A Behavioral Economics Approach to Principled Multi-Agent Reinforcement Learning

Eric Mazumdar[*†]
Caltech

Kishan Panaganti[*‡]
Caltech

Laixi Shi[*‡]
Caltech

Abstract

A significant roadblock to the development of principled multi-agent reinforcement learning is the fact that desired solution concepts like Nash equilibria may be intractable to compute. To overcome this obstacle, we take inspiration from behavioral economics and show that—by imbuing agents with important features of human decision-making like risk aversion and bounded rationality—a class of risk-averse quantal response equilibria (RQE) become 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' degree of risk-aversion and bounded rationality. To validate the richness 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 and validate our findings on a simple multi-agent reinforcement learning benchmark.

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 which may not be aligned with those of the algorithm—making the interactions *strategic*. Such situations are naturally modeled as *games* between rational agents and their prevalence in a wide range of application areas has driven a surge in research interest in learning in games [1] and multi-agent reinforcement learning [2] in recent years. Indeed, real-world applications of these problems range from the training of large language models [3] to the decentralized control of the smart grid [4], autonomous driving [5], and financial trading [6] among many others.

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) [7]: 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 [8] even for two-player matrix games.

[*] Author list is in alphabetical order.

[†] Department of Computing Mathematical Sciences and Department of Economics, California Institute of Technology, CA 91125, USA.

[‡] Department of Computing Mathematical Sciences, California Institute of Technology, CA 91125, USA.

NeurIPS 2024 Workshop on Behavioral Machine Learning.

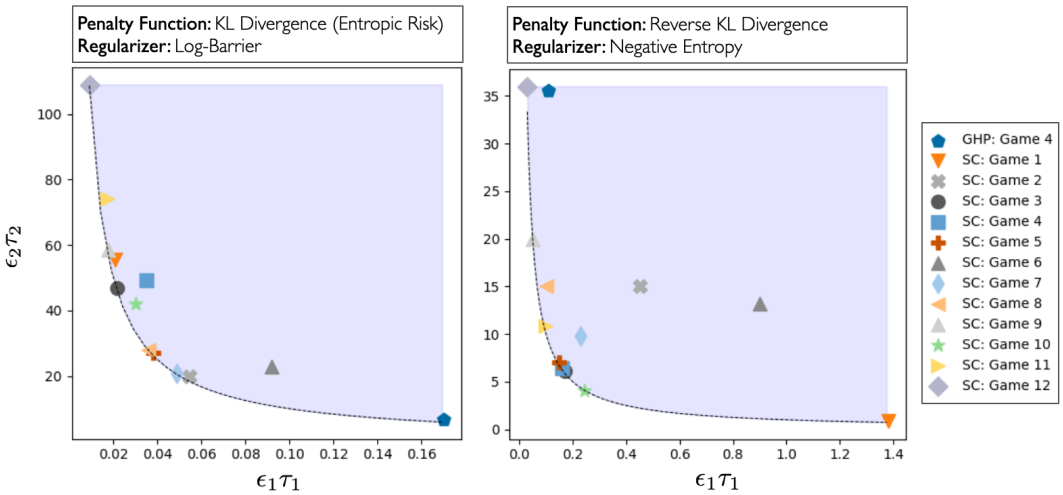

Figure 1: The shaded blue region depicts the regime of risk-aversion (τ_1 and τ_2) and bounded rationality (ϵ_1 and ϵ_2) preferences that allow for computationally tractable RQE in all 2-player games as shown in Theorem 2. The markers *GHP: Game 4* [16] and *SC: Game 1-12* [22] represent the necessary parameter values required to recreate the average strategy (up to 1% accuracy) played by people in various 2-player games in observational data.

While relaxations of NE like (coarse) correlated equilibria (CCE) are known to be more tractable to compute, they also have their limitations. Indeed while CCE can be computed through the use of no-regret learning algorithms [1], the set of CCE can be large (introducing an additional problem of equilibrium selection) and may include strictly dominated strategies [9] which means that they cannot necessarily be rationalized by individual agents [10]. Furthermore, a refinement of CCE—stationary CCE— are also known to be PPAD-hard to compute in dynamic general-sum games [11].

Beyond these hardness results, solution concepts like NE and CCE also fail to be predictive of what strategies people play in games [12,13], with people being observed to be imperfect optimizers [14,15] and risk-averse [16] 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 [17] and risk-aversion [18, 19].

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 [17]. This is often captured in games through the idea of a *quantal response equilibrium* (QRE) [12].

The second observation can be attributed to the fact that players typically face uncertainty and risk in their decisions, stemming from environmental uncertainties like unknown future events, noise, or even other players' lack of optimality. The presence of these uncertainties can lead people to prefer risk-averse strategies, i.e., strategies which give more certain outcomes at the cost of lower expected returns [20]. Interestingly, there is evidence that neither of these properties alone can account for people's patterns of play observed in controlled experiments [16, 21], and that models of decision-making that incorporate *both* of these features have the best predictive power [16].

Motivated by these findings, in this paper we formulate and study games in which agents are risk-averse and have bounded rationality and study the computational tractability of the natural equilibrium concept in this class of games: a risk-averse quantal response equilibrium (RQE). At first glance, the introduction of these features of human decision-making into games can break existing game theoretic structures. However, by relying on dual formulations of risk we show how, for a large range of degrees of risk-aversion and bounded rationality, RQE are computationally tractable in *all* n-player matrix games. Crucially, the class of computationally tractable RQE is independent of the original game structure and depends only on players' degrees of risk aversion and bounded rationality. We then extend the result to finite-horizon Markov games in which agents interact in a Markov decision process (MDP) and provide a finite-sample complexity guarantee on how many samples one would need to approximate a RQE in Markov games.

These results suggest that designing multi-agent reinforcement learning (MARL) agents with realistic features of human decision-making can allow for a computationally tractable equilibrium concept in all finite action games and finite horizon Markov games. To emphasize the practical relevance of this theoretical result, we show how that 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).

2 Main Contributions and Results

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 [8], 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 [23, 24], monotone games [25], smooth games [26], or socially concave games [27].

In games without such structure however, the natural targets for computation and learning became correlated [28] and coarse correlated equilibria [29,30] (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 [1], and may have support on dominated strategies [9]. Furthermore, in the dynamic game context of Markov games, stationary CE and CCE are also computationally intractable to compute [11].

More recently, a new equilibrium concept—a smoothed Nash equilibrium— has been proposed as an alternative to these other equilibrium concepts [31] 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 [32]. 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. In particular, we make use of a general class of convex risk measures and present two forms of risk aversion: *aggregate risk-aversion* and *action-dependent risk aversion*—with the latter yielding less conservative solutions at the cost of a more complex formulation and analysis. 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 informally state this result here.

Theorem 1 (Formal statement in Theorem 4)**.** *A risk-averse Nash equilibrium (RNE) is a joint strategy such that no player has any incentive to unilaterally deviate in the risk averse game. There always exists an RNE for all aggregate and action-dependent risk-averse games.*

Proof Sketch. Both risk-averse games we consider are constructed so as to satisfy the conditions of a convex game [33]. This nontrivial construction enables the existence of Nash equilibria [33]. □

We note that such formulations of risk-aversion are common in the literature on risk-sensitive control [34, 35] and risk-sensitive reinforcement learning [36] where agents are implicitly presumed to be risk-averse only to the randomness that is outside their control (i.e., the environment).

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., [12,13,37–39]). 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 [16,21,40]. 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 [17]. 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 [12,41,42]. 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 [42–47] 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. We state this result informally that crucially depends on different forms of risk metrics and quantal responses.

Theorem 2 (Formal statement in Theorem 5)**.** *Suppose the players are risk-averse using some risk metric (for e.g., entropic risk, also see Table 1) with parameters τ_1 and τ_2 respectively. If they respond in the space of some associated quantal responses (for e.g., Kullback-Leibler divergence, also see Table 1) with parameters ϵ_1 and ϵ_2 respectively, then the players can compute a risk-averse QRE by using no-regret learning when the parameters satisfy*

$$\epsilon_1 \tau_1 \geq \frac{1}{\epsilon_2 \tau_2}.$$

The blue region in Fig. 1 supports this result for the condition on the computability of risk-averse QRE. 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 behavioral economics [16,21].

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 [48], and more recently in risk-sensitive reinforcement learning [36]. 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 [35]. Due to classic duality results (see e.g., [49]) this line of work is closely related to the literature on robust control and distributionally robust reinforcement learning [50–53] 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., [54–57]). Several provide rigorous definitions of risk-averse equilibria and some guarantees on their computation by assuming structure on the risk-adjusted game. Oftentimes this is done by assuming that the risk-adjusted game is itself zero-sum [54], monotone [55], or that it satisfies other strong conditions [56]. Other works are more empirical in nature [57–62], showing the promise of risk-averse algorithms for MARL.

One last closely related line of work is the emerging literature on robust multi-agent reinforcement learning [63–66]. 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 [67].

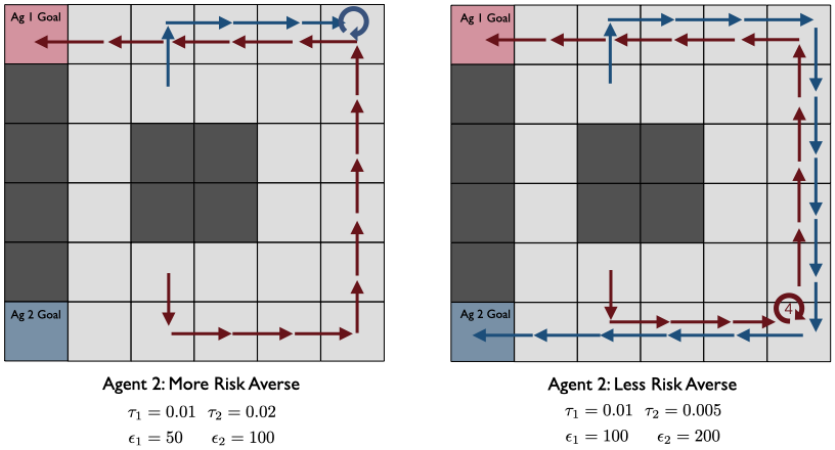

Figure 2: *Cliff Walk Description and Results*: The *Cliff Walk* grid-world is depicted here with color codes of grids: black is the cliff, blue and pink are agents/players 1 and 2's goals respectively. The agents' proximity to one another increases the randomness in the dynamics making falling into 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.

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. Indeed we showcase the effects of risk-aversion and bounded rationality in a simple Markov game mentioned below.

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.

3 Conclusion

In Section B, by incorporating risk aversion and bounded rationality into agents' decision-making processes, we introduced a computationally tractable class of risk-averse quantal response equilibria (RQE) for matrix games and finite-horizon Markov games. This approach not only makes equilibrium computation tractable but also aligns well with observed human behavior in behavioral economics. In Section C, we provided rigorous theoretical results for tractable computation of equilibrium concepts and also provided sample complexity results in the generative modeling setting for multi-agent reinforcement learning. In Sections D, we also expanded our findings on small-scale multi-agent game benchmarks: *matching pennies games* and *cliff walking environment*.

References

[1] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006. 1, 2, 3

[2] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021. 1

[3] Remi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback, 2023. 1

[4] Amir-Hamed Mohsenian-Rad, Vincent WS Wong, Juri Jatskevich, Robert Schober, and Alberto 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

[5] Sampath Kannan, Michael Kearns, Jamie Morgenstern, Mallesh Pai, Aaron Roth, Rakesh Vohra, and Zhiwei Steven Wu. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 369–386, 2017. 1

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

[7] John F Nash. Non-cooperative games. *Princeton University*, 1950. 1

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

[9] Yannick Viossat and Andriy Zapechelnyuk. No-regret dynamics and fictitious play. *Journal of Economic Theory*, 148(2):825–842, 2013. 2, 3

[10] Eddie Dekel and Drew Fudenberg. Rational behavior with payoff uncertainty. *Journal of Economic Theory*, 52(2):243–267, 1990. 2

[11] Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. The complexity of markov equilibrium in stochastic games. In Gergely Neu and Lorenzo 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 2023. 2, 3

[12] Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995. 2, 4

[13] Ido Erev and Alvin 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, 4

[14] Jacob K. Goeree and Charles 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

[15] C. Monica Capra, Jacob K. Goeree, Rosario Gomez, and Charles A. Holt. Learning and noisy equilibrium behavior in an experimental study of imperfect price competition. *International Economic Review*, 43(3):613–636, 2002. 2

[16] Jacob K Goeree, Charles A Holt, and Thomas R Palfrey. Risk averse behavior in generalized matching pennies games. *Games and Economic Behavior*, 45(1):97–113, 2003. 2, 4, 13, 15, 21

[17] R. Duncan Luce. *Individual Choice Behavior: A Theoretical analysis*. Wiley, New York, NY, USA, 1959. 2, 4

[18] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. 2

[19] Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5:297–323, 1992. 2

[20] Christian Gollier. *The economics of risk and time*. MIT press, 2001. 2

[21] K. Goeree and Theo Offerman. Efficiency in auctions with private and common values: An experimental study. *American Economic Review*, 92(3):625–643, June 2002. 2, 4

[22] Reinhard Selten and Thorsten Chmura. Stationary concepts for experimental 2x2-games. *American Economic Review*, 98(3):938–966, 2008. 2, 15, 21

[23] Yang Cai, Ozan Candogan, Constantinos Daskalakis, and Christos Papadimitriou. Zero-sum polymatrix games: A generalization of minmax. *Mathematics of Operations Research*, 41(2):648–655, 2016. 3, 15

[24] Fivos Kalogiannis and Ioannis Panageas. Zero-sum polymatrix markov games: Equilibrium collapse and efficient computation of nash equilibria, 2023. 3, 15

[25] Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. *Advances in neural information processing systems*, 33:20766–20778, 2020. 3

[26] Tim Roughgarden. Intrinsic robustness of the price of anarchy. *Journal of the ACM (JACM)*, 62(5):1–42, 2015. 3

[27] Eyal Even-Dar, Yishay Mansour, and Uri 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. 3

[28] 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. 3

[29] Robert J. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1):67–96, 1974. 3

[30] Robert J. Aumann. Correlated equilibrium as an expression of bayesian rationality. *Econometrica*, 55(1):1–18, 1987. 3

[31] Constantinos Daskalakis, Noah Golowich, Nika Haghtalab, and Abhishek Shetty. Smooth nash equilibria: Algorithms and complexity. *arXiv preprint arXiv:2309.12226*, 2023. 3

[32] Amos Fiat and Christos 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. 3, 11, 13

[33] J. B. Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica*, 33(3):520–534, 1965. 3, 13, 23

[34] Yun Shen, Wilhelm Stannat, and Klaus Obermayer. Risk-sensitive markov control processes. *SIAM Journal on Control and Optimization*, 51(5):3652–3672, 2013. 3

[35] Vivek Borkar. Risk-sensitive control, single controller games and linear programming. *Journal of Dynamics and Games*, 11, 01 2023. 3, 4, 11

[36] Yun Shen, Michael J Tobia, Tobias Sommer, and Klaus Obermayer. Risk-sensitive reinforcement learning. *Neural computation*, 26(7):1298–1328, 2014. 3, 4, 11

[37] James N Brown and Robert Rosenthal. Testing the minimax hypothesis: A re-examination of o'neill's game experiment. *Econometrica*, 58(5):1065–81, 1990. 4

[38] Barry 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. 4

[39] Richard D. McKelvey and Thomas R. Palfrey. An experimental study of the centipede game. *Econometrica*, 60(4):803–836, 1992. 4

[40] Teck Ho, Colin Camerer, and Juin-Kuan Chong. A cognitive hierarchy model games. *The Quarterly Journal of Economics*, 119:861–898, 02 2004. 4

[41] Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for extensive form games. *Experimental economics*, 1:9–41, 1998. 4

[42] Panayotis Mertikopoulos and William H. Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016. 4, 14

[43] Samuel Sokota, Ryan D'Orazio, J Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian 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. 4, 14

[44] Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34:27952–27964, 2021. 4

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

[46] Benjamin Patrick Evans and Sumitra Ganesh. Learning and calibrating heterogeneous bounded rational market behaviour with multi-agent reinforcement learning. *arXiv preprint arXiv:2402.00787*, 2024. 4

[47] Athul Paul Jacob, David J Wu, Gabriele Farina, Adam Lerer, Hengyuan Hu, Anton Bakhtin, Jacob Andreas, and Noam Brown. Modeling strong and human-like gameplay with kl-regularized search. In *International Conference on Machine Learning*, pages 9695–9728. PMLR, 2022. 4

[48] 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. 4

[49] Runyu Zhang, Yang Hu, and Na Li. Soft robust MDPs and risk-sensitive MDPs: Equivalence, policy gradient, and sample complexity. In *The Twelfth International Conference on Learning Representations*, 2024. 4

[50] Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005. 4

[51] Kishan Panaganti and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning (ICML)*, pages 511–520, 2021. 4

[52] Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*, 2022. 4, 18

[53] Zaiyan Xu*, Kishan Panaganti*, and Dileep 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. 4, 18, 20

[54] Ali Yekkehkhany, Timothy Murray, and Rakesh Nagi. Risk-averse equilibrium for games. *arXiv preprint arXiv:2002.08414*, 2020. 4

[55] Zifan Wang, Yi Shen, Michael M Zavlanos, and Karl H Johansson. Learning of nash equilibria in risk-averse games. *arXiv preprint arXiv:2403.10399*, 2024. 4

[56] Yue Gao, Kry Yik Chau Lui, and Pablo Hernandez-Leal. Robust risk-sensitive reinforcement learning agents for trading markets. *arXiv preprint arXiv:2107.08083*, 2021. 4

[57] Oliver Slumbers, David Henry Mguni, Stefano B Blumberg, Stephen Marcus Mcaleer, Yaodong Yang, and Jun Wang. A game-theoretic framework for managing risk in multi-agent systems. In *International Conference on Machine Learning*, pages 32059–32087. PMLR, 2023. 4

[58] Sumitra Ganesh, Nelson Vadori, Mengda Xu, Hua Zheng, Prashant Reddy, and Manuela Veloso. Reinforcement learning for market making in a multi-agent dealer market. *Advances in Neural Information Processing Systems*, 2019. 4

[59] Hannes Eriksson, Debabrota Basu, Mina Alibeigi, and Christos Dimitrakakis. Risk-sensitive bayesian games for multi-agent reinforcement learning under policy uncertainty. *arXiv preprint arXiv:2203.10045*, 2022. 4

[60] Shangtong Zhang, Bo Liu, and Shimon Whiteson. Mean-variance policy iteration for risk-averse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10905–10913, 2021. 4

[61] Wei Qiu, Xinrun Wang, Runsheng Yu, Rundong Wang, Xu He, Bo An, Svetlana Obraztsova, and Zinovi Rabinovich. Rmix: Learning risk-sensitive policies for cooperative reinforcement learning agents. *Advances in Neural Information Processing Systems*, 34:23049–23062, 2021. 4

[62] Siqi Shen, Chennan Ma, Chao Li, Weiquan Liu, Yongquan Fu, Songzhu Mei, Xinwang Liu, and Cheng Wang. Riskq: risk-sensitive multi-agent reinforcement learning value factorization. *Advances in Neural Information Processing Systems*, 36:34791–34825, 2023. 4

[63] Kaiqing Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. Robust multi-agent reinforcement learning with model uncertainty. *Advances in neural information processing systems*, 33:10571–10583, 2020. 4

[64] Sihong He, Songyang Han, Sanbao Su, Shuo Han, Shaofeng Zou, and Fei Miao. Robust multi-agent reinforcement learning with state uncertainty. *Transactions on Machine Learning Research*, 2023. 4

[65] Laixi Shi, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Sample-efficient robust multi-agent reinforcement learning in the face of environmental uncertainty. *arXiv preprint arXiv:2404.18909*, 2024. 4

[66] Jose Blanchet, Miao Lu, Tong Zhang, and Han 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. 4

[67] Jeremy McMahan, Giovanni Artiglio, and Qiaomin Xie. Roping in uncertainty: Robustness and regularization in markov games. In *Forty-first International Conference on Machine Learning*, 2024. 4, 13

[68] Hans Föllmer and Alexander Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4):429–447, 2002. 10, 11, 12, 14

[69] Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155:1105–1123, 2012. 12, 13

[70] Ian M. Gemp and Sridhar Mahadevan. Online monotone games. *ArXiv*, abs/1710.07328, 2017. 15

[71] Oliver 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. 15

[72] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953. 15

[73] Michael Kearns and Satinder Singh. Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in neural information processing systems*, 11, 1998. 17

[74] Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020. 17

[75] Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 9582–9602, 2022. 18, 20

[76] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 2023. 18, 20

[77] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008. 20

[78] Amir Beck. *First-order methods in optimization*. SIAM, 2017. 22

Supplementary Materials

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).

A Matrix Games

To begin, we consider 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.

We note that the expected utility in finite-action games has a multi-linear form and that $U_i(\pi_i, \pi_{-i})$ is convex (actually linear) in π_i for fixed π_{-i}. To see this, one can observe that in two player games, these utility functions take the simpler, bilinear forms:

$$\begin{aligned} U_1(\pi_1, \pi_2) &= \mathbb{E}_{\substack{a_1\sim\pi_1 \\ a_2\sim x_2}}[R_1(a_1, a_2)] = \pi_1^T R_1 \pi_2 \\ U_2(\pi_1, \pi_2) &= \mathbb{E}_{\substack{a_1\sim\pi_1 \\ a_2\sim\pi_2}}[R_2(a_1, a_2)] = \pi_2^T R_2 \pi_1, \end{aligned} \tag{2}$$

where $R_1 \in \mathbb{R}^{A_1\times A_2}$ and $R_2 \in \mathbb{R}^{A_2\times A_1}$ are payoff matrices representing the utility or reward associated to each joint pure strategy.

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 $i = 1, ..., n$:*

$$U_i(\pi_i^*, \pi_{-i}^*) \geq U_i(\pi_i, \pi_{-i}^*) \quad \forall \pi_i \in \Delta_{A_i}$$

While such games are well known to admit at least one Nash equilibrium in mixed strategies, the Nash equilibrium may be intractable to compute. Furthermore, a preponderance of empirical evidence suggests that people do *not* play their Nash strategies and in fact are risk-averse and potentially bounded rational in their decision-making— inducing new forms of equilibria.

In the next two sections, we introduce generalizations of the expected utility game that allow us to model agents as both risk-averse and imperfect optimizers—both of these being key attributes of human decision-making. We show that while the addition of risk can undo the existing structure in the game, the addition of bounded rationality in the form of quantal responses allow us to derive a class of computationally tractable equilibria in *all* games that only depends on the degree of risk aversion and bounded rationality and not on the game structure.

A.1 Risk-Aversion in Matrix Games

Given the setup of an expected utility game, we now allow agents to have risk preferences which we model through the use of a general class of convex risk metrics which have been the focus of a long line of study in mathematical finance and operations research [68]. In this framing we move into a regime where agents seek to *minimize* a measure of risk.

Definition 2 (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 allow for agents to optimize for more conservative or worst-case outcomes. For ease of exposition, we write $\rho_\pi(X)$ to denote the distribution that the measure of risk is taken with respect to a reference distribution π.

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 generalizing this to stochastic games in future sections we will also consider the case where agents are risk averse with respect to the underlying dynamics in the Markov game. Crucially, we assume that 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 [36] and it is necessary since if agents are risk-averse to their own randomness then an equilibrium may cease to exist [32]. We refer to related works for more discussion.

We consider two generalizations of the original game that differ in how risk is incorporated into the problem: *aggregate risk aversion* and *action-dependent risk aversion*. The first generalization allows for a comparatively simpler analysis but at the cost of a more conservative solution while the second is less conservative but requires a more complex analysis. Due to monotonicity and linearity of expectation both of these formulations can be seen as generalizations of risk-sensitive decision-making investigated in the literature on risk-sensitive control [35] and risk-sensitive RL [36]. To see this reduction, one can simply take the other agents to be part of an unknown environment.

Aggregate Risk Aversion: To define a game with aggregate risk aversion, we transform the player's utilities in (1) into costs f_i which take the form

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

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: To define a game with action-dependent risk aversion, we transform the player's utilities in (1) into costs f_i which take the form

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

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 2) 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.

While—at first glance— the modified games look 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. In particular, we make use of the *dual* representation theorem for convex risk measures.

Theorem 3 (Dual Representation Theorem for Convex Risk Measures [68])**.** *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 2) 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

Risk-measure	Penalty function $D(p, q)$
Entropic Risk [69]	Kullback-Leibler (KL): $KL(p, q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$
[68]	Reverse KL (RKL): $\sum_i q_i \log\left(\frac{q_i}{p_i}\right)$
ϕ-Entropic Risk [69]	ϕ-Divergence: $\sum_i p_i\, \phi\left(\frac{p_i}{q_i}\right)$
Utility-based shortfall [68]	Utility-based Shortfall (u): $\inf_{\lambda>0} \frac{1}{\lambda}[r + \sum_i p_i u^\star\left(\lambda \frac{q_i}{p_i}\right)]$

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 $u^\star$.

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 setup, we can now write the aggregate risk-adjusted game in the following form:

$$f_i(\pi_i, \pi_{-i}) = \sup_{p_i \in \mathcal{P}_{-i}} -\pi_i^T R_i p_i - D_i(p_i, \pi_{-i}) \tag{5}$$

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-adjusted game takes the form:

$$f_i(\pi_i, \pi_{-i}) = \sum_{j \in \mathcal{A}_i} \pi_i(j) \left(\sup_{p_{i,j} \in \mathcal{P}_{-i}} -\langle R_{i,j}, p_{i,j} \rangle - D_i(p_{i,j}, \pi_{-i}) \right), \tag{6}$$

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

We note that we differentiate the penalty functions D_i to allow for the different agents to have different risk preferences. In this form, one can see that in a risk-adjusted 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.

Since the Nash equilibrium of the risk-adjusted game can be qualitatively different from that of the original game, we define a risk-averse Nash equilibrium (RNE) as follows. For ease of exposition we do not differentiate between aggregate and action-dependent RNE since similar results hold for either formulation.

Definition 3 (Risk-adjusted Nash equilibrium)**.** *A risk-averse Nash equilibrium 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-adjusted games admit at least one RNE.

Theorem 4. *There always exists a RNE for all aggregate and action-dependent risk-averse games presented in* (5) *and* (6) *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 [33] and thus a Nash equilibrium must exist. The action-dependent risk aversion regime follows from the linearity of f_i in π_i in (6) and a similar invocation of the existence of Nash in convex games [33]. □

The fact that these Nash equilibria exist is a consequence of the fact that players are not risk averse to the randomness that they introduce into the game by playing a mixed strategy. Under the stronger assumption that they are risk-averse to all randomness, previous work [32] shows that RNE may not exist. Thus, our specific formulation of risk is crucial in that it will always yield a RNE.

Despite the fact that 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-adjusted 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.

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 [69].

Finally, note that the 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., [67]). This is not surprising given the fact that computing NE in general convex games can be intractable and that further structural conditions are required for computational tractability.

A.2 Bounded Rationality in Matrix Games

Since risk aversion on its own is not sufficient to guarantee computational tractability of NE, we also introduce a second important feature of human decision-making into the game: bounded rationality. Interestingly, in experimental settings risk aversion—by itself— is also not enough to model human decision-making in game theoretic scenarios [16].

To incorporate bounded rationality into agents, we resort to the notion of a *quantal response function*.

Definition 4. *A* quantal response function *is a continuous function $\sigma : \mathbb{R}^n \to \Delta_n$ such that: 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.*

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:

$$\sigma(x) = \frac{\exp(-\frac{1}{\epsilon} x_i)}{\sum_{j=1}^n \exp(-\frac{1}{\epsilon} x_j)}, \tag{7}$$

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

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 4)—the resulting equilibrium is known as a quantal response equilibrium (QRE). To incorporate quantal

responses into our risk-adjusted game, we introduce regularization to the player's losses. This can be shown to be equivalent to constraining the player's responses to quantal responses (see, e.g., [68, Proposition 7], or [42,43]). We write a regularized risk-adjusted game as:

$$f_i^{\epsilon_i}(\pi_i, \pi_{-i}) = f_i(\pi_i, \pi_{-i}) + \epsilon_i \nu_i(\pi_i) \tag{8}$$

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.

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 4.*

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 term as a risk-adjusted quantal-response equilibrium (RQE).

Definition 5. *A risk-adjusted quantal response equilibrium (RQE) 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 the that game remains a convex game and thus RQE will exist.

As we will show in the following section, if player's risk preferences and regularizers (and thus their set of quantal response functions) satisfy a simple relationship, one can show that *irrespective of the underlying payoffs $R_1, ...R_n$*, the game admits RQE that can be efficiently computed using arbitrary no-regret learning algorithms which include e.g., gradient-play and mirror descent.

B Conditions for Computational Tractability of RQE

In this section we give conditions on the bounded rationality and degree of risk aversion of the players such that computing cost of RQE in the games is tractable. Further, an immediate consequence of our result is that players can find such equilibrium outcomes in a decentralized manner using no-regret learning algorithms.

Remark 1. For ease of exposition we focus on deriving the results for the aggregate risk-aversion formulation and defer the results for action-dependent risk-aversion to Appendix F. The results for both formulations in terms of computational tractability are similar in that both allow for a class of RQE to be learned using no-regret learning. However, we remark that the analysis of the aggregate case is comparatively simpler and allows one to compute simple bounds on the set of RQE that are computationally tractable. The action-dependent case requires more work and slightly more stringent requirements on the relationship between bounded rationality and risk-aversion.

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 we denote $p_i \in \mathcal{P}_{-i}$. 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:

$$J_i(\pi_i, \pi_{-i}, p) = -\pi_i^T R_i p_i - D_i(p_i, \pi_{-i}) + \epsilon_i \nu_i(\pi_i) \tag{9}$$

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

$$\bar{J}_i(\pi, p_i, p_{-i}) = \pi_i^T R_i p_i + D_i(p_i, \pi_{-i}) - \sum_{j \neq i} \xi_{i,j} \nu_j(\pi_j). \tag{10}$$

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. We note that unlike the original risk-adjusted game, the introduction of the new players gives us additional structure that we can exploit. A key feature that we introduce are the additional parameters $\xi_{i,j} > 0$ which we will use to characterize the degree of risk aversion of each player i.

Given the definition of the $2n$-player game we show how no-regret learning algorithms can be used to compute a RQE. To do so, we show that CCEs of the $2n$-player game coincide with NEs of it

and NEs 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 [23, 24]. To prove this, we first define coarse correlated equilibria.

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 $i = 1, ..., n$:*

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

With this definition in hand we now present our results on computational tractability of RQE in 2-player matrix games. We also extend the results to n-player setting in Appendix E.3.

To simplify our results we define $H_1(p_1, \pi_2) = D_1(p_1, \pi_2) - \xi_1\nu_2(\pi_2)$ and $H_2(p_2, \pi_1) = D_2(p_2, \pi_1) - \xi_2\nu_1(\pi_1)$. 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. The parameters ξ_1^* and ξ_2^* capture the player's degrees of risk aversion as illustrated in Corollaries 8.1 and 8.2 in Appendix E respectively.

The following theorem gives conditions on ξ_1^*, ξ_2^* and ϵ_1, ϵ_2, the bounded rationality parameters under which RQE's are computationally tractable using no-regret learning. The proof is postponed to Appendix E.2.

Theorem 5. *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*

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

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

The previous theorem gives a range of risk aversion and bounded rationality parameters under which an RQE can be computationally tractable using no-regret learning. Before presenting corollaries for certain instantiations of risk measures and quantal response functions in Appendix E, we briefly remark that the result does not necessarily guarantee uniqueness, though by exploiting the connections between socially convex games and monotone games [70] such a result would follow.

To validate this result, we show that the class of RQE that is computationally tractable seems sufficiently rich to capture the aggregate strategy played by people in various matrix games studied in laboratory experiments in behavioral economics [16, 22, 71] to showcase differing human behavior from Nash equilibrium strategies. In particular, we showcase in Fig. 1 that for some choice of parameters (ϵ_j and τ_j's) satisfying conditions in Theorem 5, we recover risk-averse and rational human behaviors in various matching pennies games mentioned in Tables 2 and 3. We provide more details about the games and experiments in Appendix D.1 for reproducing our results.

C Extension to Markov Games

In this section, we extend our results to finite-horizon Markov games, or stochastic games [72] which allow us to model dynamic games which are played out over Markov decision processes.

A Markov game can be seen as a sequential matrix game with dynamics. In this paper, we consider a general-sum finite-horizon risk-aversion Markov game 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 games; $\mathcal{S} = \{1, 2, \cdots, S\}$ represents the state space of the joint environment with size S; we reuse several definitions in matrix game (see Section A), $\{\mathcal{A}_i\}_{i\in[n]}$ is the action spaces with cardinality $|\mathcal{A}_i| = A_i$, the joint action space is $\mathcal{A}$, the joint action profile is $a \in \mathcal{A}$, and $\mathcal{P}$ is the product policy space; $R_{i,h}$ is the utility function of the i-th player at time step h for any $(i, h) \in [n] \times [H]$, i.e., $R_{i,h}(s, \boldsymbol{a})$ represents the immediate reward (or utility) received by the i-th player given state-action pair $(s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}$. For simplicity, we assume $R_{i,h}$ is deterministic; $P_{i,h} : \mathcal{S} \times \mathcal{A} \mapsto \Delta_S$ denote the transition probability so that $P_{i,h}(s' \mid s, \boldsymbol{a})$ represents the probability transitioning from current state s to the next state s' conditioned on the action $\boldsymbol{a}$.

Markov policies and value functions. In this section, we focus on a class of Markov policy, 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. Specifically, for any $1 \le i \le n$, the i-th agent executes actions according to a policy $\pi_i = \{\pi_{i,h} : \mathcal{S} \mapsto \Delta_{A_i}\}_{1 \le h \le H}$, with $\pi_{i,h}(a \mid s)$ the probability of selecting action a in state s at time step h. We consider the joint policy of all agents as a product policy defined as $\pi = (\pi_1, \ldots, \pi_n) : \mathcal{S} \times [H] \mapsto \mathcal{P}$ and the corresponding policy space as Π, namely, the joint action profile $\boldsymbol{a}$ of all agents is chosen according to the distribution specified by $\pi_h(\cdot \mid s) = (\pi_{1,h}, \pi_{2,h} \ldots, \pi_{n,h})(\cdot \mid s) \in \mathcal{P}$ 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.

Then for any joint policy π, we can introduce the corresponding value function $\widetilde{V}_{i,h}(\pi) : \mathcal{S} \mapsto \mathbb{R}$ (resp. Q-function $\widetilde{Q}_{i,h}(\pi) : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$) of the i-th agent to characterize the long-term cumulative reward by as follows:

$$\begin{aligned} \forall (s, a, h) \in \mathcal{A} \times \mathcal{S} \times [H] : \quad \widetilde{V}_{i,h}(\pi; s) &\coloneqq \mathbb{E}_\pi \left[\sum_{t=h}^{H} R_{i,t}\big(s_t, \boldsymbol{a}_t\big) \mid s_h = s \right], \\ \widetilde{Q}_{i,h}(\pi; s, \boldsymbol{a}) &\coloneqq \mathbb{E}_\pi \left[\sum_{t=h}^{H} R_{i,t}\big(s_t, \boldsymbol{a}_t\big) \mid s_h = s, \boldsymbol{a}_h = \boldsymbol{a} \right], \end{aligned} \tag{11}$$

where the expectation is taken over the Markovian trajectory $\{(s_t, \boldsymbol{a}_t)\}_{h \le t \le H}$ by executing the joint policy π (i.e., $\boldsymbol{a}_t \sim \pi_t(\cdot \mid s_t)$) over the Markov dynamics.

C.1 Risk-Averse Markov games

Given these definitions, now we are ready to introduce the risk-averse Markov games (RAMGs) — considering both the uncertainty of agents' strategies and the environments (rewards and transitions). Towards this, we first introduce two important risk-measures for agents' strategies and environment respectively to form RAMGs together. Resorting to two sets of convex (convex w.r.t. the first argument with the second argument fixed) and lower-semi-continuous functions $\{D_{\mathsf{pol},i}(\cdot,\cdot)\}_{i \in [n]}$ and $\{D_{\mathsf{env},i}(\cdot,\cdot)\}_{i \in [n]}$, given any policy π, the two risk-measures are defined as:

$$\begin{aligned} \forall i \in [n] : \quad f^{\pi}_{\mathsf{pol},i}(Q_i) &= \sup_{p_i \in \mathcal{P}_i} -\pi_i^T Q_i p_i - D_{\mathsf{pol},i}(p_i, \pi_{-i}), \\ \forall i \in [n] : \quad f_{\mathsf{env},i}(R, P, V) &= \inf_{\widetilde{P} \in \Delta_S} \Big[R + \widetilde{P} V + D_{\mathsf{env},i}(\widetilde{P}, P) \Big]. \end{aligned} \tag{12}$$

Here, $Q_i \in \mathbb{R}^{A_i \times \sum_{j \neq i} A_j}$ is the payoff matrix for the i-th player, $R \in \mathbb{R}^A$, $P \in \Delta_S$, $V \in \mathbb{R}^S$. 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 $h \in [H]$, which can be extended to diverse ones for each time step.

Armed with those two penalty functions $f^{\pi}_{\mathsf{pol},i}(\cdot)$ and $f_{\mathsf{env},i}(\cdot)$, in a RAMG, the i-th agent aim to minimize its own risk-averse long-term loss function $\{V_{i,h}(\pi)\}_{h \in [H]}$, where $V_{i,h}(\pi) : \mathcal{S} \mapsto \mathbb{R}$ for any joint strategy $\pi \in \Pi$ of the i-th agent is defined recursively as follows: for all $(h, s, \boldsymbol{a}) \in [H] \times \mathcal{S} \times \mathcal{A}$,

$$V_{i,h}(\pi; s) = f^{\pi}_{\mathsf{pol},i}\left(Q_{i,h}(\pi; s, :)\right) \tag{13}$$

where

$$Q_{i,h}(\pi; s, \boldsymbol{a}) = f_{\mathsf{env},i}\left(R_{i,h}(s, \boldsymbol{a}), P_{h,s,\boldsymbol{a}}, V_{i,h+1}(\pi)\right) \quad \text{and} \quad P_{h,s,\boldsymbol{a}} \coloneqq P_h(\cdot \mid s, \boldsymbol{a}) \in \mathbb{R}^{1 \times S}. \tag{14}$$

In words, $Q_{i,h}(\pi; s, :) \in \mathbb{R}^{A_i \times \sum_{j \neq i} A_j}$ is the payoff matrix of 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 $\boldsymbol{a} = (a_i, a_{-i})$. Therefore, the long-term loss function can be represented as a non-linear sum over the future immediate rewards as below:

$$V_{i,h}(\pi; s) = f^{\pi}_{\mathsf{pol},i} \circ f_{\mathsf{env},i}\big(R_{i,h}(s, :), P_{h,s,:}, \big[f^{\pi}_{\mathsf{pol},i} \circ f_{\mathsf{env},i}\big(R_{i,h+1}, P_{h+1}, \big[f^{\pi}_{\mathsf{pol},i} \circ f_{\mathsf{env},i}\big(R_{i,h+2}, P_{h+2}, \cdots\big)\big]\big)\big]\big). \tag{15}$$

Solution concepts of RAMGs: risk-averse quantal response equilibrium (RQE). After introducing the RAMG problem formulation, we will focus on the solution concept that incorporate both risk-aversion and bounded rationality of human decision-making — RQE. To specify, we first introduce the regularized value functions for any given $\{\epsilon_i \geq 0\}_{i\in[n]}$:

$$\forall (i,h,s) \in [n]\times[H]\times\mathcal{S}: \quad V_{i,h}^{\epsilon_i}(\pi;s) = \sup_{p_i\in\mathcal{P}_i} -\pi_i^T Q_{i,h}(\pi;s,:)p_i - D_{\mathsf{pol},i}(p_i,\pi_{-i}) + \epsilon_i\nu_i(\pi_i)$$
$$= V_{i,h}(\pi;s) + \epsilon_i\nu_i(\pi_i). \tag{16}$$

In addition, for any policy π, we denote $\pi_{i,-h} := \{\pi_{i,t}\}_{t=1,2,3,\cdots,h-1,h+1,\cdots,H}$ as the i-th agent's strategy over all time steps except step h.

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

$$\forall (i,s,h) \in [n]\times\mathcal{S}\times[H]: \quad V_{i,h}^{\epsilon_i}(\pi;s) \leq \min_{\pi_h':\mathcal{S}\mapsto\Delta_{\mathcal{A}_i}} V_{i,h}^{\epsilon_i}\big((\pi_h',\pi_{i,-h})\times\pi_{-i};s\big), \tag{17}$$

where $\nu_i(\cdot)$ is a strictly convex function over the simplex and controls the class of the i-th player's QRE with certain degree of bounded rationality $\epsilon_i \geq 0$.

C.2 Algorithm and Theoretical Guarantees

RAMGs with perfect information. Notice that the risk-aversion Markov games (RAMGs) are sequential games with multiple rounds, it is natural to learn the game from the last step $h = H$ and then back forward step by step. Towards this, at each time step h, for all $(i,h,s,\boldsymbol{a}) \in [n]\times[H]\times\mathcal{S}\times\mathcal{A}$, we define the (original) payoff matrices for all agents at any state s at time step h as

$$\forall (i,s,h) \in [n]\times\mathcal{S}\times[H]: \quad \widehat{Q}_{i,h}(s,\boldsymbol{a}) = R_{i,h}(s,\boldsymbol{a}) + \inf_{\widetilde{P}\in\Delta_S} \widetilde{P}\widehat{V}_{i,h+1} + D_{\mathsf{env},i}(\widetilde{P}, P_{h,s,\boldsymbol{a}}), \tag{18}$$

which shall be used to define a risk-averse matrix game against other players' strategies later so that we can calculate a QRE for it with tractable computation cost. Here, we denote the computing routine of a RQE of a matrix game as $\mathsf{RQE}(\cdot)$. Armed with this, we summarize the entire algorithm to compute RQE for RAMGs in Algorithm 1. The following theorem shows that computing RQE using Algorithm 1 is computationally tractable when certain conditions of risk-measure and bounded rationality are satisfied, which can be directly verified based on Theorem 9.

Theorem 6. *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 any $(s,h) \in \mathcal{S}\times[H]$, we can construct a $2n$-player game with the other n-additional players' payoff as (10) with $\xi_{i,j} = \xi_{i,j}^*$ for all $(i,j) \in [n]\times[n]$. And for all $i = 1,...,n$ we have $\epsilon_i \geq \sum_{j\neq i}\xi_{j,i}^*$, then the output policy $\widehat{\pi}$ from Algorithm 1 is a RQE of $\mathcal{MG}$.*

RAMGs can be seen as a sequential of matrix games conditioned on states. As a corollary of the results in matrix games (Theorem 9), the above theorem shows that the proposed solution concept — RQE in RAMGs enables computational tractability with any well-defined risk-measures $\{D_{\mathsf{env},i}\}_{i\in[n]}$.

RAMGs with imperfect environmental information. In the previous paragraph, we focus on the cases when the dynamics of the environment of RAMG — reward and transition kernel are perfectly known. While in practice, especially in MARL tasks, the environments can be complicated and unknown that need to be learned by interacting with (sampling from) the environments. In this subsection, we focus on such scenarios and provide the theoretical guarantees of Algorithm 1.

We focus on a fundamental sampling mechanism — assuming access to a generative model or a simulator [73, 74], which allows us to collect N independent samples for each state-action pair generated based on the true environment $\{R_{i,h}, P_h\}_{i\in[n],h\in[H]}$:

$$\forall (s,\boldsymbol{a},j) \in \mathcal{S}\times\mathcal{A}\times[n]: \quad s_{i,s,\boldsymbol{a},h} \overset{i.i.d}{\sim} P_h(\cdot\,|\,s,\boldsymbol{a}), \quad r_{j,s,\boldsymbol{a},h}^i = R_{j,h}(s,\boldsymbol{a}), \quad i \in [N]. \tag{19}$$

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

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 (18).
  end for
  for $s\in\mathcal{S}$ **do**
    Get $\pi_h(s)=\{\pi_{i,h}(s)\}_{1\le i\le n}\leftarrow \mathsf{RQE}\Big(\big\{f^{\pi,\epsilon_i}_{\mathsf{pol},i}(\widehat{Q}_{i,h}(s,:)\big\}_{1\le i\le n}\Big)$.
    Set $\widehat{V}_{i,h}(s)=f^{\pi_h,\epsilon_i}_{\mathsf{pol},i}\big(\widehat{Q}_{i,h}(s,:)\big)$.
  end for
end for
output: $\widehat{\pi}=\{\pi_h\}_{1\le h\le H}$.

In such imperfect information cases, we propose a model-based approach, which first constructs an empirical reward function and nominal transition kernel based on the collected samples and then applies Algorithm 1 to learn a QNE. First, the empirical reward function and transition kernel $\widehat{P}\in\mathbb{R}^{S\prod_{i=1}^n A_i\times S}$ can be constructed on the basis of the empirical frequency of state transitions,

$$\forall(s,\boldsymbol{a},j,h)\in\mathcal{S}\times\mathcal{A}\times[n]\times[H]:\quad \widehat{R}_{j,h}(s,\boldsymbol{a})=r^1_{j,s,\boldsymbol{a},h}\quad\text{and}$$
$$\widehat{P}_h(s'\,|\,s,\boldsymbol{a}):=\frac{1}{N}\sum_{i=1}^N\mathbb{1}\big\{s_{i,s,\boldsymbol{a}}=s'\big\}. \tag{20}$$

Then with such empirical reward and transition, we can apply the oracle summarized in Algorithm 1. The following theorem provides the finite-sample guarantees of this model-based approach as below:

Theorem 7. *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*

$$\max_{(i,s,h)\in[n]\times\mathcal{S}\times[H]}\Big\{V^{\epsilon_i}_{i,h}(\widehat{\pi};s)-\min_{\pi'_h:\mathcal{S}\mapsto\Delta_{\mathcal{A}_i}}V^{\epsilon_i}_{i,h}((\pi'_h,\widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i};s)\Big\}\le\delta \tag{21}$$

as long as the total number of samples satisfying

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

Remark 2. 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 remark that 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 [52,53,75,76]. Note that computing exact RQE can be computationally expensive or may not be necessary in practice. This result provides the first finite-sample guarantees of computing RQE in RAMGs. We show that even when the environment is not perfectly known but needs to be learned by samples, we can achieve approximate-RQEs of RAMGs using sufficient finite samples of size N_{all}.

Proof of Theorem 7. Armed with the estimated reward and transition kernel in (20), 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

$$\max_{(i,s,h)\in[n]\times\mathcal{S}\times[H]}\Big\{V^{\epsilon_i}_{i,h}(\widehat{\pi};s)-\min_{\pi'_h:\mathcal{S}\mapsto\Delta_{\mathcal{A}_i}}V^{\epsilon_i}_{i,h}((\pi'_h,\widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i};s)\Big\}\le\delta. \tag{22}$$

For convenience, we denote

$$\forall (i,h,s) \in [n] \times [H] \times \mathcal{S}: \quad V_{i,h}^{\star,\epsilon_i}(\widehat{\pi}_{-i};s) = \min_{\pi'_{i,h}:\mathcal{S}\mapsto\Delta_{A_i}} V_{i,h}^{\epsilon_i}((\pi'_{i,h},\widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i};s) \tag{23}$$

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

$$\forall (i,h,s) \in [n] \times [H] \times \mathcal{S}: \quad \pi_{i,h}^\star(s) := \mathrm{argmin}_{\pi'_i:\mathcal{S}\mapsto\Delta_{A_i}} V_{i,h}^{\epsilon_i}((\pi'_{i,h},\widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i};s). \tag{24}$$

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

$$\begin{aligned}
&V_{i,h}^{\epsilon_i}(\widehat{\pi}) - V_{i,h}^{\star,\epsilon_i}(\widehat{\pi}_{-i}) \\
&= V_{i,h}^{\epsilon_i}(\widehat{\pi}) - \widehat{V}_{i,h}^{\epsilon_i}(\widehat{\pi}) + \left(\widehat{V}_{i,h}^{\epsilon_i}(\widehat{\pi}) - \widehat{V}_{i,h}^{\epsilon_i}((\pi_{i,h}^\star,\widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i})\right) \\
&\quad + \left(\widehat{V}_{i,h}^{\epsilon_i}((\pi_{i,h}^\star,\widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i}) - V_{i,h}^{\epsilon_i}((\pi_{i,h}^\star,\widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i})\right) \\
&\le V_{i,h}^{\epsilon_i}(\widehat{\pi}) - \widehat{V}_{i,h}^{\epsilon_i}(\widehat{\pi}) + \left(\widehat{V}_{i,h}^{\epsilon_i}((\pi_{i,h}^\star,\widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i}) - V_{i,h}^{\epsilon_i}((\pi_{i,h}^\star,\widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i})\right) \\
&\le \|V_{i,h}^{\epsilon_i}(\widehat{\pi}) - \widehat{V}_{i,h}^{\epsilon_i}(\widehat{\pi})\|_\infty \mathbf{1} + \left\|\widehat{V}_{i,h}^{\epsilon_i}((\pi_{i,h}^\star,\widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i}) - V_{i,h}^{\epsilon_i}((\pi_{i,h}^\star,\widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i})\right\|_\infty \mathbf{1}
\end{aligned} \tag{25}$$

where the first inequality holds by applying Theorem 6 with the estimated RAMG $\widehat{\mathcal{MG}}$ so that $\widehat{\pi}$ is a RQE of $\widehat{\mathcal{MG}}$, i.e.,

$$\widehat{V}_{i,h}^{\epsilon_i}(\widehat{\pi}) \le \min_{\pi'_{i,h}:\mathcal{S}\mapsto\Delta_{A_i}} \widehat{V}_{i,h}^{\epsilon_i}((\pi'_{i,h},\widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i}) \le \widehat{V}_{i,h}^{\epsilon_i}((\pi_{i,h}^\star,\widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i}). \tag{26}$$

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

Step 1: developing the recursion. To control the two terms in (25), we consider that for any joint policy π and time step $(s,h)\in\mathcal{S}\times[H]$,

$$\begin{aligned}
&V_{i,h}^{\epsilon_i}(\pi;s) - \widehat{V}_{i,h}^{\epsilon_i}(\pi;s) \\
&\overset{\text{(i)}}{=} g_{\mathsf{pol},i}^\pi\left(Q_{i,h}(\pi;s,:)\right) - g_{\mathsf{pol},i}^\pi\left(\widehat{Q}_{i,h}(\pi;s,:)\right) \\
&\overset{\text{(ii)}}{=} \sup_{p_i\in\mathcal{P}_i} -\pi_i(s)^T g_{\mathsf{env},i}\left(R_{i,h}(s,:),P_{h,s,:},V_{i,h+1}(\pi)\right)p_i - D_{\mathsf{pol},i}(p_i,\pi_{-i}(s)) \\
&\quad - \left[\sup_{p_i\in\mathcal{P}_i} -\pi_i(s)^T g_{\mathsf{env},i}\left(R_{i,h}(s,:),\widehat{P}_{h,s,:},\widehat{V}_{i,h+1}(\pi)\right)p_i - D_{\mathsf{pol},i}(p_i,\pi_{-i}(s))\right] \\
&\le \sup_{p_i\in\mathcal{P}_i}\left|-\pi_i(s)^T\left[g_{\mathsf{env},i}\left(R_{i,h}(s,:),P_{h,s,:},V_{i,h+1}(\pi)\right) - g_{\mathsf{env},i}\left(R_{i,h}(s,:),\widehat{P}_{h,s,:},\widehat{V}_{i,h+1}(\pi)\right)\right]p_i\right| \\
&\le \left\|g_{\mathsf{env},i}\left(R_{i,h}(s,:),P_{h,s,:},V_{i,h+1}(\pi)\right) - g_{\mathsf{env},i}\left(R_{i,h}(s,:),\widehat{P}_{h,s,:},\widehat{V}_{i,h+1}(\pi)\right)\right\|_\infty,
\end{aligned} \tag{27}$$

where (i) holds by the definition in (16) and (13), and (ii) follows from the definition of $g_{\mathsf{pol},i}^\pi(\cdot)$ in (12).

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

$$\begin{aligned}
&\left|g_{\mathsf{env},i}\left(R_{i,h}(s,\boldsymbol{a}),P_{h,s,\boldsymbol{a}},V_{i,h+1}(\pi)\right) - g_{\mathsf{env},i}\left(R_{i,h}(s,:),\widehat{P}_{h,s,\boldsymbol{a}},\widehat{V}_{i,h+1}(\pi)\right)\right| \\
&= \Big|R_{i,h}(s,\boldsymbol{a}) - \sup_{\widetilde{P}\in\Delta_S} -\widetilde{P}V_{i,h+1}(\pi) - D_{\mathsf{env},i}(\widetilde{P},P_{h,s,\boldsymbol{a}}) \\
&\quad - \left(R_{i,h}(s,\boldsymbol{a}) - \sup_{\widetilde{P}\in\Delta_S} -\widetilde{P}\widehat{V}_{i,h+1}(\pi) - D_{\mathsf{env},i}(\widetilde{P},\widehat{P}_{h,s,\boldsymbol{a}})\right)\Big| \\
&= \Big| - \sup_{\widetilde{P}\in\Delta_S}\left(-\widetilde{P}V_{i,h+1}(\pi) - D_{\mathsf{env},i}(\widetilde{P},P_{h,s,\boldsymbol{a}})\right) + \sup_{\widetilde{P}\in\Delta_S}\left(-\widetilde{P}\widehat{V}_{i,h+1}(\pi) - D_{\mathsf{env},i}(\widetilde{P},\widehat{P}_{h,s,\boldsymbol{a}})\right)\Big| \\
&\overset{\text{(i)}}{\le} \sup_{\widetilde{P}\in\Delta_S}\Big| - \left(-\widetilde{P}V_{i,h+1}(\pi) - D_{\mathsf{env},i}(\widetilde{P},P_{h,s,\boldsymbol{a}})\right) + \left(-\widetilde{P}\widehat{V}_{i,h+1}(\pi) - D_{\mathsf{env},i}(\widetilde{P},\widehat{P}_{h,s,\boldsymbol{a}})\right)\Big|
\end{aligned}$$

$$\overset{\text{(ii)}}{=} \sup_{\widetilde{P}\in\Delta_S} \left|\widetilde{P}\left(V_{i,h+1}(\pi)-\widehat{V}_{i,h+1}(\pi)\right)\right| + \sup_{\widetilde{P}\in\Delta_S}\left|D_{\mathsf{env},i}(\widetilde{P},P_{h,s,\boldsymbol{a}})-D_{\mathsf{env},i}(\widetilde{P},\widehat{P}_{h,s,\boldsymbol{a}})\right|$$
$$\leq \left\|V_{i,h+1}(\pi)-\widehat{V}_{i,h+1}(\pi)\right\|_\infty + L\left\|P_{h,s,\boldsymbol{a}}-\widehat{P}_{h,s,\boldsymbol{a}}\right\|_1. \tag{28}$$

where the first equality holds by (18), (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 [53, 75, 76] to modify this step.

Plugging in (28) back to (27) and applying the results for all $s\in\mathcal{S}$ yields

$$\left\|V^{\epsilon_i}_{i,h}(\pi)-\widehat{V}^{\epsilon_i}_{i,h}(\pi)\right\|_\infty \leq \left\|V_{i,h+1}(\pi)-\widehat{V}_{i,h+1}(\pi)\right\|_\infty + L\underbrace{\max_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}}\left\|P_{h,s,\boldsymbol{a}}-\widehat{P}_{h,s,\boldsymbol{a}}\right\|_1}_{=:l_h}$$
$$= \left\|V^{\epsilon_i}_{i,h+1}(\pi)-\widehat{V}^{\epsilon_i}_{i,h+1}(\pi)\right\|_\infty + Ll_h. \tag{29}$$

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

$$\left\|V^{\epsilon_i}_{i,h}(\pi)-\widehat{V}^{\epsilon_i}_{i,h}(\pi)\right\|_\infty \leq \left\|V^{\epsilon_i}_{i,h+2}(\pi)-\widehat{V}^{\epsilon_i}_{i,h+2}(\pi)\right\|_\infty + Ll_h + Ll_{h+1}$$
$$\leq \cdots \leq \left\|V^{\epsilon_i}_{i,H+1}(\pi)-\widehat{V}^{\epsilon_i}_{i,H+1}(\pi)\right\|_\infty + L\sum_{t=h}^{H} l_t$$
$$\leq L\sum_{t=h}^{H} l_t, \tag{30}$$

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 (30). Applying [77, 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 $1-\delta$,

$$\max_{(h,s,\boldsymbol{a})\in[H]\times\mathcal{S}\times\mathcal{A}}\left\|P_{h,s,\boldsymbol{a}}-\widehat{P}_{h,s,\boldsymbol{a}}\right\|_1 \leq \sqrt{\frac{14S}{N}\log\left(\frac{2S\prod_{i\in[n]}A_iH}{\delta}\right)}. \tag{31}$$

Applying (31), we arrive at with probability at least $1-\delta$,

$$\forall h\in[H],\ l_h = \max_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}}\left\|P_{h,s,\boldsymbol{a}}-\widehat{P}_{h,s,\boldsymbol{a}}\right\|_1 \leq \sqrt{\frac{14S}{N}\log\left(\frac{2S\prod_{i\in[n]}A_iH}{\delta}\right)}. \tag{32}$$

Inserting (32) into (30) gives

$$\left\|V^{\epsilon_i}_{i,h}(\pi)-\widehat{V}^{\epsilon_i}_{i,h}(\pi)\right\|_\infty \leq L\sum_{t=h}^{H} l_t \leq HL\sqrt{\frac{14S}{N}\log\left(\frac{2S\prod_{i\in[n]}A_iH}{\delta}\right)}. \tag{33}$$

Finally, recalling (25) yields

$$V^{\epsilon_i}_{i,1}(\widehat{\pi}) - V^{\star,\epsilon_i}_{i,1}(\widehat{\pi}_{-i}) \leq \|V^{\epsilon_i}_{i,h}(\widehat{\pi})-\widehat{V}^{\epsilon_i}_{i,h}(\widehat{\pi})\|_\infty\mathbf{1} + \left\|\widehat{V}^{\epsilon_i}_{i,h}((\pi^\star_{i,h},\widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i}) - V^{\epsilon_i}_{i,h}((\pi^\star_{i,h},\widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i})\right\|_\infty\mathbf{1}$$
$$\leq 8HL\sqrt{\frac{S}{N}\log\left(\frac{2S\prod_{i\in[n]}A_iH}{\delta}\right)}\mathbf{1}, \tag{34}$$

where the last inequality holds by applying (33) with $\pi=\widehat{\pi}$ or $\pi=(\pi^\star_{i,h},\widehat{\pi}_{i,-h})\times\widehat{\pi}_{-i}$. □

D Experiments and Evaluation

We provide experimental results and details of our experiments in this section for both matrix and Markov game settings. We provide our code on an **anonymous Github link** https://github.com/FortunaAversa/RQE containing instructions to reproduce all results in this paper.

D.1 Experiments and Evaluation in 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 [22]. 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 [16]. 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 [16, 22] 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 [78] 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.

D.2 Experiments and Evaluation in Markov Games

We consider a grid-world problem to test our algorithm (Algorithm 1) for showcasing risk-averse and bounded rational behaviors. We note that we use our matrix game method (as described in Appendix D.1) for the RQE solver at Line 8 in Algorithm 1.

Cliff Walk Environment description: A grid of size 6×6 consists of some tiles that lead to the agents falling into the cliff. The environment setup is as follows: The cliff is the black grid with rewards -2. Agents/players are rewarded 0 for taking each step and 1 for when they reach the goal. The agents move in the direction of action with probability $p_{\mathrm{d}} = 0.9$ and takes uniform step in the unintended direction. To emphasize the risky choices of agents traversing along the cliff or traversing in each others' paths, we reduce p_{d} to 0.5 when the agents are both within a grid apart. The episode horizon H is 200.

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. We also showcase results with ℓ_1 metric for the environmental uncertainty metric $D_{\mathsf{env},i}$.

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 the left figure of Fig. 2, agent 1 is risk-averse to the environmental uncertainty induced by the cliff grids by traversing the longer route shown by the red arrow. Interestingly, complementing our risk-adjusted notion, agent 2 is risk-averse w.r.t both the environmental uncertainty and evading the risk of coming into agent 1's path (which leads to more chances of falling into the cliff). It makes a less rational and more risk-averse choice by reaching the end state shown by the blue arrow. In the case of the right figure of Fig. 2. They both make a more rational choice of reaching goal states but take a riskier path, increasing the chances of falling into the cliff. Interestingly, agent 1 waits at the corner grid (at the bottom right) until agent 2 passes it, since being at this grid is a more risk-averse agnostic choice against other players. These empirical evaluations support our theoretical result (Theorem 6) w.r.t the relative choice of risk-averse and bounded rationality parameters (ϵ_j and τ_j's) for providing different level of risk-averse solutions.

Now, 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 before 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.

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.

E 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 5 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

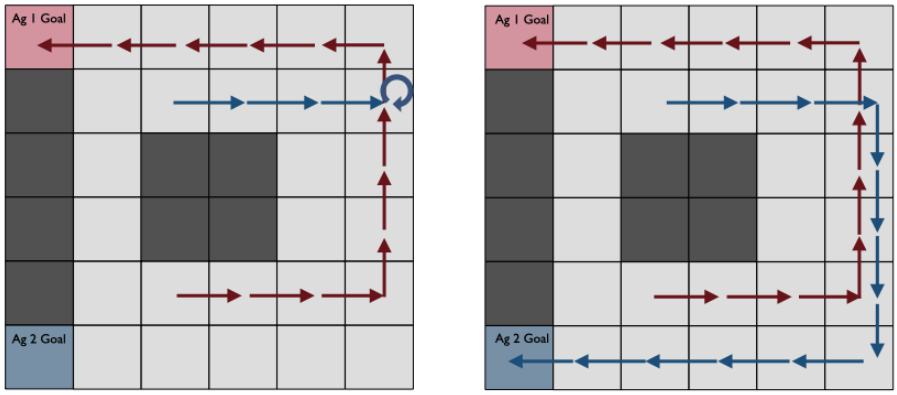

Figure 3: Cliff-Walk results for the ℓ_1 environmental uncertainty metric.

players have utilites of the form (10) and (9) and RQE of the desired game (3). We then use this result to prove Theorem 5.

E.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 [33]. 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 (8). Furthermore, if π^* is a RQE of Game (8), then (π^*, p^*) is a Nash equilibrium of the $2n$-player game, where:*

$$p_i^* = \arg\max_{p_i \in \mathcal{P}_{-i}} -\pi_i^{*T} R_i p_i - D_i(p_i, \pi_{-i}^*). \tag{35}$$

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 $i = 1, ..., n$:

$$\begin{aligned} J_i(\pi_i^*, \pi_{-i}^*, p^*) &\leq J_i(\pi_i', \pi_{-i}^*, p^*) \;\; \forall \;\; \pi_i' \in \Delta_{A_i} \\ \bar{J}_i(\pi^*, p_i^*, p_{-i}^*) &\leq \bar{J}_i(\pi^*, p_i', p_{-i}^*) \;\; \forall \;\; p_i' \in \mathcal{P}_{-i} \end{aligned} \tag{36}$$

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 (36) 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:

$$\begin{aligned} J_i(\pi_i^*, \pi_{-i}^*, p^*) &= f_i^{\epsilon}(\pi_i^*, \pi_{-i}^*) \\ &= \min_{\pi_i \in \Delta_{A_i}} \max_{p_i \in \mathcal{P}_{-i}} J(\pi_1, p_i, \pi_{-i}^*, p_{-i}^*) \\ &= \min_{\pi_i \in \Delta_{A_i}} J(\pi_1, p_i^*, \pi_{-i}^*, p_{-i}^*) \\ &\leq J(\pi_1', p_i^*, \pi_{-i}^*, p_{-i}^*) \ \ \forall \ \ \pi_i' \in \Delta_{A_i} \end{aligned}$$

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

E.2 Proof of Theorem 5 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 5. We also provide two corollaries that specialize the results to different risk-metrics and quantal response functions.

Theorem 8 (Restatement of Theorem 5)**.** *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 risk-adjusted 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:

$$\begin{aligned} \mathbb{E}_\sigma[J_1(\pi_1, p_1, \pi_2, p_2)] &\leq \mathbb{E}_{(p_1, \pi_2, p_2) \sim \sigma}[J_1(\pi_1', p_1, \pi_2, p_2)] \\ &= \mathbb{E}_{p_1 \sim \hat{p}_1}\left[\mathbb{E}_{\pi_2 \sim \sigma \,|\, p_1}\left[\mathbb{E}_{p_2 \sim \sigma \,|\, p_1, \pi_2}\left[J_1(\pi_1', p_1, \pi_2, p_2)\right]\right]\right] \\ &\leq \mathbb{E}_{p_1 \sim \hat{p}_1}\left[\mathbb{E}_{\pi_2 \sim \sigma \,|\, p_1}\left[J_1(\pi_1', p_1, \pi_2, \mathbb{E}_{\sigma \,|\, p_1, \pi_2}[p_2])\right]\right] \\ &\leq \mathbb{E}_{p_1 \sim \hat{p}_1}\left[J_1(\pi_1', p_1, \mathbb{E}_{\sigma \,|\, p_1}[\pi_2], \mathbb{E}_{\sigma \,|\, p_1}[p_2])\right] \end{aligned}$$

$$\leq J_1(\pi_1', \hat{p}_1, \hat{\pi}_2, \hat{p}_2) \quad \forall \pi_1' \in \Delta_{A_1}. \tag{37}$$

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:

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

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:

$$\frac{\lambda}{2}\left(J_1(\hat{z}) + \bar{J}_1(\hat{z})\right) + \frac{1-\lambda}{2}\left(J_2(\hat{z}) + \bar{J}_2(\hat{z})\right) \tag{38}$$

$$= \frac{\lambda}{2}\left(\mathbb{E}_\sigma\left[J_1(z)\right] + \mathbb{E}_\sigma\left[\bar{J}_1(z)\right]\right) + \frac{1-\lambda}{2}\left(\mathbb{E}_\sigma\left[J_2(z)\right] + \mathbb{E}_\sigma\left[\bar{J}_2(z)\right]\right). \tag{39}$$

By Eq. (37), we observe:

$$\mathbb{E}_\sigma\left[J_1(z)\right] \leq J_1(\hat{z}), \mathbb{E}_\sigma\left[\bar{J}_1(z)\right] \leq \bar{J}_1(\hat{z}), \mathbb{E}_\sigma\left[J_2(z)\right] \leq J_2(\hat{z}), \mathbb{E}_\sigma\left[\bar{J}_2(z)\right] \leq \bar{J}_2(\hat{z}). \tag{40}$$

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. (39) and (40), 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-adjusted game. □

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 use entropic risk and log-barrier regularizers.

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-adjusted 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 an 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-adjusted 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 an 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 5 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.

E.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-adjusted 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_1 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:

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

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

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-adjusted game. □

F 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-adjusted game of interest. In this case, the loss of the original players is given by:

$$J_i(\pi_i, \pi_{-i}, p) = \sum_{j \in \mathcal{A}_i} \pi_i(j) \left(-\langle R_{i,j}, p_{i,j} \rangle - D_i(p_{i,j}, \pi_{-i})\right) + \epsilon_i \nu_i(\pi_i). \tag{41}$$

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

$$\bar{J}_i(\pi, p_i) = \sum_{j \in \mathcal{A}_i} \pi_i(j) \left(\langle R_{i,j}, p_{i,j} \rangle + D_i(p_{i,j}, \pi_{-i})\right) - \sum_k \xi_{i,k} \nu_k(\pi_k), \tag{42}$$

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 (8). 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:*

$$p_{i,j}^* = \arg \min_{p_{i,j} \in \mathcal{P}_{-i}} \bar{J}_{i,j}(\pi^*, p_{i,j}) \tag{43}$$

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 5.

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:

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

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:

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

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.