_2) = \sup_{p_1 \in \mathcal{A}_2} -\pi_1^T R p_1 - \frac{1}{\tau_1} KL(p_1, \pi_2) = \frac{1}{\tau_1} \log\Big(\sum_{1 \le j \le A_2} \pi_2(j) \exp(-\tau_1 [R\pi_1]_j) \Big)$$

$$f_2(\pi_1, \pi_2) = \sup_{p_2 \in \mathcal{A}_1} \pi_2^T R^T p_2 - \frac{1}{\tau_2} KL(p_2, \pi_1) = \frac{1}{\tau_2} \log\Big(\sum_{1 \le j \le A_1} \pi_1(j) \exp(\tau_2 [R^T \pi_2]_j) \Big).$$

Even instantiating $R = \mathbb{I}_2$, $\tau_1 = 10$, and for any $\tau_2 > 0$, both $f_1(\pi_1, \pi_2) > f_1(\pi_1', \pi_2)$ and $f_2(\pi_1, \pi_2) > f_2(\pi_1', \pi_2)$ holds for the regions $\pi_1, \pi_2 \in \Delta_2$ satisfying $\pi_2(1) \in (0.1, 0.5)$ and $0.75 - \pi_1(1) > \pi_1'(1) > \pi_1(1)$, which implies that the game is not a strictly competitive game.

Note that the previous example introduces degrees of risk aversion into the game through the parameters $\tau_1 > 0$ and $\tau_2 > 0$. As τ increases, the game becomes less reliant on the regularization term, which makes the adversary more powerful and results in more conservative game playing by the player. As τ goes to zero we recover the risk neutral regime (Ahmadi-Javid, 2012).

B.2 Bounded Rationality in Matrix Games

Since risk aversion on its own is not sufficient to guarantee computational tractability of NE, we introduced human decision-making into the game: bounded rationality. To incorporate bounded rationality into agents, we resort to the notion of a *quantal response function* (see Definition 2).

Clearly, when players responses are constrained to be *quantal* responses, they cannot be perfect maximizers since the $\arg\max$ function does not satisfy the first desiderata of a quantal response function. Common quantal response functions include the logit response function:

$$\sigma(x) = \frac{\exp(-\frac{1}{\epsilon} x_i)}{\sum_{j=1}^n \exp(-\frac{1}{\epsilon} x_j)}, \tag{20}$$

where the sign is to account for the fact that agents may be *minimizing* their loss function.

To incorporate quantal responses into our risk-averse game, we introduced regularization to the player's losses in (6). Recalling:

$$f_i^{\epsilon_i}(\pi_i, \pi_{-i}) = f_i(\pi_i, \pi_{-i}) + \epsilon_i \nu_i(\pi_i) \tag{21}$$

where ν_i is strictly convex over the simplex and controls the class of quantal responses available to player i and ϵ_i controls the agent's degree of bounded rationality. This can be shown to be equivalent to constraining the player's responses to quantal responses (see, e.g., (Föllmer and Schied, 2002, Proposition 7), or (Sokota et al., 2023; Mertikopoulos and Sandholm, 2016)).

Example 2. *If players are constrained to* logit *response functions, one can reflect this by incorporating a negative entropy regularizer* $\nu(\pi) = \sum_i p_i \log(p_i)$. *Another class of quantal response functions would be generated by making use of e.g., a log-barrier regularizer* $\nu(\pi) = -\sum_i \log(p_i)$. *Both of these regularizers give rise to quantal response functions that satisfy Definition 2.*

This game now incorporates two key properties of human decision-making: risk aversion and bounded rationality on the part of the agents. The natural outcome of this game is what we termed a risk-averse quantal-response equilibrium (RQE).

C Further results on computational tractability of RQE in aggregate risk-averse matrix games

In this section we provide further results on the aggregate risk aversion formulation of the risk-averse problem. This section provides a proof of the main result of Theorem 3 presented in the main body of the paper as well as the generalization of the results on computationaly tractability to n-players. A key step in the proof is relating the Nash equilibrium of the $2n$-player convex game presented where players have utilites of the form (8) and (7) and RQE of the desired game (4). We then use this result to prove Theorem 3.

C.1 Relating Nash of the $2n$-player game and RQE

We first note that this $2n$ player game is a convex game played over compact convex action sets and thus a Nash equilibrium must exist (Rosen, 1965). To relate outcomes in this new game to that of the original game, we show that the strategies played by the original players in Nash equilibria of the $2n$-player game coincide with the RQE.

Proposition 1. *If* (π^*, p^*) *is a Nash equilibrium of the* $2n$*-player game, then* π^* *is a RQE of Game (6). Furthermore, if* π^* *is a RQE of Game (6), then* (π^*, p^*) *is a Nash equilibrium of the* $2n$*-player game, where:*

$$p_i^* = \arg\max_{p_i \in \mathcal{P}_{-i}} -{\pi_i^*}^T R_i p_i - D_i(p_i, \pi_{-i}^*). \tag{22}$$

Proof. To prove this result, we rely on the definitions of Nash and RQE. We begin by proving the first claim of the proposition. Recall that a Nash equilibrium is a joint strategy $(\pi^*, p^*) \in \mathcal{P} \times \bar{\mathcal{P}}$ such that, for all $i = 1, ..., n$:

$$\begin{aligned} J_i(\pi_i^*, \pi_{-i}^*, p^*) &\leq J_i(\pi_i', \pi_{-i}^*, p^*) \ \ \forall \ \ \pi_i' \in \Delta_{A_i} \\ \bar{J}_i(\pi^*, p_i^*, p_{-i}^*) &\leq \bar{J}_i(\pi^*, p_i', p_{-i}^*) \ \ \forall \ \ p_i' \in \mathcal{P}_{-i} \end{aligned} \tag{23}$$

Noting that each $J_i(\pi_i^*, \pi_{-i}^*, p_i^*, p_{-i}^*)$ does not depend on p_{-i}^*, to show the forward direction, we start by taking the supremum of the right hand side of (23) over $p_i \in \mathcal{P}_{-i}$. Thus, we find that, for all i:

$$\begin{aligned} J_i(\pi_i^*, \pi_{-i}^*, p^*) &\leq \sup_{p_i \in \mathcal{P}_{-i}} J_i(\pi_i', \pi_{-i}^*, p_i, p_{-i}^*) \ \ \forall \ \ \pi_i' \in \Delta_{A_i} \\ &= f_i^{\epsilon_i}(\pi_i', \pi_{-i}^*) \ \ \forall \ \ \pi_i' \in \Delta_{A_i}, \end{aligned}$$

where in the second line we used the fact that for any π, p_{-i}, by definition,

$$\sup_{p_i \in \mathcal{P}_{-i}} J_i(\pi, p_i, p_{-i}) = f_i^{\epsilon_i}(\pi).$$

It remains to show that for any $i = 1, ..., n$, $J_i(\pi_i^*, \pi_{-i}^*, p^*) = f_i^{\epsilon_i}(\pi_i^*, \pi_{-i}^*)$. This follows from the fact that the simplex is compact. Indeed for any fixed π, the function $J_i(\pi, p_i, p_{-i})$ is concave in p_i. Thus, the supremum is attained at p_i^* since

$$p_i^* = \arg\min_{p_i \in \mathcal{P}_{-i}} \bar{J}_i(\pi^*, p_i, p_{-i}^*) = \arg\max_{p_i \in \mathcal{P}_{-i}} J_i(\pi^*, p_i, p_{-i}^*).$$

Since the same argument holds for all i, we have shown that if (π^*, p^*) is a Nash equilibrium of the $2n$-player game, then:

$$f_i^{\epsilon_i}(\pi_i^*, \pi_{-i}^*) \leq f_i^{\epsilon_i}(\pi_i, \pi_{-i}^*) \ \ \forall \ \ \pi_i \in \Delta_{A_i}$$

which is the definition of a RQE.

To prove the second claim, suppose that π^* is a RQE. By definition, we have that if

$$p_i^* = \arg\max_{p_i \in \mathcal{P}_{-i}} -\pi_i^{*T} R_i p_i - D_i(p_i, \pi_{-i}^*).$$

then p_i^* by construction satisfies the condition of a Nash equilibrium of the $2n$-player game:

$$\bar{J}_i(\pi^*, p_i^*, p_{-i}^*) \leq \bar{J}_i(\pi^*, p_i', p_{-i}^*) \ \ \forall \ \ p_i' \in \mathcal{P}_{-i}.$$

It remains to show that π^* also satisfies the necessary conditions on J_i. To see that this must hold, by definition, we must have that (π_i^*, p_i^*) satisfies:

$$J_i(\pi_i^*, p_i^*, \pi_{-i}^*, p_{-i}^*) = f_i^{\epsilon_i}(\pi_i^*, \pi_{-i}^*) = \min_{\pi_i \in \Delta_{A_i}} \max_{p_i \in \mathcal{P}_{-i}} J(\pi_i, p_i, \pi_{-i}^*, p_{-i}^*)$$

Further manipulations allow us to show that:

$$\begin{aligned} J_i(\pi_i^*, \pi_{-i}^*, p^*) &= f_i^{\epsilon}(\pi_i^*, \pi_{-i}^*) \\ &= \min_{\pi_i \in \Delta_{A_i}} \max_{p_i \in \mathcal{P}_{-i}} J(\pi_i, p_i, \pi_{-i}^*, p_{-i}^*) \\ &= \min_{\pi_i \in \Delta_{A_i}} J(\pi_i, p_i^*, \pi_{-i}^*, p_{-i}^*) \\ &\leq J(\pi_i', p_i^*, \pi_{-i}^*, p_{-i}^*) \ \ \forall \ \ \pi_i' \in \Delta_{A_i} \end{aligned}$$

Since this holds true for all i, this completes the proof. □

C.2 Proof of Theorem 3 and immediate corollaries

Given the results showing that finding a Nash equilibrium of the $2n$-player convex game allows us to compute RQE, we present the proof of Theorem 3. We also provide two corollaries that specialize the results to different risk-metrics and quantal response functions.

Theorem 8 (Restatement of Theorem 3)**.** *Assume the penalty functions that give rise to the players' risk preferences $D_1(\cdot,\cdot)$ and $D_2(\cdot,\cdot)$ are jointly convex in both their arguments. If σ is a CCE of the four player game with $\xi_1 = \xi_1^*$ and $\xi_2 = \xi_2^*$, and*

$$\frac{\epsilon_1}{\xi_1^*} \geq \frac{\xi_2^*}{\epsilon_2},$$

then $\hat{\pi}_1 = \mathbb{E}_\sigma[\pi_1]$ and $\hat{\pi}_2 = \mathbb{E}_\sigma[\pi_2]$ constitute a RQE of the original game.

The proof follows by showing that CCE of the $2n$-player convex game coincide with Nash equilibria of the $2n$-player convex game, and then invoking results from the previous section.

Proof. To prove this, we show that $(\hat{\pi}_1, \hat{p}_1, \hat{\pi}_2, \hat{p}_2)$ is a Nash equilibrium of the four player game and then invoke Proposition 1 (where $\hat{\pi}_1 = \mathbb{E}_\sigma[\pi_1]$, $\hat{p}_2 = \mathbb{E}_\sigma[p_2]$,$\hat{\pi}_2 = \mathbb{E}_\sigma[\pi_2]$, and $\hat{p}_2 = \mathbb{E}_\sigma[p_2]$).

To begin, we focus on J_1 and $\bar{J}_1$. By symmetry, the same arguments hold for J_2 and $\bar{J}_2$. We first note that J_1 is (strictly) convex in π_1 for all fixed p_1, π_2, p_2 and jointly concave in p_1, π_2, p_2 for all fixed π_1. Now, starting with the definition of CCE for σ and via Jensen's inequality, we have that:

$$\begin{aligned} \mathbb{E}_\sigma[J_1(\pi_1, p_1, \pi_2, p_2)] &\leq \mathbb{E}_{(p_1,\pi_2,p_2)\sim\sigma}[J_1(\pi_1', p_1, \pi_2, p_2)] \\ &= \mathbb{E}_{p_1 \sim \hat{p}_1}\left[\mathbb{E}_{\pi_2 \sim \sigma \,|\, p_1}\left[\mathbb{E}_{p_2 \sim \sigma \,|\, p_1, \pi_2}\left[J_1(\pi_1', p_1, \pi_2, p_2)\right]\right]\right] \\ &\leq \mathbb{E}_{p_1 \sim \hat{p}_1}\left[\mathbb{E}_{\pi_2 \sim \sigma \,|\, p_1}\left[J_1(\pi_1', p_1, \pi_2, \mathbb{E}_{\sigma \,|\, p_1, \pi_2}[p_2])\right]\right] \\ &\leq \mathbb{E}_{p_1 \sim \hat{p}_1}\left[J_1(\pi_1', p_1, \mathbb{E}_{\sigma \,|\, p_1}[\pi_2], \mathbb{E}_{\sigma \,|\, p_1}[p_2])\right] \\ &\leq J_1(\pi_1', \hat{p}_1, \hat{\pi}_2, \hat{p}_2) \quad \forall \pi_1' \in \Delta_{A_1}. \end{aligned} \tag{24}$$

Similarly, we note that for $\xi_1 = \xi_1^*$, $\bar{J}_1$ is jointly concave in π_1, π_2, p_2, such that:

$$\mathbb{E}_\sigma[\bar{J}_1(\pi_1, p_1, \pi_2, p_2)] \leq \bar{J}_1(\hat{\pi}_1, p_1', \hat{\pi}_2, \hat{p}_2) \quad \forall p_1' \in \Delta_{A_2}.$$

Letting $\hat{z} = (\hat{\pi}_1, \hat{p}_1, \hat{\pi}_2, \hat{p}_2)$ and $z = (\pi_1, p_1, \pi_2, p_2)$ to simplify notation, we can now take a weighted sum of the four utility functions with $\lambda \in (0, 1)$ to find that:

$$\begin{aligned}
&\frac{\lambda}{2}\left(J_1(\hat{z}) + \bar{J}_1(\hat{z})\right) + \frac{1-\lambda}{2}\left(J_2(\hat{z}) + \bar{J}_2(\hat{z})\right) \\
&\quad \geq \mathbb{E}_\sigma\left[\frac{\lambda}{2}\left(J_1(z) + \bar{J}_1(z)\right) + \frac{1-\lambda}{2}\left(J_2(z) + \bar{J}_2(z)\right)\right] \\
&\quad = \frac{1}{2}\left(\lambda\epsilon_1 - (1-\lambda)\xi_2^*\right)\mathbb{E}_\sigma\left[\nu_1(\pi_1)\right] + \frac{1}{2}\left((1-\lambda)\epsilon_2 - \lambda\xi_1^*\right)\mathbb{E}_\sigma\left[\nu_2(\pi_2)\right].
\end{aligned}$$

Choosing $\lambda = \xi_2^*/(\epsilon_1 + \xi_2^*)$, we can further simplify to find that:

$$\begin{aligned}
&\mathbb{E}_\sigma\left[\frac{\lambda}{2}\left(J_1(\hat{z}) + \bar{J}_1(\hat{z})\right) + \frac{1-\lambda}{2}\left(J_2(\hat{z}) + \bar{J}_2(\hat{z})\right)\right] \\
&\quad = \frac{1}{2}\left(\frac{\epsilon_1\epsilon_2}{\epsilon_1 + \xi_2^*} - \frac{\xi_1^*\xi_2^*}{\epsilon_1 + \xi_2^*}\right)\mathbb{E}_\sigma\left[\nu_2(\pi_2)\right] \\
&\quad \geq \frac{1}{2}\left(\frac{\epsilon_1\epsilon_2}{\epsilon_1 + \xi_2^*} - \frac{\xi_1^*\xi_2^*}{\epsilon_1 + \xi_2^*}\right)\nu_2(\hat{\pi}_2) \\
&\quad = \frac{\lambda}{2}\left(J_1(\hat{z}) + \bar{J}_1(\hat{z})\right) + \frac{1-\lambda}{2}\left(J_2(\hat{z}) + \bar{J}_2(\hat{z})\right),
\end{aligned}$$

where we used the fact that $\frac{\epsilon_1}{\xi_1^*} \geq \frac{\xi_2^*}{\epsilon_2}$ by assumption and invoke Jensen's inequality for ν_2 at the second inequality. Thus we have shown that:

$$\frac{\lambda}{2}\left(J_1(\hat{z}) + \bar{J}_1(\hat{z})\right) + \frac{1-\lambda}{2}\left(J_2(\hat{z}) + \bar{J}_2(\hat{z})\right) \tag{25}$$

$$= \frac{\lambda}{2}\left(\mathbb{E}_\sigma\left[J_1(z)\right] + \mathbb{E}_\sigma\left[\bar{J}_1(z)\right]\right) + \frac{1-\lambda}{2}\left(\mathbb{E}_\sigma\left[J_2(z)\right] + \mathbb{E}_\sigma\left[\bar{J}_2(z)\right]\right). \tag{26}$$

By Eq. (24), we observe:

$$\mathbb{E}_\sigma\left[J_1(z)\right] \leq J_1(\hat{z}), \mathbb{E}_\sigma\left[\bar{J}_1(z)\right] \leq \bar{J}_1(\hat{z}), \mathbb{E}_\sigma\left[J_2(z)\right] \leq J_2(\hat{z}), \mathbb{E}_\sigma\left[\bar{J}_2(z)\right] \leq \bar{J}_2(\hat{z}). \tag{27}$$

We note a fact: $\lambda a + (1-\lambda)b = \lambda c + (1-\lambda)d$ and $a \leq c, b \leq d$ implies $a = c, b = d$ for any $a, b, c, d \in \mathbb{R}$. Using this fact for Eqs. (26) and (27), we have:

$$\begin{aligned}
\mathbb{E}_\sigma\left[J_1(z)\right] &= J_1(\hat{z}) \leq J_1(\pi_1', \hat{p}_1, \hat{\pi}_2, \hat{p}_2) \quad \forall \pi_1' \in \Delta_{A_1} \\
\mathbb{E}_\sigma\left[\bar{J}_1(z)\right] &= \bar{J}_1(\hat{z}) \leq \bar{J}_1(\hat{\pi}_1, p_1', \hat{\pi}_2, \hat{p}_2) \quad \forall p_1' \in \Delta_{A_2} \\
\mathbb{E}_\sigma\left[J_2(z)\right] &= J_2(\hat{z}) \leq J_2(\hat{\pi}_1, \hat{p}_1, \pi_2', \hat{p}_2) \quad \forall \pi_2' \in \Delta_{A_2} \\
\mathbb{E}_\sigma\left[\bar{J}_2(z)\right] &= \bar{J}_2(\hat{z}) \leq \bar{J}_2(\hat{\pi}_1, \hat{p}_1, \hat{\pi}_2, p_2') \quad \forall p_2' \in \Delta_{A_1}
\end{aligned}$$

Thus we have shown that $\hat{z} = (\hat{\pi}_1, \hat{p}_1, \hat{\pi}_2, \hat{p}_2)$ is a Nash equilibrium for the 4-player game. By invoking Propostion 1 we derive our result that $(\hat{\pi}_1, \hat{\pi}_2)$ must be a RQE for the original risk-averse game. □

We remark that the result does not necessarily guarantee uniqueness, though by exploiting the connections between socially convex games and monotone games (Gemp and Mahadevan, 2017) such a result would follow.

We now present two corollaries that specialize the results to specific risk metrics and quantal responses. In the first we look at the case where players make use of the entropic risk and log-barrier reguarlizers.

Corollary 8.1. *Suppose the players are risk-averse in the entropic risk metric with parameters τ_1 and τ_2 respectively, meaning that their risk-averse losses are given by:*

$$f_1(\pi_1, \pi_2) = \sup_{p_1 \in \mathcal{A}_2} -\pi_1^T R_1 p_1 - \frac{1}{\tau_1} KL(p_1, \pi_2) \qquad f_2(\pi_1, \pi_2) = \sup_{p_2 \in \mathcal{A}_1} \pi_2^T R_2 p_2 - \frac{1}{\tau_2} KL(p_2, \pi_1)$$

If they respond in the space of quantal responses generated by the log-barrier regularizers with parameters ϵ_1 and ϵ_2 respectively, and if $\epsilon_1\tau_1 \geq \frac{1}{\epsilon_2\tau_2}$ then, for any R_1, R_2 the players can compute a RQE by using no-regret learning.

In the second corollary we look at the case when players make use of the reverse KL penalty function and logit quantal responses.

Corollary 8.2. *Suppose the players are risk-averse and make use of the reverse-KL as a penalty function to give rise to their risk metric with parameters τ_1 and τ_2 respectively. Their risk-averse losses are given by:*

$$f_1(\pi_1, \pi_2) = \sup_{p_1 \in \mathcal{A}_2} -\pi_1^T R_1 p_1 - \frac{1}{\tau_1} RKL(p_1, \pi_2) \;\; f_2(\pi_1, \pi_2) = \sup_{p_2 \in \mathcal{A}_1} \pi_2^T R_2 p_2 - \frac{1}{\tau_2} RKL(p_2, \pi_1)$$

If they respond in the space of quantal responses generated by the negative entropy regularizor with parameters ϵ_1 and ϵ_2 respectively, and if $\epsilon_1 \tau_1 \geq \frac{1}{\epsilon_2 \tau_2}$ then, for any R_1, R_2 the players can compute a RQE by using no-regret learning.

Proof. The proof of these corollaries result comes from the fact that for $\xi^* = \frac{1}{\tau}$, the function $H(p, \pi) = \frac{1}{\tau} RKL(p, \pi) - \xi \nu(\pi)$ is concave in π for all p if $\xi \geq \frac{1}{\tau}$. Thus, choosing $\xi^* = \frac{1}{\tau}$ and invoking Theorem 3 completes the proof. □

Note that the proof of this corollary is the same as that of Corollary 8.1 and so we only provide one proof for both.

C.3 Computing RQE in n-player General-Sum Games

We now extend our result to the computation of RQE in n-player games. This requires stronger assumptions on the players' risk preferences and bounded rationality parameters. Nevertheless, we once again show that a large class of RQE is computationally tractable in this class of games.

To do so we now define $H_i(p_i, \pi_{-i}) = D_i(p_i, \pi_{-i}) - \sum_{j \neq i} \xi_{i,j} \nu_j(\pi_j)$. For all $i, j \in \{1, ..., n\}$ let $\xi^*_{i,j} > 0$ be the smallest values of $\xi_{i,j}$ such that $H_i(p_i, \pi_{-i})$ is concave in π_j. Again, the parameters $\xi^*_{i,j}$ capture the player's degrees of risk aversion. The following theorem gives a general condition under which an RQE is computable using no-regret learning.

Theorem 9. *Assume the penalty functions that give rise to the players' risk preferences $D_i(\cdot, \cdot)$ are jointly convex in both of their arguments. If σ is a CCE of the $2n$-player game with $\xi_{i,j} = \xi^*_{i,j}$ for all $i, j \in \{1, ..., n\}$, and for all $i = 1, ..., n$ we have $\epsilon_i \geq \sum_{j \neq i} \xi^*_{j,i}$, then $\hat{\pi} = \mathbb{E}_\sigma[\pi]$ is a RQE of the risk-averse n-player game.*

Proof. To prove this, we show that $(\hat{\pi}, \hat{p})$ is a Nash equilibrium of the $2n$-player game and then invoke Proposition 1 (where $\hat{\pi} = \mathbb{E}_\sigma[\pi]$, $\hat{p} = \mathbb{E}_\sigma[p]$).

We focus on J_i and $\bar{J}_i$. We first note that J_i is (strictly) convex in π_i for all fixed p, π_{-i} and jointly concave in p, π_{-i} for all fixed π_i by assumption. Thus, via Jensen's inequality, we have that:

$$\mathbb{E}_\sigma[J_i(\pi, p)] \leq J_i(\pi_i', \hat{\pi}_{-i}, \hat{p}) \quad \forall \pi_i' \in \Delta_{A_i}$$

Similarly, we note that for $\xi_{i,j} = \xi^*_{i,j}$, $\bar{J}_i$ is jointly concave in π, p_{-i}, such that:

$$\mathbb{E}_\sigma[\bar{J}_i(\pi, p)] \leq \bar{J}_i(\hat{\pi}, p_i', \hat{p}_{-i}) \quad \forall p_i' \in \mathcal{P}_{-i}.$$

We can now take a sum of the $2n$ utility functions to find that:

$$\begin{aligned}
\sum_{i=1}^n J_i(\hat{\pi}, \hat{p}) + \bar{J}_i(\hat{\pi}, \hat{p}) &\geq \sum_{i=1}^n \mathbb{E}_\sigma[J_i(\pi, p)] + \mathbb{E}_\sigma[\bar{J}_i(\pi, p)] \\
&= \sum_{i=1}^n \left(\epsilon_i - \sum_{j \neq i} \xi^*_{j,i} \right) \mathbb{E}_\sigma[\nu_i(\pi_i)] \\
&\geq \sum_{i=1}^n \left(\epsilon_i - \sum_{j \neq i} \xi^*_{j,i} \right) \nu_i(\hat{\pi}_i)
\end{aligned}$$

$$= \sum_{i=1}^{n} J_i(\hat{\pi}, \hat{p}) + \bar{J}_i(\hat{\pi}, \hat{p})$$

where we used the assumed condition on ϵ_i to guarantee convexity of the functions of π_i in the second line, allowing us to use Jensen's inequality to derive the third line.

Thus we have shown that:

$$\sum_{i=1}^{n} J_i(\hat{\pi}, \hat{p}) + \bar{J}_i(\hat{\pi}, \hat{p}) = \sum_{i=1}^{n} \mathbb{E}_\sigma[J_i(\pi, p)] + \mathbb{E}_\sigma[\bar{J}_i(\pi, p)]$$

This implies that:

$$J_i(\hat{\pi}, \hat{p}) = \mathbb{E}_\sigma[J_i(\pi, p)] \leq J_i(\pi_i', \hat{\pi}_{-i}, \hat{p}) \quad \forall \pi_i' \in \Delta_{A_i}$$
$$\bar{J}_i(\hat{\pi}, \hat{p}) = \mathbb{E}_\sigma[\bar{J}_i(\pi, p)] \leq \bar{J}_i(\hat{\pi}, p_i', \hat{p}_{-i}) \quad \forall p_i' \in \mathcal{P}_{-i}.$$

Thus we have shown that $(\hat{\pi}, \hat{p})$ is a Nash equilibrium for the $2n$-player game. By invoking Proposition 1 we can observe that $\hat{\pi}$ must be a RQE for the risk-averse game. □

D Computational tractability of RQE in action-dependent risk-averse matrix games

As in the case of aggregate risk-aversion, to prove our results we again introduce an auxiliary game that we relate to our risk-averse game of interest. In this case, the loss of the original players is given by:

$$J_i(\pi_i, \pi_{-i}, p) = \sum_{j \in \mathcal{A}_i} \pi_i(j) \left(-\langle R_{i,j}, p_{i,j} \rangle - D_i(p_{i,j}, \pi_{-i})\right) + \epsilon_i \nu_i(\pi_i). \tag{28}$$

We now associate each player i to its intermediate adversary p_i. For each player p_i their loss function is given by:

$$\bar{J}_i(\pi, p_i) = \sum_{j \in \mathcal{A}_i} \pi_i(j) \left(\langle R_{i,j}, p_{i,j} \rangle + D_i(p_{i,j}, \pi_{-i})\right) - \sum_k \xi_{i,k} \nu_k(\pi_k), \tag{29}$$

where $p_i = \{p_{i,j}\}_{j \in \mathcal{A}_i}$ where $p_{i,j} \in \mathcal{P}_{-i}$. This is once again a convex game since each player's loss is convex in its own argument. Define $\xi_{i,k}^* \geq 0$ as the minimum value of $\xi_{i,k}$ needed for $\bar{J}_i(\pi, p_i)$ to be concave in π for all values of p_i for all i. Note that due to the structure of $\bar{J}$ the values of $\xi_{i,k}^*$ only depend on properties of the risk metrics under consideration which are captured in D_i, and *not* on the payoff structure R_i.

The following proposition relates the Nash equilibrium of the $2n$-player convex game to the RQE of the action-dependent risk averse game.

Proposition 2. *If (π^*, p^*) is a Nash equilibrium of the $2n$-player game, then π^* is a RQE of Game (6). Furthermore, if π^* is a RQE of the action-dependent risk averse game, then (π^*, p^*) is a Nash equilibrium of the $2n$-player game, where:*

$$p_{i,j}^* = \arg \min_{p_{i,j} \in \mathcal{P}_{-i}} \bar{J}_{i,j}(\pi^*, p_{i,j}) \tag{30}$$

The proof of this result follows by exactly the same arguments as the proof of Proposition 1 and is therefore omitted for brevity.

Given these results we now prove the equivalent of Theorem 9 for the action-dependent formulation. We note that by a more careful accounting of terms, a stronger guarantee is possible in the 2-player regime similar to Theorem 3.

Theorem 10. *Assume the penalty functions that give rise to the players' risk preferences $D_i(\cdot, \cdot)$ are jointly convex in both their arguments. If σ is a CCE of the n-player game and for each i:*

$$\epsilon_i \geq \sum_j \xi_{i,j}^*$$

then the marginal strategies $\hat{\pi}_i = \mathbb{E}_\sigma[\pi_i]$ constitute a RQE of the action-dependent risk-averse game.

Note that one point of departure for this result from the previous results is that σ is now the CCE of the original 2-player convex game defined on the objective functions $f_1^{\epsilon_1}$ and $f_2^{\epsilon_2}$.

Proof. To begin, let σ be a CCE of the convex game played on $f_i^{\epsilon_i}$, where:

$$f_i^{\epsilon_i}(\pi_i, \pi_{-i}) = \sum_{j \in \mathcal{A}_i} \pi_i(j) \left(\sup_{p_{i,j} \in \mathcal{P}_{-i}} -\langle R_{i,j}, p_{i,j} \rangle - D_i(p_{i,j}, \pi_{-i}) \right) + \epsilon_i \nu(\pi_i).$$

By definition of a CCE, we must have that:

$$\mathbb{E}_\sigma[f_i^{\epsilon_i}(\pi_i, \pi_{-i})] \leq \mathbb{E}_\sigma[f_i^{\epsilon_i}(\pi_i', \pi_{-i})] \;\; \forall \, \pi_i' \in \Delta_{A_i},$$

for all $i = 1, ..., n$. Given σ, we can define a new distribution σ' as $(\pi, p^*(\pi))$ where

$$p_i^*(\pi) = \arg \min_{p_i \in \mathcal{P}_{-i}} \bar{J}_i(\pi, p_i),$$

with $\xi_{i,j} = \xi_{i,j}^*$. We now claim that, by construction, σ' is a CCE of the $2n$-player game. To see this, we observe that

$$\mathbb{E}_{\sigma'}[J_i(\pi_i, \pi_{-i}, p)] = \mathbb{E}_\sigma[f_i^{\epsilon_i}(\pi_i, \pi_{-i})] \leq \mathbb{E}_\sigma[f_i^{\epsilon_i}(\pi_i', \pi_{-i})] = \mathbb{E}_{\sigma'}[J_i(\pi_i', \pi_{-i}, p)] \;\; \forall \, \pi_i \in \Delta_{A_i},$$

where the first equality follows by construction of σ', the second from the definition of a CCE, and the third from the fact that the value of $p^*(\pi)$ only depends on π_{-i} and not π_i'.

Similarly, we can show the same result for all the players p_i. By simply applying Jensen's inequality, we can see that

$$\mathbb{E}_{\sigma'}[\bar{J}_i(\pi, p_i)] = \mathbb{E}_{\pi \sim \sigma} \left[\min_{p_{i,j}} \bar{J}_i(\pi, p_{i,j}) \right] \leq \mathbb{E}_{\sigma'}[\bar{J}_i(\pi, p_i')] \;\; \forall p_i' \in \mathcal{P}_{-i}.$$

Thus we can observe that σ' is a CCE of the $2n$-player game. To show that the marginals of π in this CCE are Nash equilibria of the convex game (and thus RQE), we proceed as before. Using the fact that $D(p_{i,j}, \pi_{-i})$ is jointly convex in each of its arguments, we can apply Jensen's inequality to find that:

$$\mathbb{E}_{\sigma'}[J_i(\pi_i, \pi_{-i}, p)] \leq J_i(\pi_i', \hat{\pi}_{-i}, \hat{p}) \;\; \forall \pi_i' \in \Delta_{A_i},$$

where $\hat{p}_{i,j} = \mathbb{E}_{\sigma'}[p_{i,j}]$ and $\hat{\pi}_i = \mathbb{E}_{\sigma'}[\pi_i]$. Similarly, by our choice of $\xi_{i,j} = \xi_{i,j}^*$ we can find that:

$$\mathbb{E}_{\sigma'}[\bar{J}_i(\pi, p_i)] \leq \bar{J}_i(\hat{\pi}, p_i') \;\; \forall p_i' \in \mathcal{P}_{-i}.$$

Now, we can observe that:

$$\begin{aligned}
\sum_i J_i(\hat{\pi}_i, \hat{\pi}_{-i}, \hat{p}) + \bar{J}_i(\hat{\pi}, \hat{p}_i) &\geq \mathbb{E}_{\sigma'} \left[\sum_i J_i(\pi_i, \pi_{-i}, p) + \bar{J}_i(\pi, p_i) \right] \\
&= \sum_i \mathbb{E}_{\sigma'} \left[\left(\epsilon_i - \sum_j \xi_{i,j}^* \right) \nu_i(\pi_i) \right] \\
&\geq \sum_i \left(\epsilon_i - \sum_j \xi_{i,j}^* \right) \nu_i(\hat{\pi}_i) \\
&= \sum_i J_i(\hat{\pi}_i, \hat{\pi}_{-i}, \hat{p}) + \bar{J}_i(\hat{\pi}, \hat{p}_i),
\end{aligned}$$

where the third inequality follows by the Jensen's inequality. By the same rationale as in the proof of Theorem 9 we can now conclude that $\hat{\pi}$ is a Nash equilibrium-joint strategy profile-of the $2n$ player game and thus (due to Proposition 2) an RQE of the action-dependent risk averse game. □

Similar to the aggregate risk regime, one can instantiate the previous theorem with specific risk-metrics and quantal response functions to illustrate the class of action-dependent RQE that are guaranteed to be computationally tractable. Importantly, this class once again only depends on the levels of risk averse and quantal response but *not* on the underlying structure of the game. However, note that the requirements are more stringent since the requirements on the ξ^*'s are stronger since they must ensure joint convexity of $\bar{J}_i$. Nevertheless further algebraic manipulations would allow one to recover analogues of Corollaries 8.1 and 8.2 for the action-dependent risk case as well. For brevity we leave these as exercises to the reader.

E PROOF OF THEOREM 5

Armed with the estimated reward and transition kernel in (15), we can construct an empirical MG $\widehat{\mathcal{MG}} = \{H, \mathcal{S}, \{\mathcal{A}_i\}_{i\in[n]}, \{\widehat{R}_{i,h}, \widehat{P}_{i,h}\}_{i\in[n],h\in[H]}\}$. Analogously, for any joint policy π, we denote the corresponding risk-averse loss functions or the payoff matrices as $\{\widehat{V}^{\epsilon_i}_{i,h}(\pi)\}$ and $\{\widehat{Q}^{\epsilon_i}_{i,h}(\pi)\}$, respectively.

To begin with, recall the goal is to show that

$$\max_{(i,s,h)\in[n]\times\mathcal{S}\times[H]} \left\{ V^{\epsilon_i}_{i,h}(\widehat{\pi}; s) - \min_{\pi'_h:\mathcal{S}\mapsto\Delta_{\mathcal{A}_i}} V^{\epsilon_i}_{i,h}((\pi'_h, \widehat{\pi}_{i,-h}) \times \widehat{\pi}_{-i}; s) \right\} \le \delta. \tag{31}$$

For convenience, we denote

$$\forall (i,h,s) \in [n]\times[H]\times\mathcal{S}: \quad V^{\star,\epsilon_i}_{i,h}(\widehat{\pi}_{-i}; s) = \min_{\pi'_{i,h}:\mathcal{S}\mapsto\Delta_{A_i}} V^{\epsilon_i}_{i,h}((\pi'_{i,h}, \widehat{\pi}_{i,-h}) \times \widehat{\pi}_{-i}; s) \tag{32}$$

and define the best-response policy $\pi^\star_i := \{\pi^\star_{i,h} : \mathcal{S}\mapsto\Delta_{A_i}\}_{h\in[H]}$ so that

$$\forall (i,h,s) \in [n]\times[H]\times\mathcal{S}: \quad \pi^\star_{i,h}(s) := \operatorname{argmin}_{\pi'_i:\mathcal{S}\mapsto\Delta_{A_i}} V^{\epsilon_i}_{i,h}((\pi'_{i,h}, \widehat{\pi}_{i,-h}) \times \widehat{\pi}_{-i}; s). \tag{33}$$

Then, for any $i \in [n]$, the gap $V^{\epsilon_i}_{i,1}(\widehat{\pi}) - V^{\star,\epsilon_i}_{i,1}(\widehat{\pi}_{-i})$ can be decomposed as follows:

$$\begin{aligned}
&V^{\epsilon_i}_{i,h}(\widehat{\pi}) - V^{\star,\epsilon_i}_{i,h}(\widehat{\pi}_{-i}) \\
&= V^{\epsilon_i}_{i,h}(\widehat{\pi}) - \widehat{V}^{\epsilon_i}_{i,h}(\widehat{\pi}) + \left(\widehat{V}^{\epsilon_i}_{i,h}(\widehat{\pi}) - \widehat{V}^{\epsilon_i}_{i,h}((\pi^\star_{i,h}, \widehat{\pi}_{i,-h}) \times \widehat{\pi}_{-i})\right) \\
&\quad + \left(\widehat{V}^{\epsilon_i}_{i,h}((\pi^\star_{i,h}, \widehat{\pi}_{i,-h}) \times \widehat{\pi}_{-i}) - V^{\epsilon_i}_{i,h}((\pi^\star_{i,h}, \widehat{\pi}_{i,-h}) \times \widehat{\pi}_{-i})\right) \\
&\le V^{\epsilon_i}_{i,h}(\widehat{\pi}) - \widehat{V}^{\epsilon_i}_{i,h}(\widehat{\pi}) + \left(\widehat{V}^{\epsilon_i}_{i,h}((\pi^\star_{i,h}, \widehat{\pi}_{i,-h}) \times \widehat{\pi}_{-i}) - V^{\epsilon_i}_{i,h}((\pi^\star_{i,h}, \widehat{\pi}_{i,-h}) \times \widehat{\pi}_{-i})\right) \\
&\le \|V^{\epsilon_i}_{i,h}(\widehat{\pi}) - \widehat{V}^{\epsilon_i}_{i,h}(\widehat{\pi})\|_\infty \mathbf{1} + \left\|\widehat{V}^{\epsilon_i}_{i,h}((\pi^\star_{i,h}, \widehat{\pi}_{i,-h}) \times \widehat{\pi}_{-i}) - V^{\epsilon_i}_{i,h}((\pi^\star_{i,h}, \widehat{\pi}_{i,-h}) \times \widehat{\pi}_{-i})\right\|_\infty \mathbf{1}
\end{aligned} \tag{34}$$

where the first inequality holds by applying Theorem 4 with the estimated RAMG $\widehat{\mathcal{MG}}$ so that $\widehat{\pi}$ is a RQE of $\widehat{\mathcal{MG}}$, i.e.,

$$\widehat{V}^{\epsilon_i}_{i,h}(\widehat{\pi}) \le \min_{\pi'_{i,h}:\mathcal{S}\mapsto\Delta_{A_i}} \widehat{V}^{\epsilon_i}_{i,h}((\pi'_{i,h}, \widehat{\pi}_{i,-h}) \times \widehat{\pi}_{-i}) \le \widehat{V}^{\epsilon_i}_{i,h}((\pi^\star_{i,h}, \widehat{\pi}_{i,-h}) \times \widehat{\pi}_{-i}). \tag{35}$$

To continue, we divide the proof into several key steps.

Step 1: developing the recursion. To control the two terms in (34), we consider that for any joint policy π and time step $(s,h) \in \mathcal{S}\times[H]$,

$$\begin{aligned}
&V^{\epsilon_i}_{i,h}(\pi; s) - \widehat{V}^{\epsilon_i}_{i,h}(\pi; s) \\
&\overset{\text{(i)}}{=} g^\pi_{\mathsf{pol},i}\left(Q_{i,h}(\pi; s, :)\right) - g^\pi_{\mathsf{pol},i}\left(\widehat{Q}_{i,h}(\pi; s, :)\right) \\
&\overset{\text{(ii)}}{=} \sup_{p_i\in\mathcal{P}_i} -\pi_i(s)^T g_{\mathsf{env},i}\left(R_{i,h}(s,:), P_{h,s,:}, V_{i,h+1}(\pi)\right) p_i - D_{\mathsf{pol},i}(p_i, \pi_{-i}(s)) \\
&\quad - \left[\sup_{p_i\in\mathcal{P}_i} -\pi_i(s)^T g_{\mathsf{env},i}\left(R_{i,h}(s,:), \widehat{P}_{h,s,:}, \widehat{V}_{i,h+1}(\pi)\right) p_i - D_{\mathsf{pol},i}(p_i, \pi_{-i}(s))\right] \\
&\le \sup_{p_i\in\mathcal{P}_i} \left|-\pi_i(s)^T \left[g_{\mathsf{env},i}\left(R_{i,h}(s,:), P_{h,s,:}, V_{i,h+1}(\pi)\right) - g_{\mathsf{env},i}\left(R_{i,h}(s,:), \widehat{P}_{h,s,:}, \widehat{V}_{i,h+1}(\pi)\right)\right] p_i\right| \\
&\le \left\|g_{\mathsf{env},i}\left(R_{i,h}(s,:), P_{h,s,:}, V_{i,h+1}(\pi)\right) - g_{\mathsf{env},i}\left(R_{i,h}(s,:), \widehat{P}_{h,s,:}, \widehat{V}_{i,h+1}(\pi)\right)\right\|_\infty,
\end{aligned} \tag{36}$$

where (i) holds by the definition in (11) and (10), and (ii) follows from the definition of $g^\pi_{\mathsf{pol},i}(\cdot)$ in (9).

To continue, we know that for any $(s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}$:

$$
\begin{aligned}
&\left|g_{\mathsf{env},i}\left(R_{i,h}(s,\boldsymbol{a}), P_{h,s,\boldsymbol{a}}, V_{i,h+1}(\pi)\right) - g_{\mathsf{env},i}\left(R_{i,h}(s,:), \widehat{P}_{h,s,\boldsymbol{a}}, \widehat{V}_{i,h+1}(\pi)\right)\right| \\
&= \Bigg|R_{i,h}(s,\boldsymbol{a}) - \sup_{\widetilde{P}\in\Delta_S} -\widetilde{P}V_{i,h+1}(\pi) - D_{\mathsf{env},i}(\widetilde{P}, P_{h,s,\boldsymbol{a}}) \\
&\quad - \left(R_{i,h}(s,\boldsymbol{a}) - \sup_{\widetilde{P}\in\Delta_S} -\widetilde{P}\widehat{V}_{i,h+1}(\pi) - D_{\mathsf{env},i}(\widetilde{P}, \widehat{P}_{h,s,\boldsymbol{a}})\right)\Bigg| \\
&= \left|-\sup_{\widetilde{P}\in\Delta_S}\left(-\widetilde{P}V_{i,h+1}(\pi) - D_{\mathsf{env},i}(\widetilde{P}, P_{h,s,\boldsymbol{a}})\right) + \sup_{\widetilde{P}\in\Delta_S}\left(-\widetilde{P}\widehat{V}_{i,h+1}(\pi) - D_{\mathsf{env},i}(\widetilde{P}, \widehat{P}_{h,s,\boldsymbol{a}})\right)\right| \\
&\overset{\text{(i)}}{\leq} \sup_{\widetilde{P}\in\Delta_S}\left|-\left(-\widetilde{P}V_{i,h+1}(\pi) - D_{\mathsf{env},i}(\widetilde{P}, P_{h,s,\boldsymbol{a}})\right) + \left(-\widetilde{P}\widehat{V}_{i,h+1}(\pi) - D_{\mathsf{env},i}(\widetilde{P}, \widehat{P}_{h,s,\boldsymbol{a}})\right)\right| \\
&\overset{\text{(ii)}}{=} \sup_{\widetilde{P}\in\Delta_S}\left|\widetilde{P}\left(V_{i,h+1}(\pi) - \widehat{V}_{i,h+1}(\pi)\right)\right| + \sup_{\widetilde{P}\in\Delta_S}\left|D_{\mathsf{env},i}(\widetilde{P}, P_{h,s,\boldsymbol{a}}) - D_{\mathsf{env},i}(\widetilde{P}, \widehat{P}_{h,s,\boldsymbol{a}})\right| \\
&\leq \left\|V_{i,h+1}(\pi) - \widehat{V}_{i,h+1}(\pi)\right\|_\infty + L\left\|P_{h,s,\boldsymbol{a}} - \widehat{P}_{h,s,\boldsymbol{a}}\right\|_1. \qquad (37)
\end{aligned}
$$

where the first equality holds by (13), (i) holds by the supreme operator is 1-Lipschitz, and the last inequality holds by the assumption that $D_{\mathsf{env},i}(\cdot)$ is L-Lipschitz with respect to the ℓ_1 norm for the second argument, with any fixed first argument. We mention an important note here. As remarked earlier, similar sample complexity guarantees can be shown when $\{D_{\mathsf{env},i}\}_{i\in[n]}$ are defined as ϕ-divergence and allude to analyses ideas in (Panaganti and Kalathil, 2022; Xu* et al., 2023; Shi et al., 2023) to modify this step.

Plugging in (37) back to (36) and applying the results for all $s \in \mathcal{S}$ yields

$$
\begin{aligned}
\left\|V^{\epsilon_i}_{i,h}(\pi) - \widehat{V}^{\epsilon_i}_{i,h}(\pi)\right\|_\infty &\leq \left\|V_{i,h+1}(\pi) - \widehat{V}_{i,h+1}(\pi)\right\|_\infty + L\underbrace{\max_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}}\left\|P_{h,s,\boldsymbol{a}} - \widehat{P}_{h,s,\boldsymbol{a}}\right\|_1}_{=:l_h} \\
&= \left\|V^{\epsilon_i}_{i,h+1}(\pi) - \widehat{V}^{\epsilon_i}_{i,h+1}(\pi)\right\|_\infty + Ll_h. \qquad (38)
\end{aligned}
$$

Applying above fact recursively for $h, h+1, \cdots, H$, we arrive at

$$
\begin{aligned}
\left\|V^{\epsilon_i}_{i,h}(\pi) - \widehat{V}^{\epsilon_i}_{i,h}(\pi)\right\|_\infty &\leq \left\|V^{\epsilon_i}_{i,h+2}(\pi) - \widehat{V}^{\epsilon_i}_{i,h+2}(\pi)\right\|_\infty + Ll_h + Ll_{h+1} \\
&\leq \cdots \leq \left\|V^{\epsilon_i}_{i,H+1}(\pi) - \widehat{V}^{\epsilon_i}_{i,H+1}(\pi)\right\|_\infty + L\sum_{t=h}^{H} l_t \\
&\leq L\sum_{t=h}^{H} l_t, \qquad (39)
\end{aligned}
$$

where the last inequality holds since $V^{\epsilon_i}_{i,H+1}(\pi) = \widehat{V}^{\epsilon_i}_{i,H+1}(\pi) = V_{i,h+1}(\pi) = \widehat{V}_{i,h+1}(\pi) = 0$ for any policy π.

Step 2: controlling the errors $\{l_t\}$. The remainder of the proof will focus on controlling (39). Applying (Auer et al., 2008, Lemma 17) over all $(h, s, \boldsymbol{a}) \in [H] \times \mathcal{S} \times \mathcal{A}$, we achieve the union bound that with probability at least $1 - \delta$,

$$
\max_{(h,s,\boldsymbol{a})\in[H]\times\mathcal{S}\times\mathcal{A}}\left\|P_{h,s,\boldsymbol{a}} - \widehat{P}_{h,s,\boldsymbol{a}}\right\|_1 \leq \sqrt{\frac{14S}{N}\log\left(\frac{2S\prod_{i\in[n]}A_iH}{\delta}\right)}. \qquad (40)
$$

Applying (40), we arrive at with probability at least $1 - \delta$,

$$
\forall h \in [H],\ l_h = \max_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}}\left\|P_{h,s,\boldsymbol{a}} - \widehat{P}_{h,s,\boldsymbol{a}}\right\|_1 \leq \sqrt{\frac{14S}{N}\log\left(\frac{2S\prod_{i\in[n]}A_iH}{\delta}\right)}. \qquad (41)
$$

Inserting (41) into (39) gives

$$\left\|V_{i,h}^{\epsilon_i}(\pi) - \widehat{V}_{i,h}^{\epsilon_i}(\pi)\right\|_\infty \leq L\sum_{t=h}^{H} l_t \leq HL\sqrt{\frac{14S}{N}\log\left(\frac{2S\prod_{i\in[n]}A_iH}{\delta}\right)}. \tag{42}$$

Finally, recalling (34) yields

$$\begin{aligned} V_{i,1}^{\epsilon_i}(\widehat{\pi}) - V_{i,1}^{\star,\epsilon_i}(\widehat{\pi}_{-i}) &\leq \|V_{i,h}^{\epsilon_i}(\widehat{\pi}) - \widehat{V}_{i,h}^{\epsilon_i}(\widehat{\pi})\|_\infty \mathbf{1} \qquad (43)\\ &\quad + \left\|\widehat{V}_{i,h}^{\epsilon_i}((\pi_{i,h}^\star, \widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i}) - V_{i,h}^{\epsilon_i}((\pi_{i,h}^\star, \widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i})\right\|_\infty \mathbf{1} \\ &\leq 8HL\sqrt{\frac{S}{N}\log\left(\frac{2S\prod_{i\in[n]}A_iH}{\delta}\right)}\mathbf{1}, \qquad (44)\end{aligned}$$

where the last inequality holds by applying (42) with $\pi = \widehat{\pi}$ or $\pi = (\pi_{i,h}^\star, \widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i}$. □

F Experiment Details

We provide more results and details of our experiments in this section. We provide our code in the following **Github link** https://github.com/kishanpb/Risk-averse-Quantal-Equilibria that contains instructions to reproduce all results in this paper.

F.1 Matrix Games

Game 1			**Game 2**			**Game 3**		
	L	R		L	R		L	R
U	10 10	0 18	U	9 4	0 13	U	8 6	0 14
D	9 9	10 8	D	6 7	8 5	D	7 7	10 4
Game 4			**Game 5**			**Game 6**		
	L	R		L	R		L	R
U	7 4	0 11	U	7 2	0 9	U	7 1	1 7
D	5 6	9 2	D	4 5	8 1	D	3 5	8 0
Game 7			**Game 8**			**Game 9**		
	L	R		L	R		L	R
U	10 12	4 22	U	9 7	3 16	U	8 9	3 17
D	9 9	14 8	D	6 7	11 5	D	7 7	13 4
Game 10			**Game 11**			**Game 12**		
	L	R		L	R		L	R
U	7 6	2 13	U	7 4	2 11	U	7 3	3 9
D	5 6	11 2	D	4 5	10 1	D	3 5	10 0

Table 2: Payoff Matrices from (Selten and Chmura, 2008). In each column, the number above (below) is the payoff for player 1 (player 2).

	L	R
U	200 160	160 10
D	370 200	10 370

Table 3: Game 4 Payoff Matrix from (Goeree et al., 2003). In each column, the number above (below) is the payoff for player 1 (player 2).

Matching pennies matrix games: Two players in matching pennies simultaneously choose heads or tails, and a player wins (the other player loses) a payoff if their choices *match*. The detailed payoff matrices are provided in Tables 2 and 3. These games from (Goeree et al., 2003; Selten and Chmura, 2008) are focused on showcasing risk-averse solutions by developing payoffs strategically such that either player deviating from their choice of plays will cause hefty damage in terms of payoff to the other player. So, both players prefer a 'safer' choice of plays, thus highlighting both risk-averse and bounded rational preferences.

Algorithm and Result details: We use the penalty functions KL and reverse KL with the regularizers log barrier and negative entropy, as mentioned in Table 1, for our experiments. To be precise, we consider the following two experimental setups: (a) For $j = 1, 2$, we let $\nu_j(p) = -\sum_i \log(p_i)$ and $D_j(p, q) = (1/\tau_j)\mathrm{KL}(p, q)$. (b) For $j = 1, 2$, we let $\nu_j(p) = \sum_i p_i \log(p_i)$ and $D_j(p, q) = (1/\tau_j)\mathrm{KL}(q, p)$. We note that the parameters τ_j^{-1} play the same role as ξ_j's (formalized in Corollaries 8.1 and 8.2), i.e., the players take more risk neutral decisions as $\tau_j \to 0$. With the perfect information, we use the vanilla projected gradient descent with constant stepsizes (Beck, 2017) as the no-regret algorithm to arrive at the CCE σ of the four player game. We use consistent stepsizes between 10^{-4} and 10^{-3} for all our runs iterating through 10^4 steps of gradient descent. We notice 2% deviations in sup-norm for the algorithm policies in about 20 runs.

F.2 Markov Games

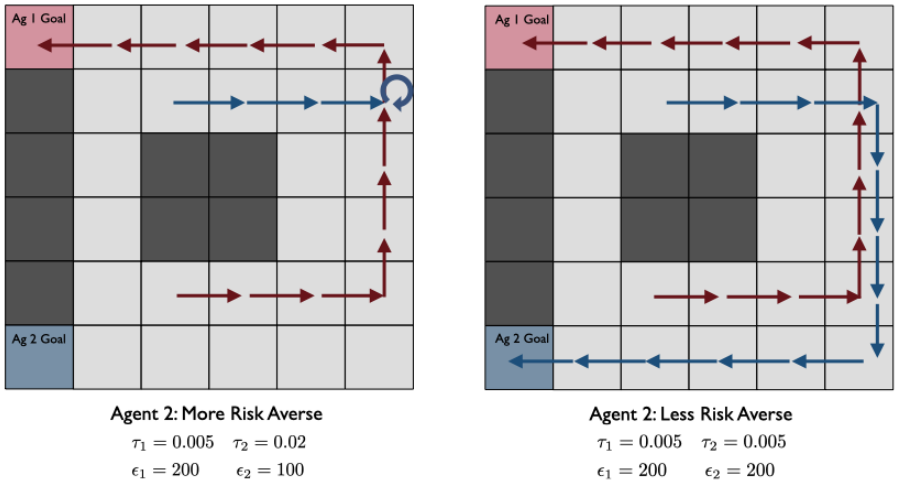

Figure 3: Cliff-Walk results for the ℓ_1 environmental uncertainty metric.

Algorithm details: We use the penalty function KL with the log-barrier regularizer, as mentioned in Table 1 and Example 2, for our experiments. We also use KL divergence for the environmental uncertainty metric $D_{\mathsf{env},i}$ with the same bounded rational parameter τ_i^{-1} used by RQE. We use the full information of this grid-world to evaluate the Q-values described in Algorithm 1. Additionally, as we use the matrix games solver for RQE, we have similar statistical deviations for different training runs. In Appendix F.2, we also showcase environmental uncertainty results with ℓ_1 metric.

Here, we showcase results corresponding to ℓ_1 metric for the environmental uncertainty metric $D_{\mathsf{env},i}$ for the grid-world problem. We consider the same *Cliff-Walk* environment described in Section 4.3 under some minor modifications discussed here. The rewards for falling into the cliff is now -100 and for reaching the goal is 20. Agents/players are rewarded a small negative reward of -0.1 for taking each step. The episode horizon H is 100. The algorithm details is the same as described in Section 4.3.

We present two results in Fig. 3 that has similar implications as in Fig. 2. We make one important observation about these agents equipped with ℓ_1 environmental uncertainty metric. These agents are less risk-averse towards the cliff in horizontal grid axis compared to the results Fig. 2 corresponding to the KL environmental uncertainty metric. We do not investigate the effects of different environmental uncertainty metrics in this work and postpone to address in future research directions.