Strategically Robust Multi-Agent Reinforcement Learning with Linear Function Approximation

Jake Gonzales, Max Horwitz, Eric Mazumdar, Lillian J. Ratliff

Keywords: strategic robustness, multi-agent reinforcement learning, game theory, regret analysis.

Summary

Provably efficient and robust equilibrium computation in general-sum Markov games remains a core challenge in multi-agent reinforcement learning. Nash equilibrium is computationally intractable in general and brittle due to equilibrium multiplicity and sensitivity to approximation error. We study Risk-Sensitive Quantal Response Equilibrium (RQRE), which yields a unique, smooth solution under bounded rationality and risk sensitivity. We propose `RQRE-OVI`, an optimistic value iteration algorithm for computing RQRE with linear function approximation in large or continuous state spaces. Through finite-sample regret analysis, we establish convergence and explicitly characterize how sample complexity scales with rationality and risk-sensitivity parameters. The regret bounds reveal a quantitative tradeoff: increasing rationality tightens regret, while risk sensitivity induces regularization that enhances stability and robustness. This exposes a Pareto frontier between expected performance and robustness, with Nash recovered in the limit of perfect rationality and risk neutrality. We further show that the RQRE policy map is Lipschitz continuous in estimated payoffs, unlike Nash, and RQRE admits a distributionally robust optimization interpretation. Empirically, we demonstrate that `RQRE-OVI` achieves competitive performance under self-play while producing substantially more robust behavior under cross-play compared to Nash-based approaches. These results suggest `RQRE-OVI` offers a principled, scalable, and tunable path for equilibrium learning with improved robustness and generalization.

Contribution(s)

1. This paper provides the first finite-sample regret guarantees for optimistic multi-agent reinforcement learning with linear function approximation under risk sensitivity and bounded rationality, and with approximate equilibrium computation.
   Context: Existing regret analyses typically assume a Nash equilibrium oracle or risk-neutral objectives (Wang et al., 2023b; Cisneros-Velarde & Koyejo, 2023).
2. A correspondence between risk-sensitive quantal response equilibrium (RQRE) and penalized distributionally robust optimization over opponent strategies is established.
   Context: This connects bounded rationality to robustness against payoff misspecification and generalizes recently proposed notions of robust equilibrium.
3. It is established that the RQRE policy map is Lipschitz continuous with respect to pay off perturbations, implying policy convergence when estimated value functions are sufficiently accurate.
   Context: Unlike Nash equilibria, RQRE policies vary smoothly with the estimated payoffs, yielding explicit bounds relating policy error to value-function estimation error.
4. Empirical evaluation of the proposed algorithm, `RQRE-OVI`, is performed on multi-agent coordination benchmarks.
   Context: Experiments demonstrate that RQRE-based learning achieves competitive self-play performance while producing substantially more robust behavior under cross-play compared to Nash-based approaches.

Strategically Robust Multi-Agent Reinforcement Learning with Linear Function Approximation

Jake Gonzales[1], **Max Horwitz**[1], **Eric Mazumdar**[2], **Lillian J. Ratliff**[1]
{jakegonz,maxh24,ratliffl}@uw.edu, mazumdar@caltech.edu

[1]**Department of Electrical & Computer Engineering, University of Washington**
[2]**Computing & Mathematical Sciences, California Institute of Technology**

Abstract

Provably efficient and robust equilibrium computation in general-sum Markov games remains a core challenge in multi-agent reinforcement learning. Nash equilibrium is computationally intractable in general and brittle due to equilibrium multiplicity and sensitivity to approximation error. We study Risk-Sensitive Quantal Response Equilibrium (RQRE), which yields a unique, smooth solution under bounded rationality and risk sensitivity. We propose `RQRE-OVI`, an optimistic value iteration algorithm for computing RQRE with linear function approximation in large or continuous state spaces. Through finite-sample regret analysis, we establish convergence and explicitly characterize how sample complexity scales with rationality and risk-sensitivity parameters. The regret bounds reveal a quantitative tradeoff: increasing rationality tightens regret, while risk sensitivity induces regularization that enhances stability and robustness. This exposes a Pareto frontier between expected performance and robustness, with Nash recovered in the limit of perfect rationality and risk neutrality. We further show that the RQRE policy map is Lipschitz continuous in estimated payoffs, unlike Nash, and RQRE admits a distributionally robust optimization interpretation. Empirically, we demonstrate that `RQRE-OVI` achieves competitive performance under self-play while producing substantially more robust behavior under cross-play compared to Nash-based approaches. These results suggest `RQRE-OVI` offers a principled, scalable, and tunable path for equilibrium learning with improved robustness and generalization.

1 Introduction

Modeling multiple strategic decision-makers in dynamic, real-world environments remains a central challenge in artificial intelligence. Multi-agent reinforcement learning provides a principled framework for such settings, where agents learn policies by interacting with the environment and adapting to the behavior of others (Zhang et al., 2021). This has enabled advances in domains ranging from autonomous driving (Cusumano-Towner et al., 2025), high-frequency trading (Mohl et al., 2025), multi-robot control (Gu et al., 2023), and alignment of large language models (Munos et al., 2024). Despite these empirical successes, developing principled algorithms with rigorous theoretical guarantees for multi-agent reinforcement learning remains an open challenge, particularly in settings with continuous or large state spaces, where equilibrium selection and brittleness lead to poor generalization and robustness.

Towards addressing scalability, Cisneros-Velarde & Koyejo (2023) proposed Nash Q-learning with Optimistic Value Iteration (NQOVI), extending the frameworks of Hu & Wellman (2003) and Liu et al. (2021) to the linear function approximation regime and providing finite-sample regret bounds that scale with feature dimension rather than the size of the state space. While this represents significant progress, a fundamental bottleneck persists: the algorithm requires solving for a Nash equilibrium at every stage game and every episode. This not only inherits the computational intractability of

Nash equilibria in general-sum games but also exposes the learning process to the inherent instability of the Nash correspondence (as we demonstrate in Example 2), where arbitrarily small perturbations in estimated payoffs can produce discontinuous jumps in the selected equilibrium strategy. In the function approximation setting, where Q-values are necessarily estimated with error, this brittleness is especially problematic: it means that the equilibrium computation at the core of the algorithm is ill-conditioned with respect to the approximation errors that the algorithm itself introduces.

These limitations motivate equilibrium concepts that are simultaneously computationally tractable, stable under payoff perturbations, and amenable to scalable reinforcement learning with formal guarantees. This leads to the central research question of this work:

> *Can we achieve provably sample-efficient learning of computationally tractable and robust equilibria in general-sum Markov games with linear function approximation?*

To address this question, we study the Risk-Sensitive Quantal Response Equilibrium (RQRE), a solution concept introduced in Mazumdar et al. (2025) which models agents as both boundedly rational and risk-averse. The RQRE combines two behavioral modeling choices, each of which independently contributes desirable properties for multi-agent learning while also capturing behaviors inherent to natural learning agents (as opposed to algorithmic ones). Indeed, *bounded rationality* replaces exact best responses with smooth, regularized response mappings, yielding well-posed and often unique equilibria while mitigating discontinuities and equilibrium multiplicity (McKelvey & Palfrey, 1995; Erev & Roth, 1998; Goeree & Holt, 1999). In multi-agent reinforcement learning, this framework improves stability under approximation error and non-stationarity, providing a principled tradeoff between optimality and robustness that aligns naturally with regularized learning objectives. At the same time, incorporating *risk sensitivity* addresses key limitations of purely expected-value objectives by discouraging policies that achieve high average performance at the expense of rare but catastrophic outcomes. In multi-agent settings, risk-sensitive formulations act as a principled form of robustification—reducing sensitivity to modeling error, noise, and opponent misspecification (Mazumdar et al., 2025; Lanzetti et al., 2025)—thereby promoting more stable, predictable, and generalizable equilibria.

Contributions. The main contributions of this work are summarized as follows:

- **Finite-sample guarantees.** We perform regret analysis for a novel algorithm, `RQRE-OVI`, and establish that regret satisfies $\mathrm{reg}(K) \leq \tilde{\mathcal{O}}\big(L_{\mathsf{env}}\, B\, \sqrt{K\, d^3\, H^3}\big) + KH\big(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}})\big)$, where $\varepsilon_{\mathsf{env}}, \varepsilon_{\mathsf{pol}}, \varepsilon_{\mathsf{eq}}$ capture environment-risk, policy-risk, and stage-equilibrium approximation errors. The bound explicitly characterizes how sample complexity scales with the rationality and risk-sensitivity parameters: the value range $B = H(1 + \log|A|/\epsilon)$ depends on the rationality parameter ϵ, while the solver error scales as $\varepsilon_{\mathsf{eq}} = O(B\sqrt{\Delta_{\mathsf{eq}}/\tau_{\min}})$, so that greater risk aversion relaxes solver accuracy requirements. Our analysis provides the first regret guarantees for optimistic MARL with linear function approximation with risk-sensitivity and approximate equilibrium computation, rather than relying on a Nash equilibrium oracle.
- **Distributional robustness.** We show that the RQRE are distributionally robust (Proposition 1), and generalize other robust notions of equilibrium. This connects the bounded rationality parameter to a formal notion of robustness against payoff misspecification.
- **Stability.** We prove that the RQRE policy map is Lipschitz continuous in estimated payoffs (Corollary 2), a stability property that Nash equilibria provably lack (Example 2). This theoretically justifies the use of RQRE over Nash. Moreover, it allows us to conclude policy convergence under additional regularity (Proposition 3).
- **Empirical evaluation.** We evaluate `RQRE-OVI` on popular multi-agent coordination benchmarks and demonstrate that it achieves competitive self-play performance while producing substantially more robust behavior under cross-play compared to Nash-based methods.

An extended related work discussion is given in Appendix A.

2 Preliminaries

In this section, we present the game-theoretic and multi-agent reinforcement learning preliminaries. We consider the framework of Markov games (also known as stochastic games), which generalizes the standard Markov decision process (MDP) to the multi-player setting (Shapley, 1953; Littman, 1994; Neyman et al., 2003).

We define an episodic Markov game of the form $\mathcal{MG} = (\mathcal{X}, \mathcal{A}, H, \mathcal{P}, r)$ with state space $\mathcal{X}$, joint action space $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$, episode length H, and transition kernels $\mathcal{P} = \{\mathcal{P}_h\}_{h\in[H]}$, where $\mathcal{P}_h(\cdot \mid x, a)$ denotes the transition kernel at step h. Agent i's reward is $r_i = \{r_{i,h}\}_{h=1}^H$ with $r_{i,h} : \mathcal{X}\times\mathcal{A} \to [0,1]$. At step h, agent i selects action $a_{i,h} \in \mathcal{A}_i$, and we write $a_h = (a_{1,h}, \ldots, a_{n,h}) \in \mathcal{A}$ for the joint action. The agents' action spaces are assumed finite; however, the state space can be arbitrarily large or continuous.

We denote agent i's policy by $\pi_i = \{\pi_{i,h}\}_{h\in[H]}$ with $\pi_{i,h} : \mathcal{X} \to \Delta(\mathcal{A}_i)$, and by an abuse of notation write $\pi_h : \mathcal{X} \to \Delta(\mathcal{A})$ for the joint policy. For agent i, define the value function $\widetilde{V}_h^i(\cdot;\pi) : \mathcal{X} \to \mathbb{R}$ at step h as $\widetilde{V}_h^i(x;\pi) := \mathbb{E}_\pi[\sum_{h'=h}^H r_{i,h'}(x_{h'}, a_{h'}) \,\big|\, x_h = x]$. The most common solution concept for games is the Nash equilibrium (Nash, 1950).

Definition 1 (Markov Nash Equilibrium)**.** A joint policy $\pi^* = (\pi_1^*, \ldots, \pi_n^*)$ is a *Nash equilibrium* if no player can improve their value by unilaterally deviating. That is, for all $i \in [n]$, all alternative policies π_i', and all initial states $s_0 \in \mathcal{X}$, the inequality $\widetilde{V}_1^i(s_0;\pi^*) \geq \widetilde{V}_1^i(s_0;\pi')$ holds, where $\pi' = (\pi_i', \pi_{-i}^*)$ denotes the joint policy in which player i deviates to π_i' while all others play π_{-i}^*. If $\widetilde{V}_1^i(s_0;\pi^*) \leq \varepsilon + \widetilde{V}_1^i(s_0;(\pi_i, \pi_{-i}^*))$ for all i and s_0, then π^* is an ε-approximate Nash equilibrium.

2.1 Bounded Rationality and Quantal Response Equilibrium

Bounded rationality provides a principled mechanism for smoothing the equilibrium correspondence, ensuring uniqueness, and regularizing the learning problem, properties that are essential for stable multi-agent learning under function approximation.

With this in mind, the following is the formalism for incorporating bounded rationality in the admissible policy class. Consider a *finite normal-form game* defined by the tuple $\mathcal{G} = ([n], \{\mathcal{A}_i\}_{i\in[n]}, \{u_i\}_{i\in[n]})$, where for each player $i \in [n]$, $u_i : \mathcal{A} \to \mathbb{R}$ is the utility function. We write a joint action as $a = (a_i, a_{-i})$, where a_{-i} represents the actions of all players except i. For a fixed profile π_{-i}, let $U_i(\pi_{-i}) \in \mathbb{R}^{|\mathcal{A}_i|}$ be the vector whose entry corresponding to action a_i is $u_i(a_i, \pi_{-i}) := \mathbb{E}_{a_{-i}\sim\pi_{-i}}[u_i(a_i, a_{-i})]$.

Definition 2 (Quantal Response Equilibrium (McKelvey & Palfrey, 1995))**.** A mixed-strategy profile $\pi^* = (\pi_1^*, \ldots, \pi_n^*)$ is a *Quantal Response Equilibrium* (QRE) if, for every player $i \in [n]$, the response is given by $\pi_i^* = \sigma_i\left(U_i(\pi_{-i}^*)\right)$, where $\sigma_i : \mathbb{R}^{|\mathcal{A}_i|} \to \Delta(\mathcal{A}_i)$ is a continuous quantal response function that is *monotonic* (i.e., for any $x \in \mathbb{R}^{|\mathcal{A}_i|}$, $x_a > x_{a'}$ implies $\sigma_i(a \mid x) > \sigma_i(a' \mid x)$) and *full support* (i.e., $\sigma_i(a \mid x) > 0$ for all $a \in \mathcal{A}_i$).

It is well known that quantal responses may be introduced into games via regularization; indeed this is shown via constraining the player's responses to quantal responses (see, e.g., Föllmer & Schied (2002); Hofbauer & Sandholm (2002)). This is straightforward via Fenchel duality.

2.2 Risk-Sensitive Markov Games

Having established the role of bounded rationality, we now incorporate risk aversion into the multi-agent framework. As discussed in Section 1, risk-sensitive objectives provide a principled mechanism for promoting safety, robustness to modeling error and strategic misspecification, and stability in settings with multiple equilibria or volatile dynamics. To formalize this, we equip each agent with a convex risk measure drawn from the class studied extensively in mathematical finance and operations research (Föllmer & Schied, 2002).

Definition 3 (Convex Risk Measure)**.** A risk functional $\rho : L^\infty(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ is a convex risk measure if it satisfies the following properties:

a. **Convexity:** For all $x, y \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ and $\lambda \in [0, 1]$, $\rho(\lambda x + (1-\lambda)y) \leq \lambda\rho(x) + (1-\lambda)\rho(y)$.
b. **Monotonicity:** For all $x, y \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ such that $x \leq y$, $\rho(x) \geq \rho(y)$.
c. **Translation Invariance:** For all $x \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ and $c \in \mathbb{R}$, $\rho(x + c) = \rho(x) + c$.

A fundamental property of convex risk measures is that they admit a dual representation as worst-case expectations over adversarial perturbations (Föllmer & Schied, 2002), providing a principled framework for modeling uncertainty in both environment dynamics and opponent strategies.

Theorem 1 (Dual Representation of Convex Risk Measures)**.** A mapping $\rho : \mathcal{Z} \to \mathbb{R}$ over a finite outcome space Ω is a convex risk measure if and only if there exists a convex, lower-semicontinuous penalty function $\varphi : \Delta(\Omega) \to (-\infty, \infty]$ such that $\rho(Z) = \sup_{p \in \Delta(\Omega)} \{\mathbb{E}_p[-Z] - \varphi(p)\}$.

Assumption 1 (Lipschitz Penalty Functions)**.** For each player $i \in [n]$, the penalty function $\varphi_i : \Delta(\Omega) \to (-\infty, \infty]$ in the dual representation (Theorem 1) is L_φ-Lipschitz continuous with respect to the ℓ_1 norm. That is, for all $p, q \in \Delta(\Omega)$, the bound holds: $|\varphi_i(p) - \varphi_i(q)| \leq L_\varphi \|p - q\|_1$.

This assumption ensures that the risk measure ρ is Lipschitz continuous with respect to the underlying distribution of returns and hence with respect to the Q-functions we estimate.

Risk-adjusted loss in matrix games. Consider an n-player matrix game with action sets $\{\mathcal{A}_i\}_{i=1}^n$ and payoff functions $u_i : \mathcal{A} \to \mathbb{R}$. Each player i minimizes a risk-adjusted loss based on the negation of their utility. For a joint mixed-strategy profile π, the risk-adjusted loss is $\ell_i(\pi_i, \pi_{-i}) := \rho_i(-u_i(a))$, where the randomness is over the joint action $a \sim (\pi_i, \pi_{-i})$. By Theorem 1 (dual representation), this loss is equivalently expressed as

$$\ell_i(\pi_i, \pi_{-i}) = \sup_{p_i \in \Delta(\mathcal{A}_{-i})} \left\{ \mathop{\mathbb{E}}_{a_{-i} \sim p_i} [-u_i(\pi_i, a_{-i})] - \varphi_i(p_i, \pi_{-i}) \right\}. \tag{1}$$

Here p_i represents an adversarial distribution over the opponents' joint actions, while $\varphi_i(\cdot, \pi_{-i})$ is a convex penalty that constrains the adversary by penalizing deviations from the reference policy π_{-i}. A canonical choice is the entropic risk measure.

Example 1 (Entropic Risk)**.** The entropic risk measure is obtained by setting the penalty function φ_i to be the Kullback–Leibler divergence scaled by a temperature parameter. The risk-adjusted loss then admits the closed form $\ell_i(\pi_i, \pi_{-i}) = \frac{1}{\tau_i} \log \left(\mathbb{E}_{a_{-i} \sim \pi_{-i}} [\exp(-\tau_i\, u_i(\pi_i, a_{-i}))] \right)$. The parameter $\tau_i > 0$ controls the degree of risk aversion: as $\tau_i \to 0$, the objective recovers the standard risk-neutral expected loss, whereas $\tau_i \to \infty$ yields the worst-case (minimax) objective.

Risk-Sensitive Quantal Response Equilibrium (RQRE). To model agents that are both risk-averse and boundedly rational, we augment the risk-adjusted loss with a regularization term. For each player i, let $\nu_i : \Delta(\mathcal{A}_i) \to \mathbb{R}$ be a strictly convex regularizer and $\epsilon_i > 0$ a precision parameter controlling the degree of bounded rationality. The *regularized risk-adjusted loss* is

$$l_i(\pi_i, \pi_{-i}) := \ell_i(\pi_i, \pi_{-i}) + \frac{1}{\epsilon_i} \nu_i(\pi_i),$$

where ℓ_i is the risk-adjusted loss in (1). A common example of a convex regularizer is negative entropy $\nu_i(p) = \sum_k p_k \log(p_k)$, for instance, if players are are constrained to logit response functions. Each player seeks to minimize this combined objective, balancing risk minimization against the entropy constraint imposed by bounded rationality.

Definition 4 (Mazumdar et al. (2025))**.** Consider an n-player normal-form game. A joint strategy profile $\pi^\star = (\pi_1^\star, \ldots, \pi_n^\star)$ is a *Risk-Averse Quantal Response Equilibrium (RQRE)* if, for every player $i \in [n]$, $\pi_i^* = \arg\min_{\pi_i \in \Delta(\mathcal{A}_i)} \{\ell_i(\pi_i, \pi_{-i}^*) + \frac{1}{\epsilon_i} \nu_i(\pi_i)\}$.

Under sufficient risk aversion and bounded rationality, Mazumdar et al. (2025) show that RQRE is computationally tractable via no-regret learning by establishing an equivalence to the coarse correlated equilibrium of a *lifted* $2n$*-player game*.

Extension to Risk-Sensitive Markov games. Now we show the extension of the static RQRE concept to the finite-horizon Markov game setting, considering both the uncertainty of agents' strategies and the environments (rewards and transitions). The key is applying the equilibrium condition stage-wise to the action-value functions.

Let us formalize the connection between the policy-side and environment-side risk operators used in the risk-sensitive Markov game and a single abstract convex risk functional ρ. As demonstrated via Theorem 1, a large class of convex risk functionals admit the dual representation $\rho(Z) = \sup_{p\in\Delta(\Omega)} \{\mathbb{E}_p[-Z] - \varphi(p)\}$. In risk-sensitive Markov games, risk operators arise from this dual form, but applied to different sources of randomness: (i) **Environment risk:** randomness in next state x_{h+1}; (ii) **Policy risk:** randomness in opponents' actions a_{-i}. The penalty functions $\varphi_{\mathsf{e},i}$ and $\varphi_{\mathsf{p},i}$ determine the particular choice of convex risk functional ρ in each case. We remark that these two sources of risk are fundamentally different and thus must be separated. Policy risk represents a form of strategic risk-aversion and changes as the opponents' strategies change over time. Environment risks are taken with respect to the randomness stemming from the random—but stationary—environment the agent is acting in.

Fixing (i, h, x, a), the randomness in the environment arises from the next state $x_{h+1} \sim \mathcal{P}_h(\cdot \mid x, a)$. Given a continuation value function $V_{i,h+1}(\cdot;\pi)$, define the random variable $Z^{\mathsf{e}}_{i,h}(x') := V_{i,h+1}(x';\pi)$. We define the environment risk operator

$$\mathcal{R}^{\mathsf{e}}_{i,h}\big(r_{i,h}, \mathcal{P}_h, V_{i,h+1}; x, a\big) := r_{i,h}(x,a) + \rho^{\mathsf{e}}_{i,h}\big(Z^{\mathsf{e}}_{i,h} \,\big|\, \mathcal{P}_h(\cdot \mid x, a)\big), \tag{2}$$

where $\rho^{\mathsf{e}}_{i,h}(Z|\mathcal{P}) = \inf_{\widetilde{\mathcal{P}}\in\mathcal{P}(\mathcal{X})}\{\int Z(x')\,\widetilde{\mathcal{P}}(dx') + \varphi_{\mathsf{e},i}(\widetilde{\mathcal{P}} \,\|\, \mathcal{P})\}$ is given in dual form. Thus, in this notation, we have action-value function $Q^i_h(x,a;\pi) = \mathcal{R}^{\mathsf{e}}_{i,h}\big(r_{i,h}, \mathcal{P}_h, V_{i,h+1}; x, a\big)$. The penalty function $\varphi_{\mathsf{e},i}$ determines the specific form of $\rho^{\mathsf{e}}_{i,h}$ (e.g. entropic risk, distributionally robust risk).

Now, let $\pi_h(\cdot \mid x)$ denote the joint mixed action. For each opponent action profile a_{-i}, define the induced cost $Z^{\mathsf{p}}_{i,h}(a_{-i}) := \sum_{a_i\in\mathcal{A}_i} \pi_{i,h}(a_i \mid x)\, Q^i_h(x, a_i, a_{-i};\pi)$. The policy-side risk operator is

$$\mathcal{R}^{\mathsf{p}}_{i,h}\big(Q^i_h, \pi_h; x\big) := \rho^{\mathsf{p}}_{i,h}\big(Z^{\mathsf{p}}_{i,h} \,\big|\, \pi_{-i,h}(\cdot \mid x)\big), \tag{3}$$

where the dual representation is $\rho^{\mathsf{p}}_{i,h}(Z \mid \pi_{-i}) = \sup_{p_i\in\mathcal{P}^{\mathsf{p}}_i}\{-\langle Z, p_i\rangle - \varphi_{\mathsf{p},i}(p_i, \pi_{-i})\}$. Hence, in this notation, the value function is $V^i_h(x;\pi) = \mathcal{R}^{\mathsf{p}}_{i,h}\big(Q^i_h, \pi_h; x\big)$.

With an additional strictly convex regularizer $\nu_i : \Delta(\mathcal{A}_i) \to \mathbb{R}$, the regularized value is

$$V^{\epsilon_i}_{i,h}(\pi; x) = \mathcal{R}^{\mathsf{p}}_{i,h}\big(Q^i_h, \pi_h; x\big) + \tfrac{1}{\epsilon_i}\nu_i(\pi_{i,h}(\cdot \mid x)). \tag{4}$$

Thus the risk-regularized Bellman recursion is defined by $Q^i_h = \mathcal{R}^{\mathsf{e}}_{i,h}\big(r_{i,h}, \mathcal{P}_h, V_{i,h+1}\big)$, and $V^i_h = \mathcal{R}^{\mathsf{p}}_{i,h}(Q^i_h, \pi_h)$.

Definition 5 (Risk Quantal Response Equilibrium for Markov Games)**.** Consider a finite-horizon risk-sensitive Markov game $\mathcal{MG}$ equipped with environment risk operators $\mathcal{R}^{\mathsf{e}}_{i,h}$ and policy risk operators $\mathcal{R}^{\mathsf{p}}_{i,h}$, as defined in (2)–(3), for each player $i \in [n]$. A Markov policy profile $\pi^\star = \{\pi^\star_h\}_{h=1}^H$ is called a *Risk Quantal Response Equilibrium (RQRE)* if for every stage $h \in [H]$, every state $x \in \mathcal{X}$, and every player $i \in [n]$, the mixed action $\pi^\star_{i,h}(\cdot \mid x)$ satisfies the fixed-point condition

$$\pi^\star_{i,h}(\cdot \mid x) \in \arg\min_{\mu_i\in\Delta(\mathcal{A}_i)} \Big\{\mathcal{R}^{\mathsf{p}}_{i,h}\big(Q^i_h(\cdot;\pi^\star), (\mu_i, \pi^\star_{-i,h}); x\big) + \frac{1}{\epsilon_i}\nu_i(\mu_i)\Big\}, \tag{5}$$

where the continuation action-value functions satisfy the risk-sensitive Bellman recursion $Q^i_h(x,a;\pi^\star) = \mathcal{R}^{\mathsf{e}}_{i,h}\big(r_{i,h}, \mathcal{P}_h, V_{i,h+1}(\cdot;\pi^\star); x, a\big)$, and $V^{\epsilon_i}_{i,h}(\pi^\star; x) = \mathcal{R}^{\mathsf{p}}_{i,h}\big(Q^i_h, \pi^\star_h; x\big) + \frac{1}{\epsilon_i}\nu_i(\pi^\star_{i,h}(\cdot \mid x))$, with terminal condition $V^{\epsilon_i}_{i,H+1}(\cdot;\pi^\star) \equiv 0$.

3 Risk-Sensitive QRE Optimistic Value Iteration

Consider an n-player episodic Markov game of the form $\mathcal{MG} = (\mathcal{X}, \mathcal{A}, H, \mathcal{P}, r)$ as defined in Section 2. To address large or continuous state spaces, we consider the setting of linear Markov games,

Algorithm 1 Risk QRE-Optimistic Value Iteration (RQRE-OVI)

Require: Risk-sensitivity $\tau_i > 0$, policy regularization (bounded rationality) $\epsilon_i > 0$.
Require: finite sets $\widehat{\mathcal{P}}_i^{\mathsf{p}}$ (policy risk) and $\widehat{\mathcal{P}}_i^{\mathsf{e}}$ (environment risk). Set $Q_h^{i,1} \equiv H$ for all i, h.

for $k = 1, 2, \ldots, K$ **do** ▷ Episode k
  for $h = H, H-1, \ldots, 1$ **do** ▷ Backward pass
    $\Lambda_h^k \leftarrow \lambda I_d + \sum_{t=1}^{k-1} \phi(x_h^t, a_h^t, h)\phi(x_h^t, a_h^t, h)^\top$
    $\pi_{h+1}^k(x) \leftarrow \texttt{RQRE}_\varepsilon(Q_{h+1}^{1,k}(x,\cdot), \ldots, Q_{h+1}^{n,k}(x,\cdot))$
    for $i = 1, \ldots, n$ **do**
      $\widehat{V}_{i,h+1}^{\epsilon_i}(\pi; x) := \max_{p_i \in \widehat{\mathcal{P}}_i^{\mathsf{p}}} \{-\pi_i(\cdot|x)^\top Q_{h+1}^{i,k}(x,:)\, p_i - \varphi_{\mathsf{p},i}(p_i, \pi_{-i})\} + \frac{1}{\epsilon_i}\nu_i(\pi_i)$.
      $\widehat{y}_{i,h}^t := r_{i,h}(x_h^t, a_h^t) + \widehat{\rho}_{i,h}^{\mathsf{e},\mathsf{k}}\left(\widehat{V}_{i,h+1}^{\epsilon_i}(\pi_{h+1}^k; x_{h+1}) \,\middle|\, x_h^t, a_h^t\right), \qquad t = 1, \ldots, k-1.$
      $w_h^{i,k} \leftarrow (\Lambda_h^k)^{-1} \sum_{t=1}^{k-1} \phi(x_h^t, a_h^t, h)\, \widehat{y}_{i,h}^t$
      $Q_h^{i,k}(x,a) \leftarrow \min\left\{(w_h^{i,k})^\top \phi(x,a,h) + \beta\sqrt{\phi(x,a,h)^\top (\Lambda_h^k)^{-1}\phi(x,a,h)},\ B\right\}$
  Execute episode k with $\pi_h^k(x) \leftarrow \texttt{RQRE}_\varepsilon(Q_h^{1,k}(x,\cdot), \ldots, Q_h^{n,k}(x,\cdot))$.

in which the transition kernels and reward functions are assumed to be linear. This assumption implies that the action-value functions are linear, as we will show. As discussed in Jin et al. (2020), this is not an assumption on the policy class but rather a statistical modeling assumption on how the data are generated.

Assumption 2. Consider a Markov game $\mathcal{MG} = (\mathcal{X}, \mathcal{A}, H, \mathcal{P}, r)$, and let $\pi^\star$ be an RQRE.

a. **The transition kernels and reward functions are linear.** That is, there exists a feature map $\phi : \mathcal{X} \times \mathcal{A} \times [H] \to \mathbb{R}^d$ such that for every $h \in [H]$, there exist d unknown signed measures $(\mu_h^1, \ldots, \mu_h^d)$ over $\mathcal{X}$ and an unknown vector $\theta_{i,h} \in \mathbb{R}^d$ satisfying $\mathcal{P}_h(x' \mid x, a) = \langle \phi(x,a,h), \mu_h(x') \rangle$, and $r_{i,h}(x,a) = \langle \phi(x,a,h), \theta_{i,h} \rangle$ for all $x, x' \in \mathcal{X}$, $a \in \mathcal{A}$, and $h \in [H]$. Without loss of generality, we assume that $\|\phi(x,a,h)\| \le 1$ for all $(x,a,h) \in \mathcal{X} \times \mathcal{A} \times [H]$ and $\max\{\max_{x \in \mathcal{X}} \|\mu_h(x)\|, \|\theta_{i,h}\|\} \le \sqrt{d}$ for all $h \in [H]$.

b. **Realizability.** For each player $i \in [n]$ and stage $h \in [H]$, there exists a vector $w_{i,h}^\star \in \mathbb{R}^d$ such that for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, the action-value map $Q_{i,h}^{\pi^\star}(x,a) = \langle \phi(x,a,h),\, w_{i,h}^\star \rangle$.

Critically, this assumption implies that for any policy, the action-value functions are linear in the feature map ϕ. Therefore, when designing multi-agent reinforcement learning algorithms, it suffices to focus on linear action-value functions. We refer the reader to Jin et al. (2020) for common examples of games encompassed by Assumption 2.

Regret Notion. Let $\pi_{i,h}^{\mathrm{br}}(\cdot \mid x) \in \arg\max_{\mu_i \in \Delta(\mathcal{A}_i)} V_{i,h}^{\epsilon_i}\big((\mu_i, \pi_{-i,h}(\cdot \mid x)); x\big)$, be the smoothed best response such that future continuation uses π_i^{br} recursively in

$$Q_{i,h}^{(\pi_i', \pi_{-i})}(x,a) := r_{i,h}(x,a) + \rho_{i,h}^{\mathsf{e}}(V_{i,h+1}^{\epsilon_i}((\pi_i', \pi_{-i}); X_{h+1}) \mid x, a)$$
$$V_{i,h}^{\epsilon_i}((\pi_i', \pi_{-i}); x) := \mathcal{R}_{i,h}^{\mathsf{pol}}\big(Q_{i,h}^{(\pi_i', \pi_{-i})}, (\pi_i', \pi_{-i})_h; x\big) + \tfrac{1}{\epsilon_i}\nu_i(\pi_i').$$

with terminal condition $V_{i,H+1}^{\epsilon_i} \equiv 0$. Define the so-called *exploitability regret* by

$$\mathrm{reg}(K) := \sum_{k=1}^K \max_{i \in [n]} \Big(V_{i,1}^{\epsilon_i}\big(\pi_i^{\mathrm{br}}(\pi_{-i}^k), \pi_{-i}^k;\, s_0\big) - V_{i,1}^{\epsilon_i}\big(\pi^k;\, s_0\big)\Big),$$

where the episode-wise regret term is zero precisely when the policy π_k is an RQRE, which itself is commonly unique under sufficient regularity (Mazumdar et al., 2025). This regret is also analogous to the regret notion considered in prior works (Cisneros-Velarde & Koyejo, 2023). Bounding the exploitability $\mathrm{reg}(K)$ implies that each π^k is *approximately unilaterally stable*: no player can gain much by deviating in the risk-regularized sense.

RQRE-OVI (Algorithm 1). The algorithm performs optimistic value iteration over episodes, solving an approximate RQRE at each stage game via the subroutine $\texttt{RQRE}_\varepsilon$. This replaces the Nash

oracle used in prior work (Cisneros-Velarde & Koyejo, 2023) with a unique, Lipschitz stable, and computationally tractable equilibrium computation, and risk measures for environmental, policy and opponent uncertainty. The environment-risk operator $\rho^{\mathsf{e}}_{i,h}$ is estimated from samples, for instance via finite dual discretization or closed-form evaluation as in the entropic case (Example 1); concrete instantiations and error analysis are provided in Appendix C.

Theorem 2 (Regret bound)**.** Consider Algorithm 1, and let Assumption 2 hold. Suppose that there exists constant $B > 0$ such that $0 \le V^{\epsilon_i}_{i,h}(\pi; x) \le B$ for all (i, h, x, π), and suppose the following approximations are given:

- **Environment & Policy Risk.** There exists $\varepsilon_{\mathsf{env}} \ge 0$ and $\varepsilon_{\mathsf{pol}}$ such that for all bounded X, $0 \le \rho^{\mathsf{env}}_{i,h}(X) - \widehat{\rho}^{k}_{i,h}(X) \le \varepsilon_{\mathsf{env}}$, and for all (i, h, x, π), $0 \le V^{\epsilon_i}_{i,h}(\pi; x) - \widehat{V}^{\epsilon_i}_{i,h}(\pi; x) \le \varepsilon_{\mathsf{pol}}$.
- **Stage RQRE Approximation.** For each (h, x, Q) the computed stage equilibrium $\tilde{\pi}_h(\cdot|x; Q)$ satisfies $|V^{\epsilon_i}_{i,h}(\tilde{\pi}_h; x, Q) - V^{\epsilon_i}_{i,h}(\pi^\star_h; x, Q)| \le \varepsilon_{\mathsf{eq}}$ for each i, where $\pi^\star_h$ is the exact stage equilibrium.

Then with probability at least $1 - \delta$, the estimate holds:

$$\mathrm{regret}(K) \;\le\; \widetilde{\mathcal{O}}\big(L_{\mathsf{env}}\, B\, \sqrt{K\, d^3\, H^3}\big) \;+\; KH\big(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}\big(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}}\big)\big), \tag{6}$$

where, as per usual, $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors in $(n, d, H, K, 1/\delta)$.

This theorem builds on the analysis of optimistic value iteration with linear function approximation (Cisneros-Velarde & Koyejo, 2023; Zhou et al., 2021; Jin et al., 2020), sharing the same elliptical potential and statistical machinery, but requires several new ingredients to handle the approximation errors introduced by the risk estimation and approximate equilibrium computation.

First, the covering number for the risk induced value class (Lemma 1) must be established for the RQRE smoothed best-response structure rather than for an exact Nash oracle. Unlike single-agent linear MDPs, the continuation values depend on the data-dependent mapping $Q \mapsto \tilde{\pi}(Q)$ and are subsequently transformed by nonlinear risk operators. As a result, the regression targets do not form a simple linear class, preventing a direct linear-bandit-style concentration argument. This requires controlling the complexity of the risk-induced value-function class via a covering argument. Second, the optimism guarantee (Lemma 2) must account for the combined effect of the confidence bonus at each episode and the approximate RQRE—rather than an exact Nash equilibrium—across all stages, translating the environment-risk, policy-risk, and equilibrium approximation errors into an additive optimism deficit. To this end, Lemma 3 relates the deviation gap at each stage to an optimistic Q-advantage plus the exploration bonus. Finally, a stage-wise performance difference recursion propagates these per-step errors through the risk-sensitive Bellman backup, where the Lipschitz constant L_{env} of the environment risk operator governs how approximation errors compound across stages. We provide more detail on novelty in addition to the full proof in Appendix B.

3.1 Regret Bounds in terms of Rationality and Risk Preferences

In Algorithm 1, we compute stage-wise $\varepsilon_{\mathsf{eq}}$–approximate RQRE. Here we give some sense in which this approximation depends on problem parameters. For example, a natural class of algorithms are *no-regret*; if agents run generic no-regret dynamics, then $\varepsilon_{\mathsf{eq}}$ is $\varepsilon_{\mathrm{eq}}(T) = \widetilde{\mathcal{O}}(\sum_i \sqrt{\log(|\mathcal{A}_i|)/T})$ to a coarse correlated equilibrium (Cesa-Bianchi & Lugosi, 2006; Blum & Mansour, 2007).

In the special case where the *policy-space* risk penalty is entropic (KL) and the player policy is entropy-regularized, there are a number of methods one can employ, including no-regret learning methods such as extra-gradient or mirror-prox, multiplicative weights, or iterative best response.

To this end, fix a stage h and state x. For player i, let $\nu_i \equiv \Phi$ be an entropy regularizer, and let $\varphi_{\mathsf{p},i} \equiv \frac{1}{\tau_i}\mathrm{KL}$. Then the policy-risk value functional admits the closed form

$$V^{\epsilon_i,\tau_i}_{i,h}(\pi; x; Q) = -\tfrac{1}{\tau_i} \log\big(\textstyle\sum_{a_{-i}} \pi_{-i}(a_{-i}|x)\, \exp(-\tau_i\, u^\pi_i(x, a_{-i}; Q))\big) + \tfrac{1}{\epsilon_i}\Phi(\pi_i(\cdot|x)), \tag{7}$$

where $u^\pi_i(x, a_{-i}; Q) := \sum_{a_i} \pi_i(a_i|x)\, Q_i(x, a_i, a_{-i})$. Here $Q_i(x, :)$ denotes the tensor slice $a \mapsto Q_i(x, a)$, and $\pi_i^\top Q_i\, p_i$ denotes $\sum_{a_i, a_{-i}} \pi_i(a_i|x)\, Q_i(x, a_i, a_{-i})\, p_i(a_{-i})$.

The entropy regularization also impacts the range of values for the value function. Indeed, assume stage rewards satisfy $r_{i,h}(x,a) \in [0,1]$ and $V_{i,H+1}^{\epsilon_i,\tau_i} \equiv 0$. Then for any policy profile π and any (h,x), the cumulative (undiscounted) reward-to-go is at most H. Moreover, the entropy satisfies $0 \le \Phi(\pi_i(\cdot|x)) \le \log|\mathcal{A}_i|$. Therefore the additive regularization term is bounded as $0 \le \Phi(\pi_i(\cdot|x))/\epsilon_i \le \log(|\mathcal{A}_i|)/\epsilon_i$. Since the entropic risk term in (7) is a (KL-regularized) log-sum-exp aggregation of bounded Q-values, it preserves the same scale as the underlying payoff (up to the horizon factor). Consequently, when Q is clipped to $[0,B]$ with $B := \max_{i\in[n]} B_i$ where $B_i := H\,(1+\log(|\mathcal{A}_i|)/\epsilon_i)$, we have $0 \le V_{i,h}^{\epsilon_i,\tau_i}(\pi;x;Q) \le B$ for all i,h,x,π. Here the factor H matches the horizon scaling of cumulative costs, while $\log|\mathcal{A}_i|/\epsilon_i$ captures the maximum magnitude of entropy regularization.

Corollary 1 (Regret under entropic policy-risk and regularization)**.** Under the assumptions of Theorem 2, specialize the policy-side risk/value operator to (7) and set the policy regularizer to $\nu_i(\mu_i) = \Phi(\mu_i)$. Suppose the stage solver returns a policy with duality gap at most Δ_{eq} uniformly over (h,x,Q) so that $\varepsilon_{\mathsf{eq}} \le c_{\mathsf{eq}}\, B\sqrt{\Delta_{\mathsf{eq}}/\tau_{\min}}$ where $\tau_{\min} := \min_{i\in[n]} \tau_i$, and for an absolute constant $c_{\mathsf{eq}} > 0$ the dependence on $\tau_{\min}$ is explicit. With probability at least $1-\delta$,

$$\mathrm{reg}(K) \le \widetilde{\mathcal{O}}(L_{\mathsf{env}}\, B\,\sqrt{K\, d^3\, H^3}) + KH\big(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}\varepsilon_{\mathsf{pol}} + L_{\mathsf{env}}\, c_{\mathsf{eq}}\, B\sqrt{\Delta_{\mathsf{eq}}/\tau_{\min}}\big). \tag{8}$$

For example, if agents run no-regret as mentioned above, then $\Delta_{\mathsf{eq}} \le \widetilde{\mathcal{O}}(\sum_i \sqrt{\log(|\mathcal{A}_i|)/T})$ where the equilibrium solver $\texttt{RQRE}_\varepsilon$ is run for T steps. If the variational inequality is strongly monotone with modulus $\mu = \mu(\tau)$ (cf. Corollary 2) and Lipschitz constant $L = L(\tau)$, then an extragradient/Mirror-Prox solver run for T steps yields $\varepsilon_{\mathrm{eq}}(T) \le C\exp(-T\mu/L)$. Plugging this into Theorem 2 replaces the additive equilibrium-solver term by $KHL_{\mathsf{env}}C\exp(-T\mu/L)$. In Appendix B.2, we provide more precise examples of these approximations.

Corollary 1 reveals the following parameter tradeoffs. Increasing ϵ_i (weaker entropy regularization) decreases $B(\epsilon)$ and thus both the leading statistical term and the equilibrium-solver error. For fixed Δ_{eq}, increasing $\tau_{\min}$ decreases the solver-error contribution as $\mathcal{O}(\sqrt{\Delta_{\mathsf{eq}}/\tau_{\min}})$. However, recovering risk neutrality ($\tau \to 0$) requires scaling $\Delta_{\mathsf{eq}} = \mathcal{O}(\tau_{\min}\varepsilon_{\mathsf{eq}}^2/B^2)$ to maintain a fixed $\varepsilon_{\mathsf{eq}}$. In Appendix B.4 we analyze the risk approximation errors as a function of ϵ and τ.

4 Distributional Robustness & Stability of RQRE

Beyond the ability to capture phenomena of natural learners, a key motivation for RQRE over Nash is that the additional structure introduced by bounded rationality and risk sensitivity yields both distributional robustness and stability under approximation errors.

4.1 Distributional Robustness of RQRE

In parallel to RQRE, several distributionally robust equilibrium concepts have been proposed, including strategically robust equilibrium via optimal transport (Lanzetti et al., 2025) and distributionally robust Nash equilibria (Alizadeh et al., 2025). The dual representation of the risk-adjusted loss (Section 2) reveals a direct connection between RQRE and distributionally robust optimization (DRO). In particular, the risk-adjusted loss (1) admits a natural interpretation as a *penalized DRO* objective: player i evaluates a strategy under an adversarial distribution over opponents' actions while paying a convex penalty for deviating from the reference distribution π_{-i}. Consequently, RQRE corresponds to a fixed point of regularized best responses under penalized distributional robustness. Classical ambiguity-set DRO arises as a special case of (1).

Proposition 1. Consider an n-player normal-form game with utilities $u_i : \mathcal{A} \to \mathbb{R}$.

a. **Any hard-DRO equilibrium is an RQRE.** The risk-adjusted loss ℓ_i defined for each fixed π_{-i} by a proper convex and lower semi-continuous map $p_i \mapsto \varphi_i(p_i,\pi_{-i})$ reduces to the ambiguity-set DRO loss $\ell_i(\pi_i,\pi_{-i}) = \sup_{p_i\in\mathcal{P}_i(\pi_{-i})} \mathbb{E}_{a_{-i}\sim p_i}\big[-u_i(\pi_i,a_{-i})\big]$.

b. **Convex-penalty RQRE strictly generalize indicator-penalty RQRE.** There exist proper convex penalties $\varphi_i(\cdot, \pi_{-i})$ for which the induced loss $\ell_i(\cdot, \pi_{-i})$ cannot be represented as a hard-DRO loss for any ambiguity set correspondence $\mathcal{P}_i(\cdot)$.

The first claim holds by deriving the stated reduction using $\varphi_i(\cdot, \pi_{-i}) = \mathbf{1}_{\mathcal{P}_i(\pi_{-i})}$ and the convex measure duality (cf. Theorem 1). In particular, the strategically robust equilibrium concept based on ambiguity sets proposed by Lanzetti et al. (2025) arises as a special case of RQRE obtained by choosing $\mathcal{P}_i(\pi_{-i})$ to be an optimal-transport ball around π_{-i}.

To see that the second claim holds, consider the KL penalty $\varphi_i(p_i, \pi_{-i}) = \frac{1}{\tau_i} \mathrm{KL}(p_i \| \pi_{-i})$ with $\tau_i > 0$ which induces the entropic risk functional of Example 1.

This is *not representable* as $\sup_{p_i \in \mathcal{P}_i(\pi_{-i})} \mathbb{E}_{p_i}[-u_i(\pi_i, a_{-i})]$ for any set $\mathcal{P}_i(\pi_{-i})$. Any hard-DRO mapping has the form $Z \mapsto \sup_{p \in \mathcal{P}} \mathbb{E}_p[Z]$ and is positively homogeneous: $\sup_{p \in \mathcal{P}} \mathbb{E}_p[tZ] = t \sup_{p \in \mathcal{P}} \mathbb{E}_p[Z]$ for all $t \geq 0$. The entropic mapping $Z \mapsto (1/\tau) \log \mathbb{E}_{\pi_{-i}}[\exp(\tau Z)]$ is not positively homogeneous in general: e.g., take $Z \in \{0, 1\}$ with probability $1/2$ each under π_{-i}, so that $\frac{1}{\tau} \log((1 + e^{2\tau})/2) \neq \frac{2}{\tau} \log((1 + e^{\tau})/2)$. Hence no ambiguity set $\mathcal{P}$ can make these mappings coincide for all Z, proving strictness. In fact, the following inclusions are generally true:

$$\text{RQRE} \supset \text{Penalized DRO Eq.} \supset \text{Hard ambiguity-set DRO Eq.} \supset \text{Nash Eq.} \tag{9}$$

Remark 1 (Existence versus expressivity)**.** Indicator penalties $\mathbf{1}_{\mathcal{P}}$ are convex but not Lipschitz on $\Delta(\mathcal{A}_{-i})$. The Lipschitz assumption on φ_i is used to obtain existence (and stability) of RQRE. Proposition 1 is an *expressivity* statement: it shows that hard ambiguity-set robustness is contained in the RQRE class, even though such penalties lie outside the Lipschitz-penalty subclass.

Now for Markov games, we show that RQRE are distributionally robust (see Appendix E for details).

Proposition 2 (RQRE in Markov Games are Distributionally Robust)**.** Let $\pi^\star$ be an RQRE of the risk-sensitive Markov game $\mathcal{MG}$ with penalty functions $\varphi_{\mathsf{p},i}$, $\varphi_{\mathsf{e},i}$ as defined in the dual representation (Theorem 1) and satisfying Assumption 1. Then $\pi^\star$ is distributionally robust with respect to the agents policy, the opponents' strategies, and the stochastic environment transitions.

4.2 RQRE Admit Stability Properties that Yield Performance Guarantees

Another main advantage of RQRE over Nash equilibrium for value-iteration-based algorithms with function approximation is stability: small errors in the estimated Q-functions should produce small changes in the computed equilibrium policy.

Since RQRE is unique by construction, the map from payoff tables to equilibrium policy is a well-defined function. In contrast, the Nash correspondence is set-valued in general-sum games, and any single-valued selection can be discontinuous (Fudenberg & Tirole, 1991; van Damme, 1987). This distinction is the source of the stability gap.

Corollary 2 (Lipschitz Stability of RQRE)**.** Suppose each player $i \in [n]$ employs an α_ν-strongly convex regularizer ν_i , and an α_{p}-strongly convex policy-risk penalty $\varphi_{\mathsf{p},i}(\cdot, q)$ for any q. The regularized risk-adjusted objective is $\mu := \frac{1}{\epsilon} + \tau\alpha_{\mathsf{p}} > 0$ strongly concave in π_i, and there exists a constant $c > 0$ such that for any $Q, \widetilde{Q}$, the estimate holds: $\|\pi^{\texttt{RQRE}}(Q) - \pi^{\texttt{RQRE}}(\widetilde{Q})\|_1 \leq \frac{c}{\mu} \|Q - \widetilde{Q}\|_\infty$. In particular, for entropy regularization and policy risk, $\alpha_\nu = 1/\epsilon$ and $\alpha_{\mathsf{p}} = \tau$, so $\mu = 1/\epsilon + \tau^2$.

The Lipschitz constant c/μ is controlled entirely by the bounded rationality and the risk-aversion parameters, both of which are design choices of `RQRE-OVI`. Stronger regularization (larger ϵ) or greater risk aversion (smaller τ) yields a smaller c/μ, improving stability at the cost of regret performance. In contrast, Nash equilibria in general-sum games are not guaranteed to be unique, so any algorithm relying on stage-wise equilibrium selection inherits the resulting instability.

Example 2 (Instability of Nash Equilibrium under Multiplicity)**.** Consider the two-player symmetric coordination game with payoff matrix $Q(\alpha) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & \alpha \end{smallmatrix}\right)$ for any $\alpha > 0$. This game has three Nash equilibria, including two pure-strategy equilibria (e_1, e_1) and (e_2, e_2). Standard equilibrium selection

criteria (e.g., risk-dominance or payoff-dominance) select (e_1, e_1) when $\alpha < 1$ and (e_2, e_2) when $\alpha > 1$. Hence any Lipschitz constant L would have to satisfy $L \geq \frac{\|\pi^{\text{NE}}(Q(1-\varepsilon)) - \pi^{\text{NE}}(Q(1+\varepsilon))\|_1}{\|Q(1-\varepsilon) - Q(1+\varepsilon)\|_\infty} = 1/\varepsilon$, which diverges as $\varepsilon \to 0$. Thus no finite uniform Lipschitz constant exists.

The instability is not due to a particular selection rule, but a consequence of equilibrium multiplicity itself. Under function approximation, estimated payoff tables are necessarily perturbed, and when the stage game admits multiple equilibria, arbitrarily small errors may cause a Nash oracle to return different policies across episodes (Cisneros-Velarde & Koyejo, 2023). `RQRE-OVI` *avoids this failure mode*: the regularization inherent in the formulation guarantees a unique equilibrium at every stage, and Corollary 2 ensures it varies smoothly with estimated payoffs. Such stability properties also translate to a form of policy convergence.

Proposition 3 (Robustness of `RQRE-OVI` under Linear Function Approximation)**.** Consider Algorithm 1 with parameter estimates $\widehat{w}_h$ satisfying $\max_{i\in[n]} \|\widehat{w}_h^i - w_h^{i,\star}\|_2 \leq \delta_h$ for all $h \in [H]$ where $w_h^{i,\star}$ are the true parameters (Assumption 2). The induced RQRE policies satisfy $\|\widehat{\pi}_h - \pi_h^\star\|_1 \leq \frac{c}{\mu}\delta_h$ at each stage where $\pi^\star = \{\pi_h^\star\}_{h\in[H]}$ is the RQRE of the Markov game and $c > 0$ is an absolute constant, and $\sup_x |V_1^{\widehat{\pi}}(x) - V_1^{\pi^\star}(x)| = O(\frac{B\bar{\delta}}{\mu} \sum_{t=0}^{H-1} L_{\text{env}}^t)$ when $\delta_h \leq \bar{\delta}$.

Observe, if the environment risk operator is non-expansive in $\|\cdot\|_\infty$, i.e., $L_{\text{env}} \leq 1$ (as holds for entropic risk and many coherent risk maps), then $\sup_x |V_1^{\widehat{\pi}}(x) - V_1^{\pi^\star}(x)| = O(HB\bar{\delta}/\mu)$. The complete finite-horizon analysis is in Appendix D.

5 Numerical Experiments

We evaluate `RQRE-OVI` on two multi-agent coordination benchmarks[1]: a *dynamic* Stag Hunt environment modified from the Melting Pot suite (Agapiou et al., 2022), and Overcooked (Gessler et al., 2025), specifically using the implementation from JaxMARL (Rutherford et al., 2024). For each environment, we compare `RQRE-OVI` across a range of risk-aversion parameters against `QRE-OVI` (risk-neutral bounded rationality) and `NQ-OVI` (Cisneros-Velarde & Koyejo, 2023).

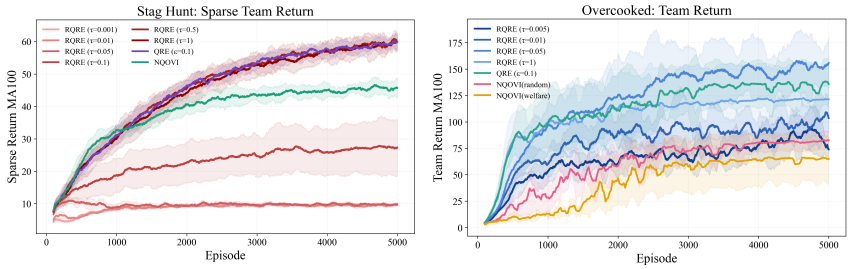

Figure 1: **Self-play team return during training.** Team return over episodes for Stag-Hunt (left) and Overcooked (right). Curves show mean ± 1 standard deviation over five seeds. In Stag-Hunt, higher τ drives agents toward the payoff-dominant (`stag`, `stag`) outcome, while lower τ yields the more robust risk-dominant (`hare`, `hare`) outcome. In Overcooked, all RQRE variants and QRE converge to comparable team returns, with Nash variants reaching similar or slightly lower levels.

We assess performance under three scenarios: **(i)** self-play, where agents train and evaluate with their own partner; **(ii)** cross-play with a perturbed partner, where an ego agent's partner deviates to a random or fixed action with probability δ at test time; **(iii)** cross-play with an unseen partner, where agents trained under different algorithms are paired at test time. Across both environments, we find RQRE achieves competitive self-play performance while producing *substantially more robust* behavior in cross-play, with the risk-aversion parameter τ governing a tradeoff between peak self-play return and worst-case robustness. A detailed description of the environments, training procedures, and hyperparameters is in Appendix F.

[1]Code is available at `https://jakeagonzales.github.io/linear-rqe-website/`.

Env. 1: Dynamic Stag Hunt. Stag Hunt is a classical coordination game in which two agents must choose between a safe, low-payoff action (hare) and a risky, high-payoff action (stag) that succeeds only under mutual coordination. We implement a dynamic, spatial variant of this game leveraging the Melting Pot framework (Agapiou et al., 2022). In our environment, agents navigate a 9×9 grid to collect resources and interact to resolve payoffs. The payoff matrix follows the standard structure: mutual stag yields $(4, 4)$, mutual hare $(2, 2)$, and stag-hare mismatch gives $(0, 2)$.

Self-play reveals a risk-return tradeoff. Under self-play (Figure 1, left) the risk aversion parameter τ directly controls which equilibrium agents coordinate on. High τ agents ($\tau \geq 0.5$) converge to the payoff-dominate outcome with team returns near 60, while low τ agents ($\tau \leq 0.05$) settle on the risk dominant outcome near 20. This is the tradeoff predicted by the theory—greater risk aversion biases agents toward strategies robust to coordination failure, at the cost of forgoing the higher-payoff but fragile stag equilibrium. `QRE-OVI` and `NQ-OVI` achieve intermediate returns converging to the payoff-dominant strategy.

Risk-averse agents degrade gracefully under partner perturbation. Under cross-play with a perturbed partner (Figure 2, left), more risk-averse agents (low τ) maintain high retention across all noise levels, as their convergence to the risk-dominant hare equilibrium renders them inherently insensitive to partner deviations. In contrast, less risk-averse agents coordinate on the fragile stag equilibrium and suffer sharp retention drops as δ increases. `NQ-OVI` and `QRE-OVI` exhibit similar fragility, having also converged to the payoff-dominant strategy.

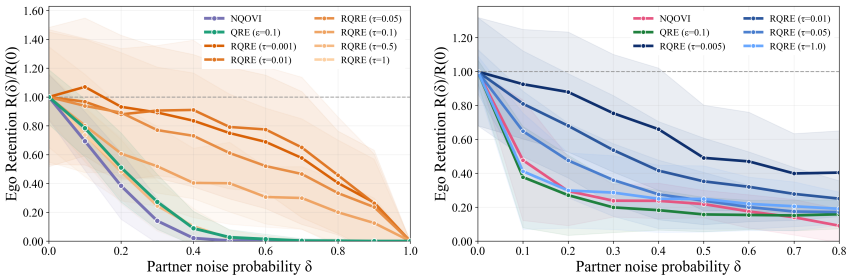

Figure 2: **Cross-play retention** $(R(\delta)/R(0))$ **as a function of perturbed partner noise** (δ)**:** Stag-Hunt (left) and Overcooked (right). At each evaluation step, the partner's action is a fixed deterministic action (e.g., always move in one direction) with probability δ and otherwise follows its trained policy. This produces high-signal deviations to emphasize the robustness phenomena. Curves are normalized by the $\delta = 0$ baseline, so higher values indicate strong robustness and lower values indicate performance degradation. Results are averaged over 200 evaluation rollouts per noise level.

Env. 2: "Overcooked" Cooperative Dynamic Game. The second environment on which we evaluate is Overcooked (Rutherford et al., 2024; Gessler et al., 2025): two agents cooperate in a small kitchen to prepare and deliver onion soup. The task requires a sequential chain of subtasks—picking onions, loading a pot, retrieving cooked soup with a plate, and delivering it—with a sparse team reward of 20 per delivery and shaped rewards for completing intermediate steps. The tight layout makes coordination essential, as agents must divide labor while avoiding blocking one another.

RQRE and QRE outperform NQOVI in self-play. Under self-play (Figure 1, right), all `RQRE-OVI` variants and `QRE-OVI` converge to team returns in medium to high ranges while `NQ-OVI` converges more slowly and to notably lower returns. We attribute this gap to the equilibrium selection problem inherent to Nash computation. Overcooked requires consistent coordination across many time steps, as agents must implicitly agree on a division of labor and maintain this role assignment throughout the episode. Because Nash equilibria are non-unique in general-sum games, solving for Nash at each stage game introduces the possibility of inconsistent selections across stages and episodes, disrupting the sustained coordination the task demands. `RQRE-OVI` and `QRE-OVI` avoid this failure mode entirely because the RQRE ensures a unique equilibrium at every stage, producing consistent

and coherent multi-step behavior. Among the RQRE variants, performance is relatively stable across risk levels indicating that moderate risk aversion does not substantially reduce self-play returns.

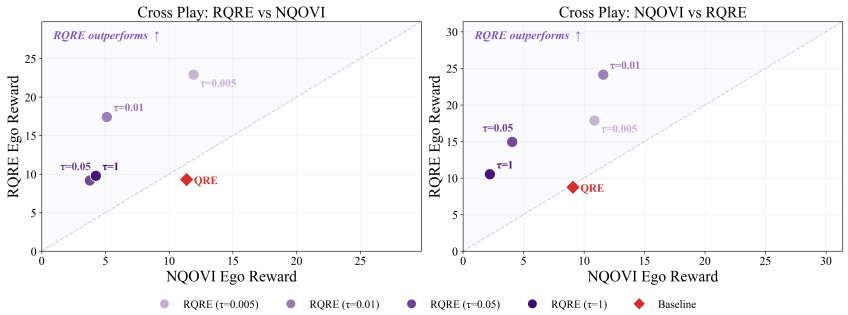

Figure 3: **Cross-play with unseen partners in Overcooked.** Each point represents the reward of two agents trained under a different algorithm and paired at test time without ever seeing each other before. The left panel shows `RQRE-OVI` agent's reward (vertical axis) versus `NQ-OVI` agent's reward (horizontal axis) for each pairing; the right panel reverses the roles. Points above the diagonal indicate that the agent on the vertical axis (in this case RQRE) achieves higher return than its partner. Labels denote the τ value of the corresponding `RQRE-OVI` agent, and the red diamond marks the `QRE-OVI` baseline. Across *all* pairings, `RQRE-OVI` agents achieve equal or higher ego reward than their cross-play partner, with moderate τ values (e.g., $\tau = 0.01$) yielding the strongest advantage.

RQRE retains performance under partner perturbation. Figure 2 shows that the moderate risk aversion retains $60 - 80\%$ of performance at $\delta = 0.3$, while `NQ-OVI`, `QRE-OVI`, and higher τ RQRE variants drop below 40%. Combined with the cross play with unseen partners results that follow, this suggests that risk aversion produces policies that are fundamentally more adaptive to partner deviation, rather than overfitting to the specific partner encountered during training.

RQRE dominates cross-play with unseen partners. Figure 3 shows rewards when agents trained under different algorithms are paired at test time without ever seeing each other before. Across nearly all pairings, `RQRE-OVI` agents achieve equal or higher ego reward than their partner, with points consistently above the diagonal. The advantage is most pronounced at moderate risk levels, where the `RQRE-OVI` agent achieves around $2 - 3\times$ the reward of the `NQ-OVI` partner. Notably, these results hold across both pairing directions, indicating that the robustness advantage reflects a genuinely more adaptive policy rather than an artifact of which agent is designated as ego. Moreover, the `QRE-OVI` agent performs comparably to `NQ-OVI` in cross-play, suggesting bounded rationality alone does not account for improved robustness and risk is the key driver of cross-play advantage.

6 Conclusion

In this paper, we study the problem of learning robust equilibria in general-sum Markov games with linear function approximation. We provide the first finite-sample regret guarantees for Risk-Sensitive Quantal Response Equilibria that explicitly characterize how the rationality and risk-sensitivity parameters govern sample complexity, establish that RQRE strictly generalizes existing distributionally robust equilibrium concepts, and prove Lipschitz stability of the equilibrium map under payoff perturbation—a property we leverage to obtain policy convergence guarantees under approximation error. The theoretical and empirical results suggest that robust and behaviorally grounded solution concepts offer a promising path for scalable multi-agent reinforcement learning beyond Nash equilibrium. Natural future directions include decentralizing the bonus term, which may be feasible under the monotonicity conditions established for RQRE in the infinite-horizon setting (Zhang & Mazumdar, 2025), and a more fine-grained understanding of the role of risk sensitivity, including risk-seeking agents and asymmetric risk profiles across players.

Acknowledgments

JG is supported by an NSF Graduate Research Fellowship under Grant No. DGE-2140004. LJR, MH, and JG are supported in part by NSF Award 2312775.

References

John P Agapiou, Alexander Sasha Vezhnevets, Edgar A Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, DJ Strouse, Michael B Johanson, Sukhdeep Singh, Julia Haas, Igor Mordatch, Dean Mobbs, and Joel Z Leibo. Melting pot 2.0. *arXiv preprint arXiv:2211.13746*, 2022.

Zeinab Alizadeh, Azadeh Farsi, and Afrooz Jalilzadeh. Distributionally robust nash equilibria via variational inequalities. *OPT2025 Workshop on Optimization for ML (arXiv preprint arXiv:2510.17024)*, 2025.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999. DOI: https://doi.org/10.1111/1467-9965.00068. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9965.00068.

Robert J. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1):67–96, 1974. ISSN 0304-4068. DOI: https://doi.org/10.1016/0304-4068(74)90037-8. URL https://www.sciencedirect.com/science/article/pii/0304406874900378.

Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(6), 2007.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player zero-sum linear mixture markov games, 2022. URL https://arxiv.org/abs/2102.07404.

Pedro Cisneros-Velarde and Sanmi Koyejo. Finite-sample guarantees for Nash Q-learning with linear function approximation. In *Uncertainty in Artificial Intelligence*, pp. 424–432. PMLR, 2023. URL https://arxiv.org/abs/2303.00177.

Qiwen Cui, Kaiqing Zhang, and Simon Du. Breaking the curse of multiagents in a large state space: Rl in markov games with independent linear function approximation. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2651–2652. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/cui23a.html.

Marco Francis Cusumano-Towner, David Hafner, Alexander Hertzberg, Brody Huval, Aleksei Petrenko, Eugene Vinitsky, Erik Wijmans, Taylor W. Killian, Stuart Bowers, Ozan Sener, Philipp Kraehenbuehl, and Vladlen Koltun. Robust autonomy emerges from self-play. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=yOXoJpG6Qy.

Yan Dai, Qiwen Cui, and Simon S. Du. Refined sample complexity for markov games with independent linear function approximation (extended abstract). In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1260–1261. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/dai24a.html.

Constantinos Daskalakis. On the complexity of approximating a nash equilibrium. *ACM Trans. Algorithms*, 9(3), June 2013. ISSN 1549-6325. DOI: 10.1145/2483699.2483703. URL https://doi.org/10.1145/2483699.2483703.

Eric Eaton, Marcel Hussing, Michael Kearns, Aaron Roth, Sikata Bela Sengupta, and Jessica Sorrell. Replicable reinforcement learning with linear function approximation, 2025. URL https://arxiv.org/abs/2509.08660.

Ido Erev and Alvin E Roth. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, 88(4):848–881, September 1998. DOI: None. URL https://ideas.repec.org/a/aea/aecrev/v88y1998i4p848-81.html.

Hans Föllmer and Alexander Schied. Convex measures of risk and trading constraints. *Finance and stochastics*, 6(4):429–447, 2002.

Drew Fudenberg and Jean Tirole. *Game Theory*. MIT Press, Cambridge, MA, 1991.

Tobias Gessler, Tin Dizdarevic, Ani Calinescu, Benjamin Ellis, Andrei Lupu, and Jakob Nicolaus Foerster. Overcookedv2: Rethinking overcooked for zero-shot coordination. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=h1vLM3GX8R.

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. DOI: 10.1073/pnas.96.19.10564. URL https://www.pnas.org/doi/abs/10.1073/pnas.96.19.10564.

Shangding Gu, Jakub Grudzien Kuba, Yuanpei Chen, Yali Du, Long Yang, Alois Knoll, and Yaodong Yang. Safe multi-agent reinforcement learning for multi-robot control. *Artif. Intell.*, 319(C), June 2023. ISSN 0004-3702. DOI: 10.1016/j.artint.2023.103905. URL https://doi.org/10.1016/j.artint.2023.103905.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Matthias Heger. Consideration of risk in reinforcement learning. In William W. Cohen and Haym Hirsh (eds.), *Machine Learning Proceedings 1994*, pp. 105–111. Morgan Kaufmann, San Francisco (CA), 1994. ISBN 978-1-55860-335-6. DOI: https://doi.org/10.1016/B978-1-55860-335-6.50021-0. URL https://www.sciencedirect.com/science/article/pii/B9781558603356500210.

Daniel Hernández-Hernández and Steven I. Marcus. Risk sensitive control of markov processes in countable state space. *Syst. Control Lett.*, 29(3):147–155, November 1996. ISSN 0167-6911. DOI: 10.1016/S0167-6911(96)00051-5. URL https://doi.org/10.1016/S0167-6911(96)00051-5.

Josef Hofbauer and William H Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294, 2002.

Ronald A. Howard and James E. Matheson. Risk-sensitive markov decision processes. *Management Science*, 18(7):356–369, 1972. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/2629352.

Junling Hu and Michael P Wellman. Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/jin20a.html.

Jonathan Lacotte, Yinlam Chow, Mohammad Ghavamzadeh, and Marco Pavone. Risk-sensitive generative adversarial imitation learning. *CoRR*, abs/1808.04468, 2018. URL http://arxiv.org/abs/1808.04468.

Nicolas Lanzetti, Sylvain Fricker, Saverio Bolognani, Florian Dörfler, and Dario Paccagnan. Strategically robust game theory via optimal transport, 2025. URL https://arxiv.org/abs/2507.15325.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning*, ICML'94, pp. 157–163, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 1558603352.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pp. 7001–7010. PMLR, 2021. URL https://arxiv.org/abs/2010.01604.

Eric Mazumdar, Lillian J Ratliff, Tanner Fiez, and S Shankar Sastry. Gradient-based inverse risk-sensitive reinforcement learning. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pp. 5796–5801. IEEE, 2017.

Eric Mazumdar, Kishan Panaganti, and Laixi Shi. Tractable multi-agent reinforcement learning through behavioral economics. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=stUKwWBuBm.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995. ISSN 0899-8256. DOI: https://doi.org/10.1006/game.1995.1023. URL https://www.sciencedirect.com/science/article/pii/S0899825685710238.

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. URL https://openreview.net/forum?id=y6y2HauOpR.

Panayotis Mertikopoulos and William H. Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/26179987.

Valentin Mohl, Sascha Frey, Reuben Leyland, Kang Li, George Nigmatulin, Mihai Cucuringu, Stefan Zohren, Jakob Foerster, and Anisoara Calinescu. Jaxmarl-hft: Gpu-accelerated large-scale multi-agent reinforcement learning for high-frequency trading, 2025. URL https://arxiv.org/abs/2511.02136.

Hervé Moulin and Jean-Philippe Vial. Strategically zero-sum games: The class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory*, 7(3–4):201–221, 1978. ISSN 1432-1270. DOI: 10.1007/BF01769190. URL https://doi.org/10.1007/BF01769190.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(27):815–857, 2008. URL http://jmlr.org/papers/v9/munos08a.html.

Remi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36743–36768. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/munos24a.html.

John F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950. DOI: 10.1073/pnas.36.1.48. URL https://www.pnas.org/doi/abs/10.1073/pnas.36.1.48.

Abraham. Neyman, S. Sorin, and North Atlantic Treaty Organization. Scientific Affairs Division. *Stochastic Games and Applications*. Nato Science Series C:, Mathematical and Physical Sciences ; 570. Springer Netherlands, Dordrecht, 1st ed. 2003. edition, 2003. ISBN 94-010-0189-8.

Shuang Qiu, Jieping Ye, Zhaoran Wang, and Zhuoran Yang. On reward-free rl with kernel and neural function approximations: Single-agent mdp and markov game, 2022. URL https://arxiv.org/abs/2110.09771.

Wei Qiu, Xinrun Wang, Runsheng Yu, Xu He, Rundong Wang, Bo An, Svetlana Obraztsova, and Zinovi Rabinovich. Rmix: learning risk-sensitive policies for cooperative reinforcement learning agents. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Lillian J Ratliff and Eric Mazumdar. Inverse risk-sensitive reinforcement learning. *IEEE Transactions on Automatic Control*, 65(3):1256–1263, 2019.

Andrzej Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Math. Program.*, 125(2):235–261, October 2010. ISSN 0025-5610.

Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Gardar Ingvarsson, Timon Willi, Ravi Hammond, Akbir Khan, Christian Schroeder de Witt, Alexandra Souly, Saptarashmi Bandyopadhyay, Mikayel Samvelyan, Minqi Jiang, Robert Tjarko Lange, Shimon Whiteson, Bruno Lacerda, Nick Hawes, Tim Rocktäschel, Chris Lu, and Jakob Nicolaus Foerster. Jaxmarl: Multi-agent rl environments and algorithms in jax. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

L. S. Shapley. Stochastic games*. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, 1953. DOI: 10.1073/pnas.39.10.1095. URL https://www.pnas.org/doi/abs/10.1073/pnas.39.10.1095.

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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Yun Shen, Michael J. Tobia, Tobias Sommer, and Klaus Obermayer. Risk-sensitive reinforcement learning. *Neural Computation*, 26(7):1298–1328, 07 2014. ISSN 0899-7667. DOI: 10.1162/NECO_a_00600. URL https://doi.org/10.1162/NECO_a_00600.

Laixi Shi, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Sample-efficient robust multi-agent reinforcement learning in the face of environmental uncertainty. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Sumeet Singh, Jonathan Lacotte, Anirudha Majumdar, and Marco Pavone. Risk-sensitive inverse reinforcement learning via semi- and non-parametric methods. *Int. J. Rob. Res.*, 37(13–14):

1713–1740, December 2018. ISSN 0278-3649. DOI: 10.1177/0278364918772017. URL https://doi.org/10.1177/0278364918772017.

Oliver Slumbers, David Henry Mguni, Stefano B. Blumberg, Stephen McAleer, Yaodong Yang, and Jun Wang. A game-theoretic framework for managing risk in multi-agent systems. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

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. URL https://openreview.net/forum?id=DpE5UYUQzZH.

John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. In M.C. Mozer, M. Jordan, and T. Petsche (eds.), *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. URL https://proceedings.neurips.cc/paper_files/paper/1996/file/e00406144c1e7e35240afed70f34166a-Paper.pdf.

Eric van Damme. *Stability and Perfection of Nash Equilibria*. Springer-Verlag, Berlin, 1987.

Bingyan Wang, Yuling Yan, and Jianqing Fan. Sample-efficient reinforcement learning for linearly-parameterized MDPs with a generative model. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=J2YvvXDp7H.

Yuanhao Wang, Qinghua Liu, Yu Bai, and Chi Jin. Breaking the curse of multiagency: Provably efficient decentralized multi-agent rl with function approximation. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2793–2848. PMLR, 12–15 Jul 2023a. URL https://proceedings.mlr.press/v195/wang23b.html.

Yuanhao Wang, Qinghua Liu, Yu Bai, and Chi Jin. Breaking the curse of multiagency: Provably efficient decentralized multi-agent rl with function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 2793–2848. PMLR, 2023b.

Zifan Wang, Yi Shen, Michael M. Zavlanos, and Karl H. Johansson. Learning of nash equilibria in risk-averse games. In *2024 American Control Conference (ACC)*, pp. 3270–3275, 2024. DOI: 10.23919/ACC60939.2024.10644891.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6995–7004. PMLR, 2019. URL http://proceedings.mlr.press/v97/yang19b.html.

Ali Yekkehkhany, Timothy Murray, and Rakesh Nagi. Risk-averse equilibrium for games, 2020. URL https://arxiv.org/abs/2002.08414.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1954–1964. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/zanette20a.html.

Kaiqing Zhang, TAO SUN, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. Robust multi-agent reinforcement learning with model uncertainty. In H. Larochelle,

M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 10571–10583. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/774412967f19ea61d448977ad9749078-Paper.pdf.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. *Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms*, pp. 321–384. Springer International Publishing, Cham, 2021. ISBN 978-3-030-60990-0. DOI: 10.1007/978-3-030-60990-0_12. URL https://doi.org/10.1007/978-3-030-60990-0_12.

Yizhou Zhang and Eric Mazumdar. Convergent q-learning for infinite-horizon general-sum markov games through behavioral economics, 2025. URL https://arxiv.org/abs/2508.08669.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly Minimax Optimal Reinforcement Learning for Linear Mixture Markov Decision Processes. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4532–4576. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/zhou21a.html.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Extended Discussion & Related Work

In this section we provide a detailed discussion of related literature. Our work studies the Risk-Sensitive Quantal Response Equilibrium (RQRE) introduced by Mazumdar et al. (2025), which combines risk aversion with bounded rationality to yield a class of equilibria that is computationally tractable via no-regret learning in *all* finite-action, finite-horizon Markov games, provided agents exhibit a sufficient degree of risk aversion and bounded rationality. Zhang & Mazumdar (2025) extended these results to discounted infinite-horizon Markov games, proving contraction of the risk-averse quantal response Bellman operator and convergence of a Q-learning algorithm under monotonicity conditions.

In this work we build on this foundation by providing the first finite-sample regret guarantees for learning RQRE with linear function approximation, enabling scalability to large or continuous state spaces. Beyond this extension to function approximation, our analysis reveals an explicit dependence of the regret on the rationality and risk-sensitivity parameters, establishes that RQRE strictly generalizes existing distributionally robust equilibrium concepts, and proves Lipschitz stability of the RQRE policy map—a property that Nash equilibria provably lack—which we leverage to obtain policy convergence guarantees under approximation error.

Nash equilibrium methods in Markov games. The standard solution concept when studying multi-agent interactions from a game-theoretic perspective is the Nash equilibrium (Nash, 1950): a joint policy from which no agent can improve their outcome by unilaterally deviating. In a seminal work, Hu & Wellman (2003) introduced *Nash Q-learning*, the first extension of single-agent Q-learning to general-sum stochastic games. The core mechanism involves agents maintaining Q-functions over joint actions and performing updates based on the assumption that all participants will adopt Nash equilibrium behavior in the *stage games*. However, this original formulation provides only asymptotic guarantees. Building on this, Liu et al. (2021) provided the first finite-sample guarantees by incorporating optimism into the online learning process, an exploration bonus term that allows agents to efficiently learn an accurate model of the environment from a limited number of interactions. Despite these advancements, both works are restricted to the tabular setting, whose computational and memory costs scale prohibitively with the size of the state space. Moreover, computing Nash equilibria in general-sum games is computationally intractable in general (Daskalakis, 2013). This intractability has motivated alternative solution concepts such as correlated equilibria (Aumann, 1974) and coarse correlated equilibria (Moulin & Vial, 1978), but these rely on a shared correlation device and lack individual rationalizability, meaning each agent's learned policy carries no meaningful optimality guarantees unless all agents jointly follow the correlated recommendation. To address the scalability issues of tabular methods, Cisneros-Velarde & Koyejo (2023) proposed Nash Q-learning with Optimistic Value Iteration (NQOVI), extending the framework of Hu & Wellman (2003) to the linear function approximation regime and providing finite-sample regret bounds that scale with feature dimension rather than the size of the state space. As discussed in Section 1 & 4, a fundamental bottleneck persists: the algorithm requires solving for a Nash equilibrium at every stage game and every episode, inheriting both computational intractability and the instability of the Nash correspondence under payoff perturbations (Example 2). In the function approximation setting, where Q-values are necessarily estimated with error, this brittleness is especially problematic.

Bounded Rationality and Risk Aversion. Bounded rationality has a long history in behavioral game theory, where it was introduced to better predict the strategic behavior of human decision-makers known to be imperfect optimizers (McKelvey & Palfrey, 1995; Erev & Roth, 1998; Goeree & Holt, 1999). The standard formalization replaces exact best responses with *quantal responses*—stochastic policies that assign higher probability to higher payoff actions. In the context of multi-

agent reinforcement learning, this modeling choice offers several important benefits. First, it provides a principled equilibrium selection since entropy-regularized response functions yield a unique, smooth equilibrium for any finite game, resolving the multiplicity and discontinuity issues inherent to Nash equilibrium (McKelvey & Palfrey, 1995). Second, bounded rationality reduces sensitivity to the approximation errors and non-stationarity unavoidable in learning settings, producing policies that trade off optimality against stability. Third, this formulation aligns naturally with modern reinforcement learning, where entropy-regularized objectives are widely employed as components of coherent decision-theoretic frameworks (Haarnoja et al., 2018; Sokota et al., 2023; Mertikopoulos & Sandholm, 2016).

Incorporating risk aversion addresses fundamental limitations of purely expected utility maximizing agents. Risk-sensitive objectives provide a principled mechanism for promoting safety and reliability by discouraging policies that achieve high average performance at the cost of rare but catastrophic outcomes, a failure mode that is especially pronounced in strategic, non-stationary environments. In game-theoretic settings, risk aversion acts as a form of robustification: by penalizing outcome variability, agents become less sensitive to modeling error, finite-sample noise, and strategic misspecification of opponents (Mazumdar et al., 2025; Lanzetti et al., 2025). This can further stabilize learning by biasing agents towards equilibria with lower variance and more predictable joint behavior, ultimately improving generalization across opponents and environments.

We refer the reader to Mazumdar et al. (2025) who provide a thorough exposition and history of these concepts, and their use in Markov games.

Risk sensitive multi-agent reinforcement learning. Our work leverages risk-aversion, where policies prefer actions with more certain outcomes in exchange for potentially lower expected return. Risk sensitivity in decision-making has its origins in Markov decision processes (Howard & Matheson, 1972), dynamic programming (Ruszczyński, 2010), and optimal control (Hernández-Hernández & Marcus, 1996), with a subsequent line of work integrating risk sensitivity into reinforcement learning (Shen et al., 2014; Heger, 1994; Mazumdar et al., 2017; Ratliff & Mazumdar, 2019; Singh et al., 2018; Lacotte et al., 2018). This work focuses on risk in *multi-agent reinforcement learning*. From a theoretical point of view, most existing works differ in the source of risk they address and make varying structural assumptions. Wang et al. (2024) considers risk-aversion to randomness in the stochastic payoffs via the conditional value at risk and requires strong monotonicity of the risk-averse game to guarantee convergence to a unique Nash equilibrium. Yekkehkhany et al. (2020) studies risk-aversion to the randomness induced by a population of mixed strategies of other agents in stochastic games, proving existence of equilibria in all finite games. Slumbers et al. (2023) considers risk-aversion to opponents' strategies by regularizing each agent's utility with the variance induced by the other agents' actions, and proves existence of their equilibrium concept, but the resulting formulation lacks tractability for general games. Lanzetti et al. (2025) takes a complementary approach, proposing a strategically robust equilibrium in which each agent optimizes against the worst-case opponent behavior within an ambiguity set, defined using optimal transport, and shows that such equilibria exist under the same assumptions as Nash equilibria and interpolate between Nash and maximin strategies. While these works focus on theoretical guarantees, recent empirical studies suggest risk-sensitive techniques extend to function approximators such as neural networks (Qiu et al., 2021; Shen et al., 2023). A related but distinct line of work is robust multi-agent reinforcement learning (Shi et al., 2024; Zhang et al., 2020; McMahan et al., 2024), which primarily addresses environmental uncertainty and worst-case parameters.

Linear function approximation in RL and Markov games. Foundational work by Tsitsiklis & Van Roy (1996) established asymptotic convergence guarantees in reinforcement learning algorithms with linear function approximation. Early results in providing *finite-sample* guarantees with linear function approximation was provided by Munos & Szepesvári (2008) studying fitted-value iteration under the assumption of access to a generative model of the environment. More recently, Jin et al. (2020) provided a provably sample efficient algorithm for linear MDPs in an online setting using optimism. Zhou et al. (2021) further refined these results, obtaining nearly minimax-optimal

guarantees. Subsequently, linear function approximation has been extensively studied in the single-agent setting (Yang & Wang, 2019; Wang et al., 2021; Eaton et al., 2025; Zanette et al., 2020). Although less explored, extending these approaches to Markov games has attracted growing attention. The first line of works accomplished this in the two-player zero-sum setting (Chen et al., 2022; Qiu et al., 2022). Most recently, concurrent works Wang et al. (2023a) and Cui et al. (2023) introduced independent linear function approximation to simultaneously address large state spaces while breaking the curse of multi-agents, achieving sample complexities polynomial in the number of agents rather than exponential. Dai et al. (2024) further refined these results with improved convergence rates. Most relevant to our work, Cisneros-Velarde & Koyejo (2023) extended Nash Q-learning (Hu & Wellman, 2003) to the linear function approximation regime, providing finite-sample regret bounds that scale with the feature dimension rather than the size of the state space. However, their algorithm requires solving for a Nash equilibrium at every stage game, inheriting both the computational intractability of Nash in general-sum games and its brittleness under payoff perturbations (as we show in Example 2). Moreover, Nash equilibria are not guaranteed to be unique in general-sum games, introducing an equilibrium selection problem that is particularly problematic under function approximation, where Q-values are necessarily estimated with error. In contrast, RQRE yields a unique, smooth equilibrium that is Lipschitz continuous in estimated payoffs (Corollary 2), and is individually rationalizable in the sense that each agent's learned Q-function carries meaningful optimality guarantees independent of whether other agents follow the equilibrium. Beyond the choice of equilibrium concept, our analysis introduces new technical challenges not present in Cisneros-Velarde & Koyejo (2023): we must bound the equilibrium approximation error incurred by computing an approximate RQRE at each stage in terms of the risk-aversion and bounded rationality parameters, and account for the additional approximation error introduced by the risk operators.

B RQRE-OVI Regret Bounds

In this appendix section, we prove the main regret results (Theorem 2, Corollary 1).

B.1 Proof Roadmap & Comparison to Prior Art

This section provides a high-level overview of the proof of Theorem 2 and highlights the key differences relative to prior analyses of optimistic value iteration in single-agent reinforcement learning and Nash Q-learning in Markov games (Cisneros-Velarde & Koyejo, 2023).

Regret definition. Recall that the regret is defined via the episode-wise *exploitability gap*:

$$\Delta_k = \max_{i\in[n]} \Big(V_{i,1}(\mathrm{sBR}_i(\pi^k_{-i}), \pi^k_{-i}; s_0) - V_{i,1}(\pi^k; s_0) \Big), \qquad \mathrm{regret}(K) = \sum_{k=1}^{K} \Delta_k. \tag{10}$$

Proof Sketch in Four Steps. The proof proceeds through four main steps.

Step 1: Optimistic Value Estimation under Risk Operators. The algorithm maintains a linear approximation of the action-value functions

$$Q^k_{i,h}(x,a) = \min\Big\{ \langle w^k_{i,h}, \phi(x,a,h)\rangle + \beta\sqrt{\phi(x,a,h)^\top (\Lambda^k_h)^{-1}\phi(x,a,h)},\ B \Big\}, \tag{11}$$

where the second term is the exploration bonus and Λ^k_h is the ridge regression covariance matrix.

The first key technical step establishes an *optimism property*. Lemma 2 shows that with high probability

$$Q^k_{i,h}(x,a) \geq r_{i,h}(x,a) + \rho^{\mathsf{env}}_{i,h}\big(V_{i,h+1}(\pi^k_{h+1};\cdot)\big) - \Big(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}})\Big). \tag{12}$$

Compared to the standard optimistic Bellman inequality used in risk-neutral RL, this bound includes additional approximation terms arising from (i) estimation error $\varepsilon_{\mathsf{env}}$ in the environment risk operator, (ii) approximation error $\varepsilon_{\mathsf{pol}}$ in policy-risk evaluation, (iii) error in the stage equilibrium solver $\varepsilon_{\mathsf{eq}}$. These errors propagate through the environment risk operator via its Lipschitz constant L_{env}.

Step 2: Uniform Concentration via Covering Numbers. The second step controls the statistical complexity of the induced value-function class. This is formalized in Lemma 1. In single-agent linear MDP analyses, concentration bounds often rely on self-normalized martingale inequalities that lead to a linear-bandit style argument. In contrast, the present setting requires controlling a much richer class of value functions.

Specifically, the regression targets depend on the composition

$$Q \;\mapsto\; \tilde{\pi}_h(\cdot \mid x; Q) \;\mapsto\; V_{i,h+1}(\cdot; \tilde{\pi}) \;\mapsto\; \rho^{\mathsf{e}}_{i,h}(V_{i,h+1}), \tag{13}$$

where $\tilde{\pi}_h$ is the approximate RQRE policy computed from Q.

Since the equilibrium policy depends on the current estimate Q, the induced value functions form a data-dependent function class. Controlling concentration therefore requires a uniform bound over this class.

To accomplish this, the proof constructs an ε-cover of the induced value-function class $\mathcal{V}^k_{i,h}$ and bounds its covering number $\mathcal{N}(\cdot)$. Indeed, since the equilibrium policy depends on the current estimate Q, the resulting value functions form a data-dependent class. Lemma 1 bounds the covering number of this class, yielding

$$\log \mathcal{N}(\epsilon, \mathcal{V}^k_{i,h}, \|\cdot\|_\infty) = \tilde{\mathcal{O}}(d^2) \tag{14}$$

reflecting the need to simultaneously control (i) the linear predictor w, (ii) the bonus term involving $(\Lambda^k_h)^{-1}$, and (iii) the equilibrium mapping $Q \mapsto \tilde{\pi}$. This bound determines the confidence radius β used in the optimistic value estimates. This larger covering number leads to a larger confidence radius β, which ultimately yields a regret bound with higher dependence on the feature dimension d than in the single-agent linear MDP case.

Step 3: Incorporating Equilibrium Approximation.

The third ingredient shows how the stage-game equilibrium solver error enters the regret bound. Lemma 3 proves that the value accuracy guarantee of the approximate RQRE solver translates into an additive error term in the exploitability gap.

Combining this lemma with the optimism property yields a bound on the per-episode exploitability gap in terms of the cumulative exploration bonuses along the trajectory plus an additive approximation budget.

Step 4: Performance Difference Decomposition & Sum of Bounes. Fix episode k and let

$$i^\star \in \arg\max_i \Big(V_{i,1}(\mathrm{sBR}_i(\pi^k_{-i}), \pi^k_{-i}; s_0) - V_{i,1}(\pi^k; s_0) \Big). \tag{15}$$

Define the deviation policy $\pi^{k,\mathrm{br}} = (\mathrm{sBR}_{i^\star}(\pi^k_{-i^\star}), \pi^k_{-i^\star})$. For each stage h, define $\Delta_{k,h}(x) = V_{i^\star,h}(\pi^{k,\mathrm{br}}; x) - V_{i^\star,h}(\pi^k; x)$. A stage-wise recursion shows

$$\Delta_{k,h}(x) \le \text{(advantage term)} + \mathbb{E}[\Delta_{k,h+1}(x_{h+1})]. \tag{16}$$

Using the optimism property and the approximate equilibrium guarantee, the advantage term can be bounded by

$$\beta \sqrt{\phi(x^k_h, a^k_h, h)^\top (\Lambda^k_h)^{-1} \phi(x^k_h, a^k_h, h)} + \varepsilon_{\mathsf{eq}}. \tag{17}$$

This yields a bound on the *exploitability gap* in terms of the cumulative bonuses along the trajectory.

Next, we bound the sum of bonuses. Indeed, summing the recursion over $h = 1, \ldots, H$ yields

$$\Delta_k \lesssim \sum_{h=1}^{H} \beta \sqrt{\phi(x^k_h, a^k_h, h)^\top (\Lambda^k_h)^{-1} \phi(x^k_h, a^k_h, h)} + H\big(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}})\big). \tag{18}$$

Summing over episodes and applying the elliptical potential lemma gives

$$\sum_{k,h} \sqrt{\phi(x_h^k, a_h^k, h)^\top (\Lambda_h^k)^{-1} \phi(x_h^k, a_h^k, h)} = \tilde{\mathcal{O}}(\sqrt{KdH}). \tag{19}$$

Final Bound Construction. The remaining step of the proof bounds the cumulative exploration bonuses along the trajectory. This is done using the standard elliptical potential argument commonly used in analyses of optimistic value iteration (Jin et al., 2020; Cisneros-Velarde & Koyejo, 2023). Indeed, combining this with the covering-number bound on β yields the final regret bound

$$\mathrm{regret}(K) \leq \tilde{\mathcal{O}}\big(L_{\mathsf{env}} B \sqrt{Kd^3H^3}\big) + KH\big(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}})\big). \tag{20}$$

Comparison to Nash Q-Learning. The closest related analysis is that of Cisneros-Velarde & Koyejo (2023), which studies optimistic value iteration for risk-neutral Markov games with linear function approximation.

Three key differences distinguish the present analysis:

1. **Explicit equilibrium approximation.** Their algorithm assumes access to a Nash equilibrium oracle for each stage game. In contrast, our algorithm computes an approximate RQRE which are known to be efficiently computable via no-regret algorithms, introducing an additional error term $\varepsilon_{\mathsf{eq}}$ that must be propagated through the dynamic programming recursion.
2. **Risk-sensitive Bellman operators.** The Bellman operators involve nonlinear risk operators acting on the continuation values, as opposed to risk-neutral expectations. As a result, the analysis must control the propagation of estimation and equilibrium errors through these operators, leading to the Lipschitz amplification factor L_{env} and additional approximation terms in the regret bound.
3. **Data-Dependent Risk-Induced Value Function Class.** Since the equilibrium policy is computed from the current value estimates, the continuation values depend on the data-dependent mapping $Q \mapsto \tilde{\pi}(Q)$, which then are transformed via the risk measure ρ^{e}. This requires constructing the appropriate vale function class that is induced by the risk operators. Controlling the resulting function class requires a covering-number argument for the induced value functions.

Together these features lead to a regret analysis that differs from the Nash-Q case (and single agent case) while yielding guarantees for an algorithm that is computationally realizable in practice.

Comparison to Single-Agent Linear MDP Analyses. In single-agent linear MDP analyses, the Bellman target depends only on a linear expectation of the next-stage value function. This structure enables sharper self-normalized concentration arguments and typically leads to regret bounds with smaller dimension dependence.

In contrast, the regression targets in the present setting involve the composition

$$Q \;\mapsto\; \tilde{\pi}_h(\cdot \mid x; Q) \;\mapsto\; V_{i,h+1}(\cdot; \tilde{\pi}) \;\mapsto\; \rho_{i,h}^{\mathsf{env}}(V_{i,h+1}), \tag{21}$$

where $\tilde{\pi}_h$ is the approximate RQRE policy computed from Q. Controlling concentration uniformly over this induced value class requires a covering-number argument, which leads to a larger confidence radius and the higher d-dependence appearing in the regret bound.

B.2 Technical Lemmas

In this section, we have all the technical lemmas for the main regret results (Theorem 2).

Covering Number Bounds for the Value Functions.

Lemma 1 (Covering number of the induced value class)**.** Fix (i, h) and an episode index k. Let $\Lambda_h^k \succeq \lambda I_d$ be the (random but fixed conditional on $\mathcal{F}_{k,h-1}$) design matrix used by Algorithm 1 at

stage h in episode k. Define the optimistic Q-class at stage h as follows:

$$\mathcal{Q}_{i,h}^k := \Big\{(x,a) \mapsto \min\Big\{\langle w, \phi(x,a,h)\rangle + \beta\sqrt{\phi(x,a,h)^\top(\Lambda_h^k)^{-1}\phi(x,a,h)},\ B\Big\} :\ \|w\|_2 \le B\Big\}.$$

Additionally, define the induced policy-evaluated value class as follows:

$$\mathcal{V}_{i,h}^k := \Big\{x \mapsto \widehat{V}_{i,h}^{\epsilon_i}(\pi; x; Q) :\ Q \in \mathcal{Q}_{i,h}^k,\ \pi(\cdot|x) \in \Delta(\mathcal{A})\Big\}.$$

Then for any $\varepsilon \in (0, B]$, under the metric $\mathrm{dist}_\infty(V, V') := \sup_x |V(x) - V'(x)|$,

$$\log \mathcal{N}\left(\varepsilon, \mathcal{V}_{i,h}^k, \mathrm{dist}_\infty\right) \le d \log\left(1 + \frac{4B}{\varepsilon}\right) + d^2 \log\left(1 + \frac{8\sqrt{d}\,\beta^2}{\lambda\,\varepsilon^2}\right). \tag{22}$$

Proof. We proceed in four steps.

Lipschitz composition reduces the problem to covering $\mathcal{Q}_{i,h}^k$. Fix any $Q, Q' \in \mathcal{Q}_{i,h}^k$ and any Markov π. First observe that for each fixed (i,h) and any Markov π, the map $Q \mapsto \widehat{V}_{i,h}^{\epsilon_i}(\pi; \cdot; Q)$ is 1-Lipschitz in sup-norm:

$$\sup_{x \in \mathcal{X}} \left|\widehat{V}_{i,h}^{\epsilon_i}(\pi; x; Q) - \widehat{V}_{i,h}^{\epsilon_i}(\pi; x; Q')\right| \le \|Q - Q'\|_\infty .$$

For the dual-representation-based $\widehat{V}$, this holds because it is a supremum of expectations of Q over a set of distributions plus an additive regularizer. By this 1-Lipschitz property, we have the bound

$$\mathrm{dist}_\infty\left(\widehat{V}_{i,h}^{\epsilon_i,\tau_i}(\pi; \cdot; Q),\ \widehat{V}_{i,h}^{\epsilon_i,\tau_i}(\pi; \cdot; Q')\right) \le \|Q - Q'\|_\infty .$$

Therefore, any ε-cover of $\mathcal{Q}_{i,h}^k$ in $\|\cdot\|_\infty$ pushes forward to an ε-cover of $\mathcal{V}_{i,h}^k$ in dist_∞, and hence

$$\mathcal{N}(\varepsilon, \mathcal{V}_{i,h}^k, \mathrm{dist}_\infty) \le \mathcal{N}(\varepsilon, \mathcal{Q}_{i,h}^k, \|\cdot\|_\infty).$$

So it suffices to bound the covering number of $\mathcal{Q}_{i,h}^k$.

Separate the parametric linear part and the bonus part. For each w define the unclipped function

$$f_w(x,a) := \langle w, \phi(x,a,h)\rangle + b_h^k(x,a), \qquad b_h^k(x,a) := \beta\sqrt{\phi(x,a,h)^\top(\Lambda_h^k)^{-1}\phi(x,a,h)}.$$

Let $T(u) := \min\{u, B\}$ be the clipping operator. Then $Q_w := T \circ f_w \in \mathcal{Q}_{i,h}^k$. Since T is 1-Lipschitz on $\mathbb{R}$, we have that

$$\|Q_w - Q_{w'}\|_\infty = \|T \circ f_w - T \circ f_{w'}\|_\infty \le \|f_w - f_{w'}\|_\infty .$$

Also, since b_h^k does not depend on w, we deduce that

$$\|f_w - f_{w'}\|_\infty = \sup_{x,a} |\langle w - w', \phi(x,a,h)\rangle| \le \|w - w'\|_2 \cdot \sup_{x,a} \|\phi(x,a,h)\|_2 \le \|w - w'\|_2 ,$$

using $\|\phi\| \le 1$. Thus, if we cover the Euclidean ball $\{w : \|w\|_2 \le B\}$ in ℓ_2 to radius ε, we automatically get an ε-cover of $\mathcal{Q}_{i,h}^k$ in $\|\cdot\|_\infty$ *provided the bonus term is fixed.*

However, because the proof must hold uniformly over all *possible* random matrices Λ_h^k that can arise adaptively along the algorithm's trajectory, we incorporate a standard operator-covering argument: we bound the number of distinguishable bonus shapes $\phi^\top(\Lambda_h^k)^{-1}\phi$ over all Λ in the admissible set.

Cover the linear predictors (the $d\log(1 + 4B/\varepsilon)$ term). Let $\mathcal{W} := \{w \in \mathbb{R}^d : \|w\|_2 \le B\}$. A standard volumetric bound for Euclidean balls gives an ε-net $\mathcal{W}_\varepsilon \subset \mathcal{W}$ of size at most

$$|\mathcal{W}_\varepsilon| \le \left(1 + \frac{2B}{\varepsilon}\right)^d \le \left(1 + \frac{4B}{\varepsilon}\right)^d ,$$

since $\varepsilon \leq B$ For any $w \in \mathcal{W}$, pick $\tilde{w} \in \mathcal{W}_\varepsilon$ with $\|w - \tilde{w}\|_2 \leq \varepsilon$. Then for all (x, a), we have that

$$|\langle w, \phi \rangle - \langle \tilde{w}, \phi \rangle| = |\langle w - \tilde{w}, \phi \rangle| \leq \|w - \tilde{w}\|_2 \|\phi\|_2 \leq \varepsilon.$$

Thus, if the bonus term were fixed, this would already provide an ε-cover.

Cover the bonus shapes over admissible Λ (the $d^2 \log(\cdot)$ term). Define the normalized feature vectors

$$u(x, a) := \phi(x, a, h) \in \mathbb{R}^d, \qquad \|u(x, a)\|_2 \leq 1.$$

For any positive semi-definite matrix $M \succeq 0$, define the quadratic form

$$q_M(u) := u^\top M u.$$

Here $M = (\Lambda_h^k)^{-1}$ and $\Lambda_h^k \succeq \lambda I_d$ implies

$$0 \preceq M \preceq \lambda^{-1} I_d, \qquad \|M\|_2 \leq \lambda^{-1}.$$

We need a finite set $\mathcal{M}_\eta$ such that for every admissible M there exists $\tilde{M} \in \mathcal{M}_\eta$ with

$$\sup_{\|u\| \leq 1} \left| u^\top M u - u^\top \tilde{M} u \right| \leq \eta.$$

Note that for symmetric $M - \tilde{M}$, the bound holds:

$$\sup_{\|u\| \leq 1} |u^\top (M - \tilde{M}) u| = \left\| M - \tilde{M} \right\|_2.$$

Thus it suffices to build an η-net in operator norm for the set $\{M : 0 \preceq M \preceq \lambda^{-1} I\}$. A standard covering bound for bounded subsets of $\mathbb{R}^{d \times d}$ gives an η-net in Frobenius norm of size at most

$$\left(1 + \frac{2R}{\eta}\right)^{d^2}, \quad R := \sup \|M\|_F \leq \sqrt{d}\, \lambda^{-1}.$$

Since $\|A\|_2 \leq \|A\|_F$, an η-net in Frobenius norm is also an η-net in operator norm. Therefore there exists $\mathcal{M}_\eta$ with

$$|\mathcal{M}_\eta| \leq \left(1 + \frac{2\sqrt{d}\, \lambda^{-1}}{\eta}\right)^{d^2}.$$

Next we connect η to the bonus scale. First, the bonus uses $\sqrt{u^\top M u}$. For any nonnegative scalars s, t, we have that

$$\left| \sqrt{s} - \sqrt{t} \right| \leq \sqrt{|s - t|}.$$

Hence, if $\sup_{\|u\| \leq 1} |u^\top M u - u^\top \tilde{M} u| \leq \eta$, then

$$\sup_{\|u\| \leq 1} \left| \sqrt{u^\top M u} - \sqrt{u^\top \tilde{M} u} \right| \leq \sqrt{\eta}.$$

Multiplying by β gives a bonus error at most $\beta\sqrt{\eta}$.

To make this bonus error at most ε, choose $\eta = (\varepsilon/\beta)^2$. Then

$$\log |\mathcal{M}_\eta| \leq d^2 \log \left(1 + \frac{2\sqrt{d}\, \lambda^{-1}}{(\varepsilon/\beta)^2}\right) = d^2 \log \left(1 + \frac{2\sqrt{d}\, \beta^2}{\lambda\, \varepsilon^2}\right).$$

Adjusting constants—replacing 2 by 8—to absorb the earlier ε-splitting and the clipping/Lipschitz steps gives

$$\log |\mathcal{M}_\eta| \leq d^2 \log \left(1 + \frac{8\sqrt{d}\, \beta^2}{\lambda\, \varepsilon^2}\right).$$

Combine nets. Take the Cartesian product net $\mathcal{W}_\varepsilon \times \mathcal{M}_{(\varepsilon/\beta)^2}$. For any admissible (w, M), choose $(\tilde{w}, \tilde{M})$ from the nets. Then the linear part incurs at most ε error in $\|\cdot\|_\infty$, and the bonus part incurs at most ε error in $\|\cdot\|_\infty$. By triangle inequality, the unclipped f differs by at most 2ε, and clipping is 1-Lipschitz so the clipped Q differs by at most 2ε. Replacing ε by $\varepsilon/2$ yields the stated bound (22). □

Optimism via bonus term and approximations.

Lemma 2 (Uniform self-normalized confidence and optimism)**.** Fix (i, h). Let (x_h^t, a_h^t) be the sequence of state-action pairs observed at stage h up to the current episode index $k-1$ (across episodes), and define

$$\Lambda_h^k := \lambda I_d + \sum_{t<k} \phi(x_h^t, a_h^t, h)\phi(x_h^t, a_h^t, h)^\top.$$

Define the empirical targets as in Algorithm 1:

$$y_{i,h}^t := r_{i,h}(x_h^t, a_h^t) + \widehat{\rho}_{i,h}^t\left(\widehat{V}_{i,h+1}^{\epsilon_i,\tau_i}(\pi_{h+1}^t;\cdot)\right),$$

and let $w_{i,h}^k$ be the ridge regression estimate

$$w_{i,h}^k := (\Lambda_h^k)^{-1}\sum_{t<k}\phi(x_h^t, a_h^t, h)\, y_{i,h}^t.$$

Let the optimistic estimate be

$$Q_{i,h}^k(x,a) := \min\left\{\left\langle w_{i,h}^k, \phi(x,a,h)\right\rangle + \beta\sqrt{\phi(x,a,h)^\top(\Lambda_h^k)^{-1}\phi(x,a,h)},\ B\right\}.$$

Then for $\beta = cB\sqrt{d^2H\,\iota}$ as in Theorem 2, with probability at least $1-\delta$ simultaneously for all (i, h, k, x, a),

$$Q_{i,h}^k(x,a) \geq r_{i,h}(x,a) + \rho_{i,h}^{\mathsf{env}}\left(V_{i,h+1}^{\epsilon_i}(\pi_{h+1}^k;\cdot)\right) - \left(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}})\right). \tag{23}$$

Proof. The proof has three parts: (1) express the true Bellman target in the linear form, (2) establish a uniform self-normalized confidence bound using Lemma 1, (3) translate the approximation errors and risk-Lipschitz errors into the additive term.

Throughout, fix (i, h) and condition on the filtration up to the start of stage h in episode k so that Λ_h^k is fixed.

Part 1: linear Bellman target under realizability. By Assumption (a), for any bounded measurable function $g : \mathcal{X} \to \mathbb{R}$,

$$\mathbb{E}_{x'\sim\mathcal{P}_h(\cdot|x,a)}[g(x')] = \int g(x')\mathcal{P}_h(dx'|x,a) = \left\langle \phi(x,a,h),\ \int g(x')\mu_h(x')dx'\right\rangle.$$

Define the vector

$$\zeta_h(g) := \int g(x')\mu_h(x')\,dx' \in \mathbb{R}^d.$$

Then the *expected* one-step (non-risk) backup is linear in ϕ:

$$r_{i,h}(x,a) + \mathbb{E}[g(x')|x,a] = \left\langle \phi(x,a,h), \theta_{i,h} + \zeta_h(g)\right\rangle.$$

Now the environment risk operator $\rho_{i,h}^{\mathsf{env}}$ is applied to the random next-state value $g(x_{h+1})$ (or an equivalent lifted representation). We do not need $\rho_{i,h}^{\mathsf{env}}$ itself to be linear; we only need concentration for the scalar targets and Lipschitzness to control errors, which we do below.

Part 2: self-normalized confidence uniformly over the induced value class. Define the (ideal) target at time t:

$$\bar{y}_{i,h}^t := r_{i,h}(x_h^t, a_h^t) + \rho_{i,h}^{\mathsf{env}}\left(V_{i,h+1}^{\epsilon_i,\tau_i}(\pi_{h+1}^t;\cdot)\right).$$

Define the target noise:

$$\eta_t := y_{i,h}^t - \bar{y}_{i,h}^t.$$

We will show η_t is uniformly bounded and forms a martingale difference sequence, then apply the self-normalized inequality with a union bound over a cover.

First, by assumption, the rewards lie in $[0,1]$ and all value functions are in $[0,B]$. The monotonicity plus Lipschitz continuity assumption imply $\rho_{i,h}^{\mathsf{env}}(X)$ is bounded whenever X is bounded; in particular, there is a constant—absorbed into B by clipping—such that $|\rho_{i,h}^{\mathsf{env}}(V)| \le B$ when $0 \le V \le B$. Similarly, by assumption, we have that $\widehat{\rho}_{i,h}^t(V) \le \rho_{i,h}^{\mathsf{env}}(V)$. Thus both $\bar{y}_{i,h}^t$ and $y_{i,h}^t$ lie in an interval of length $O(B)$, hence $|\eta_t| \le B$ (up to an absolute constant factor absorbed into c).

Second, η_t is measurable with respect to the randomness at episode/stage t and satisfies the standard conditional mean-zero property needed for the self-normalized bound once we condition on the sigma-field generated by all past data: the only randomness in η_t is through the empirical dual approximation and policy approximation, both of which are treated as bounded perturbations; hence we can apply a bounded-differences martingale concentration. Concretely, we use the standard self-normalized bound stated as: for any fixed $\delta' \in (0,1)$, with probability at least $1-\delta'$,

$$\forall x, a: \quad \left|\left\langle w_{i,h}^k - w_{i,h}^\star, \phi(x,a,h)\right\rangle\right| \le \beta\sqrt{\phi(x,a,h)^\top(\Lambda_h^k)^{-1}\phi(x,a,h)}, \tag{24}$$

where $w_{i,h}^\star$ is the (unknown) linear parameter corresponding to the ideal targets $\bar{y}_{i,h}^t$ and β scales as $B\sqrt{d\log(\cdot)}$.

To make (24) hold *uniformly* over all time indices and the induced value class (because the targets depend on $\widehat{V}$ which depends on past data), we apply a union bound over an ε-net of the induced value class. Lemma 1 gives the required finite cover size, which yields the logarithmic term

$$\iota = \log\left(\frac{nHK}{\delta}\right) + d\log\left(1+\frac{K}{\lambda}\right)$$

appearing in the theorem statement. Choosing $\beta = cB\sqrt{d^2H\,\iota}$ (absorbing all absolute constants) ensures (24) holds simultaneously for all (i,h,k,x,a) with probability at least $1-\delta$.

Part 3: translate approximation errors and derive optimism. By construction, we have that $Q_{i,h}^k(x,a) \ge \left\langle w_{i,h}^k, \phi(x,a,h)\right\rangle$ and

$$Q_{i,h}^k(x,a) \ge \left\langle w_{i,h}^k, \phi(x,a,h)\right\rangle + \beta\sqrt{\phi^\top(\Lambda_h^k)^{-1}\phi} - (U(x,a)-B)_+,$$

where $U(x,a) := \langle w_{i,h}^k, \phi(x,a,h)\rangle + \beta\sqrt{\phi^\top(\Lambda_h^k)^{-1}\phi}$ and $(z)_+ := \min\{z,0\}$ as usual. Since clipping only decreases values above B and all true quantities are within $[0,B]$ by assumption, clipping does not harm the lower bound we are establishing.

Using (24), we have that

$$\left\langle w_{i,h}^k, \phi(x,a,h)\right\rangle + \beta\sqrt{\phi^\top(\Lambda_h^k)^{-1}\phi} \ge \left\langle w_{i,h}^\star, \phi(x,a,h)\right\rangle.$$

But $\langle w_{i,h}^\star, \phi(x,a,h)\rangle$ is the linear prediction of the *ideal target* $\bar{y}$ at (x,a), i.e.

$$\left\langle w_{i,h}^\star, \phi(x,a,h)\right\rangle = r_{i,h}(x,a) + \rho_{i,h}^{\mathsf{env}}\left(V_{i,h+1}^{\epsilon_i,\tau_i}(\pi_{h+1}^k;\cdot)\right),$$

by definition of $w_{i,h}^\star$ for the ideal target. It remains to incorporate the approximation errors (d,e,f) to relate the algorithm's *implemented* target to the ideal target.

The assumption on the risk measure gives us that for any bounded X, the bound holds:

$$0 \leq \rho^{\mathsf{env}}_{i,h}(X) - \widehat{\rho}^{k}_{i,h}(X) \leq \varepsilon_{\mathsf{env}}.$$

Hence, we deduce that

$$\rho^{\mathsf{env}}_{i,h}(X) \geq \widehat{\rho}^{k}_{i,h}(X).$$

The bounds on the value function imply that

$$0 \leq V^{\epsilon_i}_{i,h+1}(\pi; x) - \widehat{V}^{\epsilon_i}_{i,h+1}(\pi; x) \leq \varepsilon_{\mathsf{pol}},$$

so by L_{env}-Lipschitzness of $\rho^{\mathsf{env}}_{i,h}$,

$$\left|\rho^{\mathsf{env}}_{i,h}(V^{\epsilon_i}_{i,h+1}(\pi;\cdot)) - \rho^{\mathsf{env}}_{i,h}(\widehat{V}^{\epsilon_i}_{i,h+1}(\pi;\cdot))\right| \leq L_{\mathsf{env}}\varepsilon_{\mathsf{pol}}.$$

Finally, Assumption (f) (Model A) says that the stage equilibrium computation error changes the policy-evaluated value by at most $\varepsilon_{\mathsf{eq}}$; again, when this error is propagated through the environment operator, it contributes at most $L_{\mathsf{env}}\varepsilon_{\mathsf{eq}}$.

Collecting these additive losses yields exactly (23). □

Technical Lemmas on Equilibrium Approximation. Fix an episode k, stage h, and the realized state $x := x^k_h$. Let Q^k_h denote the optimistic stage Q-table at (h, x) for all players. Let $\tilde{\pi}_h(\cdot \mid x)$ be the policy returned by the stage solver, and let $\pi^\star_h(\cdot \mid x)$ denote the exact stage equilibrium policy for the same stage game induced by Q^k_h.

Lemma 3 (Using the stage-solver accuracy in the regret proof)**.** Assume the stage solver satisfies the *value-accuracy* condition from Theorem 2: for every player $i \in [n]$,

$$\left|V^{\epsilon_i}_{i,h}\big(\tilde{\pi}_h; x, Q^k_h\big) - V^{\epsilon_i}_{i,h}\big(\pi^\star_h; x, Q^k_h\big)\right| \;\leq\; \varepsilon_{\mathsf{eq}}. \tag{25}$$

Further assume the policy-risk value functional $V^{\epsilon_i}_{i,h}(\cdot; x, Q)$ is *1-Lipschitz in the payoff table* in the following sense: for any two payoff tables Q, Q' on (h, x) and any fixed policy profile π,

$$\left|V^{\epsilon_i}_{i,h}\big(\pi; x, Q\big) - V^{\epsilon_i}_{i,h}\big(\pi; x, Q'\big)\right| \leq \sup_{a \in \mathcal{A}} \left|Q_i(x, a) - Q'_i(x, a)\right|. \tag{26}$$

Let $a^k_h \sim \tilde{\pi}_h(\cdot \mid x)$ be the action actually played by the algorithm at (h, x), and let $a^{k,\mathrm{br}}_h$ denote a (possibly randomized) stage-wise best-response action for the distinguished player $i^\star$ against $\tilde{\pi}_{-i^\star,h}(\cdot \mid x)$ with respect to the *true* stage payoff table $\bar{Q}^k_h$ (defined below). Suppose the optimistic table admits the usual decomposition

$$Q^k_{i,h}(x, a) = \bar{Q}^k_{i,h}(x, a) + b^k_h(x, a), \qquad b^k_h(x, a) := \beta\sqrt{\phi(x, a, h)^\top (\Lambda^k_h)^{-1}\phi(x, a, h)}, \tag{27}$$

with $b^k_h(x, a) \geq 0$ for all a. Then the following bound holds:

$$\mathbb{E}\left[Q^k_{i^\star,h}(x, a^{k,\mathrm{br}}_h) - Q^k_{i^\star,h}(x, a^k_h)\right] \;\leq\; \mathbb{E}\left[b^k_h(x, a^k_h)\right] + \varepsilon_{\mathsf{eq}}, \tag{28}$$

where the expectation is over the randomness in the stage solver (if any) and the action sampling.

Proof. The proof proceeds in two parts: **(part 1)** we first show the best-response improvement is controlled by the stage equilibrium error, and then **(part 2)** we relate the deviation gap to an optimistic Q-advantage plus a bonus term.

Part 1: Best-response improvement is controlled by the stage equilibrium error. By definition of $\pi_h^\star(\cdot \mid x)$ as an *exact* stage equilibrium for the game induced by Q_h^k and the value functional $V_{i,h}^{\epsilon_i}(\cdot; x, Q_h^k)$, player $i^\star$ cannot improve its (policy-risk) value by a unilateral deviation at (h, x):

$$V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \pi^\star_{-i^\star,h}); x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big(\pi_h^\star; x, Q_h^k\big) \;\le\; 0, \qquad \forall \sigma_{i^\star} \in \Delta(\mathcal{A}_{i^\star}). \tag{29}$$

Now evaluate the deviation $\sigma_{i^\star}$ that achieves the best response against $\tilde{\pi}_{-i^\star,h}(\cdot \mid x)$—i.e., the deviation that generates $a_h^{k,\mathrm{br}}$. Add and subtract $V_{i^\star,h}^{\epsilon_{i^\star}}(\tilde{\pi}_h; x, Q_h^k)$ and $V_{i^\star,h}^{\epsilon_{i^\star}}(\pi_h^\star; x, Q_h^k)$ to get that

$$\begin{aligned}
&V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \tilde{\pi}_{-i^\star,h}); x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big(\tilde{\pi}_h; x, Q_h^k\big) \\
&= \big(V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \tilde{\pi}_{-i^\star,h}); x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \pi^\star_{-i^\star,h}); x, Q_h^k\big)\big) && (30)\\
&\quad + \big(V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \pi^\star_{-i^\star,h}); x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big(\pi_h^\star; x, Q_h^k\big)\big) && (31)\\
&\qquad + \Big(V_{i^\star,h}^{\epsilon_{i^\star}}\big(\pi_h^\star; x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big(\tilde{\pi}_h; x, Q_h^k\big)\Big). && (32)
\end{aligned}$$

Observe that (31) is less than or equal to zero by (29), and (32) is less than or equal to $\varepsilon_{\mathsf{eq}}$ by (25). Thus, we deduce that

$$\begin{aligned}
&V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \tilde{\pi}_{-i^\star,h}); x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big(\tilde{\pi}_h; x, Q_h^k\big) \\
&\quad \le V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \tilde{\pi}_{-i^\star,h}); x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \pi^\star_{-i^\star,h}); x, Q_h^k\big) + \varepsilon_{\mathsf{eq}}.
\end{aligned} \tag{33}$$

Step 2: Relating the deviation gap to an optimistic Q-advantage plus bonus. The regret proof uses the standard "optimism" relation that the optimistic table Q_h^k can be written as the (model-based / fitted) estimate $\bar{Q}_h^k$ plus the confidence bonus b_h^k, as in (27). Since $V_{i,h}^{\epsilon_i}(\cdot; x, Q)$ is 1-Lipschitz in Q (cf. (26)), swapping $\bar{Q}_h^k$ for Q_h^k changes any value by at most $\sup_a b_h^k(x, a)$, and in particular for the realized action a_h^k we obtain the pointwise bound

$$\begin{aligned}
&V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \tilde{\pi}_{-i^\star,h}); x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big(\tilde{\pi}_h; x, Q_h^k\big) \\
&\qquad \ge \mathbb{E}\left[\bar{Q}_{i^\star,h}^k(x, a_h^{k,\mathrm{br}}) - \bar{Q}_{i^\star,h}^k(x, a_h^k)\right] - \mathbb{E}\left[b_h^k(x, a_h^k)\right],
\end{aligned} \tag{34}$$

where the expectation is with respect to the sampling of $a_h^k \sim \tilde{\pi}_h(\cdot \mid x)$ and $a_h^{k,\mathrm{br}}$ induced by $\sigma_{i^\star}$ and $\tilde{\pi}_{-i^\star,h}(\cdot \mid x)$.

Now we combine (33) and (34), and use $Q_h^k = \bar{Q}_h^k + b_h^k$ from (27) to convert back to Q_h^k as follows:

$$\begin{aligned}
&\mathbb{E}\left[Q_{i^\star,h}^k(x, a_h^{k,\mathrm{br}}) - Q_{i^\star,h}^k(x, a_h^k)\right] \\
&\quad = \mathbb{E}\left[\bar{Q}_{i^\star,h}^k(x, a_h^{k,\mathrm{br}}) - \bar{Q}_{i^\star,h}^k(x, a_h^k)\right] + \mathbb{E}\left[b_h^k(x, a_h^{k,\mathrm{br}}) - b_h^k(x, a_h^k)\right] \\
&\quad \le \mathbb{E}\left[\bar{Q}_{i^\star,h}^k(x, a_h^{k,\mathrm{br}}) - \bar{Q}_{i^\star,h}^k(x, a_h^k)\right] + \mathbb{E}\left[b_h^k(x, a_h^{k,\mathrm{br}})\right] \\
&\quad \le \Big(V_{i^\star,h}^{\epsilon_{i^\star}}\big((\sigma_{i^\star}, \tilde{\pi}_{-i^\star,h}); x, Q_h^k\big) - V_{i^\star,h}^{\epsilon_{i^\star}}\big(\tilde{\pi}_h; x, Q_h^k\big)\Big) + \mathbb{E}\left[b_h^k(x, a_h^k)\right] \\
&\quad \le \mathbb{E}\left[b_h^k(x, a_h^k)\right] + \varepsilon_{\mathsf{eq}},
\end{aligned}$$

where the last inequality uses (33) and drops the term corresponding to (30), which is non-positive when $\sigma_{i^\star}$ is chosen as the stage-wise best response induced by the equilibrium operator, and the stage game is evaluated at the same optimistic table Q_h^k; this is precisely the sense in which the solver returns an approximate stage equilibrium for the optimistic table. This yields (28). □

Remark 2 (How $\varepsilon_{\mathsf{eq}}$ scales under common stage solvers)**.** The theorem assumes a *value-accuracy* guarantee of the form (25). How $\varepsilon_{\mathsf{eq}}$ depends on the number of inner iterations T depends on the solver and on structural properties of the stage game.

(i) Strongly monotone VI (Mirror-Prox / extragradient). Suppose the stage equilibrium (RQRE) can be written as the unique solution of a variational inequality $\mathrm{VI}(F, \Pi)$ over the product simplex $\Pi := \prod_i \Delta(\mathcal{A}_i)$, and that the operator F is μ-strongly monotone and L-Lipschitz (in a fixed norm). Then standard results for Mirror-Prox/extragradient yield a geometric decay of the VI gap (and hence of the value residual):

$$\varepsilon_{\mathsf{eq}}(T) \le C_0 \exp\left(-\frac{T\mu}{L}\right),$$

for a constant C_0 depending on initialization. In entropy-regularized settings, μ typically scales with the regularization level (e.g., $\mu \asymp 1/\epsilon_{\max}$ under the common "$\frac{1}{\epsilon}\Phi$" convention which we adopt), while L depends on the smoothness of the payoff map and, in entropic-risk models, on τ through the curvature of the log-sum-exp operator. Consequently, the condition number $\kappa := L/\mu$, and thus the required T, inherits explicit τ-dependence.

Indeed, let us specialize to see the dependence. Specialize to the entropic policy-risk value (7) and entropy regularization $\nu_i(\pi_i) = \Phi(\pi_i)$. Fix a stage (h, x) and write, for each player i,

$$u_i^\pi(a_{-i}) \;:=\; \sum_{a_i} \pi_i(a_i)\, Q_i(x, a_i, a_{-i}), \;\; g_i(\pi) \;:=\; -\frac{1}{\tau_i} \log\Big(\sum_{a_{-i}} \pi_{-i}(a_{-i}) e^{-\tau_i u_i^\pi(a_{-i})}\Big) + \frac{1}{\epsilon_i}\Phi(\pi_i).$$

Let the stage RQRE be characterized as the unique solution of the (regularized) VI over $\Pi := \prod_i \Delta(\mathcal{A}_i)$ associated with the first-order optimality conditions of $\{\min_{\pi_i} g_i(\pi)\}_{i=1}^n$ (equivalently, a strongly monotone operator F on Π).

Assume $Q_i(x, a) \in [0, B]$ for all i and a. Then:

(a) **Strong monotonicity.** The entropy regularizer contributes strong convexity, yielding strong monotonicity modulus

$$\mu \ge \frac{1}{\epsilon_{\max}}, \qquad \epsilon_{\max} := \max_{i \in [n]} \epsilon_i,$$

with respect to the ℓ_1 geometry on each simplex.[2]

(b) **Lipschitzness.** The entropic-risk term is a log-sum-exp smoothing of a linear map $\pi_i \mapsto u_i^\pi(\cdot)$. Its gradient w.r.t. π_i is a convex combination of payoff vectors and therefore satisfies the uniform ℓ_∞ bound

$$\|\nabla_{\pi_i} g_i(\pi)\|_\infty \;\le\; B,$$

and, moreover, its Jacobian with respect to (π_i, π_{-i}) is Lipschitz with constant scaling as

$$L \le c_L \Big(B\,\tau_{\max} \;+\; \frac{B}{\tau_{\min}}\Big), \qquad \tau_{\min} := \min_i \tau_i, \quad \tau_{\max} := \max_i \tau_i,$$

for an absolute constant $c_L > 0$. The $\frac{1}{\tau_{\min}}$ term comes from the curvature of the log-sum-exp map and dominates as $\tau_{\min} \to 0$.[3]

Consequently, the condition number $\kappa := L/\mu$ satisfies

$$\kappa \le c_L\, \epsilon_{\max} \Big(B\,\tau_{\max} + \frac{B}{\tau_{\min}}\Big),$$

and Mirror-Prox/extragradient yields the geometric solver error bound

$$\varepsilon_{\mathsf{eq}}(T) \le C_0 \exp\left(-\frac{T}{\kappa}\right) \le C_0 \exp\left(-\frac{T}{c_L\, \epsilon_{\max}\big(B\tau_{\max} + B/\tau_{\min}\big)}\right).$$

[2] More precisely, $-\Phi(\cdot)$ is 1-strongly convex with respect to $\|\cdot\|_1$, so the term $(1/\epsilon_i)\Phi(\pi_i)$ induces modulus $1/\epsilon_i$; taking the worst case gives $1/\epsilon_{\max}$.

[3] One convenient way to see the τ-dependence is to use that the Hessian of $\tau^{-1} \log \sum_j e^{\tau z_j}$ has operator norm $\le \tau$ (with respect to ℓ_∞/ℓ_1), while the induced sensitivity of the soft minimization weights with respect to π_{-i} contributes a τ^{-1} factor when π_{-i} enters inside the log.

In particular, holding all else fixed, achieving a target $\varepsilon_{\mathsf{eq}}$ requires

$$T \gtrsim \epsilon_{\max}\Big(B\tau_{\max} + \frac{B}{\tau_{\min}}\Big)\log\Big(\frac{C_0}{\varepsilon_{\mathsf{eq}}}\Big),$$

so the number of inner iterations grows on the order of $1/\tau_{\min}$ as $\tau_{\min} \to 0$.

(ii) Vanilla No-Regret. Fix a stage (h, x) and a bounded action–value function Q. Define the induced finite stage payoff

$$u_i(\pi) \;:=\; V_{i,h}^{\epsilon_i}(\pi; x, Q), \qquad i \in [n].$$

Assume that for all i and all π,

$$0 \le u_i(\pi) \le U$$

for some constant $U > 0$. Consider the associated lifted $2n$-player game $\widetilde{\mathcal{G}}(h, x, Q)$ as in Mazumdar et al. (2025). Suppose each player in the lifted game runs an external-regret algorithm for T rounds, producing joint play (π^t, p^t) for $t = 1, \dots, T$, and let

$$\sigma_T \;:=\; \frac{1}{T}\sum_{t=1}^{T} \delta_{(\pi^t, p^t)}$$

denote the empirical distribution.

Then σ_T is an $\varepsilon_{\mathrm{CCE}}(T)$-coarse correlated equilibrium of the lifted game, where

$$\varepsilon_{\mathrm{CCE}}(T) \;\le\; \frac{1}{T}\sum_{j=1}^{2n} \operatorname*{regret}_j(T).$$

In particular, if each player uses Hedge (or any algorithm with regret bound $\operatorname{regret}_j(T) \le U\sqrt{2T\log|\mathcal{A}_j|}$), then

$$\varepsilon_{\mathrm{CCE}}(T) \;\le\; U\sum_{j=1}^{2n}\sqrt{\frac{2\log|\mathcal{A}_j|}{T}}.$$

Under the equilibrium-collapse result of Mazumdar et al. (2025), the induced policy $\tilde{\pi}_h(\cdot \mid x; Q)$ obtained from σ_T is an $\varepsilon_{\mathrm{CCE}}(T)$-approximate stage RQRE of the original game. Consequently, for every $i \in [n]$,

$$\Big|V_{i,h}^{\epsilon_i}\big(\tilde{\pi}_h; x, Q\big) - V_{i,h}^{\epsilon_i}\big(\pi_h^\star; x, Q\big)\Big| \;\le\; \varepsilon_{\mathsf{eq}},$$

with the specialization

$$\varepsilon_{\mathsf{eq}} \;:=\; \varepsilon_{\mathrm{CCE}}(T).$$

In particular, to ensure $\varepsilon_{\mathsf{eq}} \le \varepsilon$, it suffices to take

$$T \gtrsim \frac{U^2}{\varepsilon^2}\Big(\sum_{j=1}^{2n}\sqrt{\log|\mathcal{A}_j|}\Big)^2,$$

up to universal constants.

B.3 Proof of Theorem 2

Now with the main technical lemmas in place, we prove the regret bound.

Proof of Theorem 2. Fix the high-probability event $\mathcal{E}$ on which Lemma 2 holds simultaneously for all (i, h, k, x, a); by Lemma 2, $\Pr(\mathcal{E}) \geq 1 - \delta$.

For each episode k, define the (episode) exploitability gap:

$$\Delta_k := \max_{i \in [n]} \left(V_{i,1}^{\epsilon_i, \tau_i}\left(\mathrm{sBR}_i(\pi_{-i}^k), \pi_{-i}^k;\, s_0\right) - V_{i,1}^{\epsilon_i, \tau_i}\left(\pi^k;\, s_0\right)\right), \qquad \mathrm{regret}(K) = \sum_{k=1}^{K} \Delta_k.$$

Fix an episode k and let $i^\star$ attain the max above.

A stage-wise performance-difference recursion. Let x_h^k be the state at stage h of episode k and let a_h^k be the joint action sampled from $\pi_h^k(\cdot|x_h^k)$. Fix a player $i^\star \in [n]$, and define the one-step best response to $\pi_{-i^\star}^k$ as

$$\mathrm{sBR}_{i^\star}(\pi_{-i^\star}^k) \in \arg\max_{\sigma_{i^\star}} V_{i^\star,1}^{\epsilon_{i^\star}}\left((\sigma_{i^\star}, \pi_{-i^\star}^k); s_0\right).$$

Define the deviating profile

$$\pi^{k,\mathrm{br}} := (\mathrm{sBR}_{i^\star}(\pi_{-i^\star}^k),\, \pi_{-i^\star}^k).$$

For any stage h and state x, define the value gap

$$\Delta_{k,h}(x) := V_{i^\star,h}^{\epsilon_{i^\star}}(\pi^{k,\mathrm{br}}; x) - V_{i^\star,h}^{\epsilon_{i^\star}}(\pi^k; x).$$

In particular, $\Delta_k = \Delta_{k,1}(s_0)$.

We now bound $\Delta_{k,h}(x)$ by a one-step advantage term plus the next-stage gap. By the definition of the true environment backup and policy evaluation, the stage-h value depends on the stage-h Q table and the policy-risk operator. Indeed, by the risk-sensitive Bellman backup definition, we have that

$$V_{i^\star,h}^{\epsilon_{i^\star}}(\pi; x) = \rho_{i^\star,h}^{\mathsf{e}}\Big(\mathbb{E}_{a \sim \pi_h(\cdot|x)} \left[r_{i^\star,h}(x,a) + V_{i^\star,h+1}^{\epsilon_{i^\star}}(\pi; x') \right] \Big). \tag{35}$$

Apply this to both policies to get that

$$\Delta_{k,h}(x) = \rho_{i^\star,h}^{\mathsf{e}}(Z^{\mathrm{br}}) - \rho_{i^\star,h}^{\mathsf{e}}(Z^k), \tag{36}$$

where we have defined the random variables

$$Z^{\mathrm{br}} := \mathbb{E}\left[r_{i^\star,h}(x, a_h^{k,\mathrm{br}}) + V_{i^\star,h+1}^{\epsilon_{i^\star}}(\pi^{k,\mathrm{br}}; x') \right],$$
$$Z^k := \mathbb{E}\left[r_{i^\star,h}(x, a_h^{k}) + V_{i^\star,h+1}^{\epsilon_{i^\star}}(\pi^{k}; x') \right].$$

Now since $\rho_{i,h}^{\mathsf{e}}$ is L_{env} Lipschitz continuous, we have that $\Delta_{k,h}(x) \leq L_{\mathsf{env}}\,(Z^{\mathrm{br}} - Z^k)$. Expanding the difference of random variables, we have that

$$Z^{\mathrm{br}} - Z^k = \mathbb{E}\left[r(x, a_h^{k,\mathrm{br}}) - r(x, a_h^k) \right] + \mathbb{E}\left[V_{h+1}(\pi^{k,\mathrm{br}}; x') - V_{h+1}(\pi^k; x') \right].$$

The second expectation is exactly $\mathbb{E}\left[\Delta_{k,h+1}(x') \mid x\right]$.

Now define the true stage h Q-function as follows:

$$Q_{i^\star,h}(x,a) := r_{i^\star,h}(x,a) + \rho_{i^\star,h}^{\mathsf{e}}\left(V_{i^\star,h+1}^{\epsilon_{i^\star}}(\pi^k; \cdot)\right).$$

Then using add and subtract to introduce the next-stage term under π^k. Indeed, adding and subtracting $V_{h+1}(\pi^k; x')$ inside the first expectation, we get that

$$\begin{aligned} Z^{\mathrm{br}} - Z^k &= \mathbb{E}\left[r(x, a^{k,\mathrm{br}}) + V_{h+1}(\pi^k; x') + \Big(V_{h+1}(\pi^{k,\mathrm{br}}; x') - V_{h+1}(\pi^k; x') \Big) \right] \\ &\quad - \mathbb{E}\left[r(x, a^k) + V_{h+1}(\pi^k; x') \right]. \\ &= \mathbb{E}\left[r(x, a^{k,\mathrm{br}}) + V_{h+1}(\pi^k; x') \right] - \mathbb{E}\left[r(x, a^k) + V_{h+1}(\pi^k; x') \right] \\ &\quad + \mathbb{E}\left[V_{h+1}(\pi^{k,\mathrm{br}}; x') - V_{h+1}(\pi^k; x') \right]. \end{aligned}$$

The second expectation is exactly $\mathbb{E}\left[\Delta_{k,h+1}(x') \mid x\right]$. The first line, on the other hand, corresponds to

$$\mathbb{E}\Big[Q_{i^\star,h}(x, a^{k,\mathrm{br}}) - Q_{i^\star,h}(x, a^k) \mid x\Big].$$

where we define the true Q-function evaluated using the continuation under π^k as

$$Q_{i^\star,h}(x,a) := r_{i^\star,h}(x,a) + \rho^{\mathsf{e}}_{i^\star,h}\big(V^{\epsilon_{i^\star}}_{i^\star,h+1}(\pi^k;\cdot)\big).$$

Combining these we have

$$\Delta_{k,h}(x) \le L_{\mathsf{env}}\left(\mathbb{E}\big[Q_{i^\star,h}(x, a^{k,\mathrm{br}}) - Q_{i^\star,h}(x, a^k) \mid x\big] + \mathbb{E}\big[\Delta_{k,h+1}(x') \mid x\big]\right).$$

Now observe that the algorithm uses an approximate policy evaluation (due to the estimation of the associated risk measure) with error $\varepsilon_{\mathsf{pol}}$ and an approximate stage equilibrium with error $\varepsilon_{\mathsf{eq}}$. Hence we have that

$$\mathbb{E}\left[Q_{i^\star,h}(x, a_h^{k,\mathrm{br}}) - Q_{i^\star,h}(x, a_h^k) \mid x\right] \le \mathbb{E}\left[Q^k_{i^\star,h}(x, a_h^{k,\mathrm{br}}) - Q^k_{i^\star,h}(x, a_h^k) \mid x\right] + (\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}}).$$

Therefore using monotonicity of $\rho^{\mathsf{env}}_{i^\star,h}$ and the stage-wise best-response definition, and applying the standard telescoping argument as above, we have that

$$\begin{aligned}\Delta_{k,h}(x_h^k) \le L_{\mathsf{env}}\ \mathbb{E}\left[Q^k_{i^\star,h}(x_h^k, a_h^{k,\mathrm{br}}) - Q^k_{i^\star,h}(x_h^k, a_h^k) \,\middle|\, x_h^k\right] \\ + L_{\mathsf{env}}\ \mathbb{E}\left[\Delta_{k,h+1}(x_{h+1}^k) \mid x_h^k\right] + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}}).\end{aligned} \tag{37}$$

where $a_h^{k,\mathrm{br}} \sim \pi_h^{k,\mathrm{br}}(\cdot|x_h^k)$ and $a_h^k \sim \pi_h^k(\cdot|x_h^k)$.

Why the factor L_{env} appears here: the one-step difference in value-to-go is a difference of the form $\rho^{\mathsf{e}}_{i^\star,h}(X) - \rho^{\mathsf{e}}_{i^\star,h}(Y)$, where X, Y are next-state value random variables induced by different action mixtures. By assumption we have

$$|\rho^{\mathsf{e}}_{i^\star,h}(X) - \rho^{\mathsf{e}}_{i^\star,h}(Y)| \le L_{\mathsf{env}}\|X - Y\|_\infty,$$

so that every time we convert a value-function deviation into a risk-to-go deviation, we incur the factor L_{env}.

Unrolling (37) from $h = 1$ to H and using $\Delta_{k,H+1} \equiv 0$ gives

$$\Delta_k \le L_{\mathsf{env}} \sum_{h=1}^{H} \mathbb{E}\left[Q^k_{i^\star,h}(x_h^k, a_h^{k,\mathrm{br}}) - Q^k_{i^\star,h}(x_h^k, a_h^k)\right] + H\, L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}}). \tag{38}$$

Replace $Q^k_{i^\star,h}$ by the optimistic estimates and use optimism. On the event $\mathcal{E}$, Lemma 2 gives, for all (x,a), us that $Q^k_{i^\star,h}(x,a) \le Q^k_{i^\star,h}(x,a)$ and $Q^k_{i^\star,h}(x,a)$ is optimistic for the true backup up to $\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}})$. More concretely, rearranging (23) yields

$$r_{i^\star,h}(x,a) + \rho^{\mathsf{e}}_{i^\star,h}(V^{\epsilon}_{i^\star,h+1}(\pi^k_{h+1})) \le Q^k_{i^\star,h}(x,a) + \varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}}).$$

This implies that the *true* advantage of $a_h^{k,\mathrm{br}}$ over a_h^k is controlled by the *optimistic* advantage plus the additive approximation error; plugging into (38) yields

$$\Delta_k \le L_{\mathsf{env}} \sum_{h=1}^{H} \mathbb{E}\left[Q^k_{i^\star,h}(x_h^k, a_h^{k,\mathrm{br}}) - Q^k_{i^\star,h}(x_h^k, a_h^k)\right] + H\Big(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}})\Big). \tag{39}$$

Use that π_h^k is an approximate stage equilibrium for the optimistic Q table. By the assumption on the equilibrium approximation, at each (h, x_h^k), $\pi_h^k(\cdot|x_h^k)$ is $\varepsilon_{\mathsf{eq}}$-optimal for the stage game defined

by the optimistic Q table. In particular, as shown in Lemma 3, the improvement obtained by any alternative (including the stage-wise best response) is bounded by $\varepsilon_{\mathsf{eq}}$ in the policy-risk value, and thus the expected optimistic advantage satisfies

$$\mathbb{E}\left[Q^k_{i^\star,h}(x^k_h, a^{k,\mathrm{br}}_h) - Q^k_{i^\star,h}(x^k_h, a^k_h)\right] \le \mathbb{E}\left[\beta\sqrt{\phi(x^k_h, a^k_h, h)^\top(\Lambda^k_h)^{-1}\phi(x^k_h, a^k_h, h)}\right] + \varepsilon_{\mathsf{eq}}.$$

Substituting into (39) gives

$$\Delta_k \le L_{\mathsf{env}}\beta\sum_{h=1}^{H}\mathbb{E}\left[\sqrt{\phi(x^k_h, a^k_h, h)^\top(\Lambda^k_h)^{-1}\phi(x^k_h, a^k_h, h)}\right] + H\Big(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}})\Big).$$

Sum over episodes and apply the elliptical potential lemma. Summing over $k = 1, \dots, K$ and applying Cauchy–Schwarz,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\phi^\top_{k,h}(\Lambda^k_h)^{-1}\phi_{k,h}} \le \sqrt{KH}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\phi^\top_{k,h}(\Lambda^k_h)^{-1}\phi_{k,h}},$$

where $\phi_{k,h} := \phi(x^k_h, a^k_h, h)$. For each fixed h, the standard elliptical potential argument yields

$$\sum_{k=1}^{K}\phi^\top_{k,h}(\Lambda^k_h)^{-1}\phi_{k,h} \le 2\log\frac{\det(\Lambda^{K+1}_h)}{\det(\lambda I)} \le 2d\log\left(1 + \frac{K}{\lambda}\right).$$

Summing over $h = 1, \dots, H$ gives

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\phi^\top_{k,h}(\Lambda^k_h)^{-1}\phi_{k,h} \le 2Hd\log\left(1 + \frac{K}{\lambda}\right).$$

Therefore,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\phi^\top_{k,h}(\Lambda^k_h)^{-1}\phi_{k,h}} \le \sqrt{KH}\sqrt{2Hd\log\left(1 + \frac{K}{\lambda}\right)} = \tilde{O}\left(H\sqrt{Kd}\right).$$

Putting everything together,

$$\mathrm{reg}(K) = \sum_{k=1}^{K}\Delta_k \le \tilde{O}\left(L_{\mathsf{env}}\beta\, H\sqrt{Kd}\right) + KH\Big(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}(\varepsilon_{\mathsf{pol}} + \varepsilon_{\mathsf{eq}})\Big).$$

Get Final Regret Bound by Plugging in $\beta = cB\sqrt{d^2H\,\iota}$. Using $\beta = \tilde{O}(B\sqrt{d^2H})$, we have that

$$L_{\mathsf{env}}\beta\, H\sqrt{Kd} = \tilde{O}\left(L_{\mathsf{env}}B\sqrt{d^2H}\cdot H\sqrt{Kd}\right) = \tilde{O}\left(L_{\mathsf{env}}B\sqrt{Kd^3H^3}\right),$$

which yields exactly (6) on the event $\mathcal{E}$. Since $\Pr(\mathcal{E}) \ge 1 - \delta$, the theorem holds with probability at least $1 - \delta$. □

B.4 Regret with Explicit Error Dependence: Entropic Risks & Regularizers

In Section 3.1 we specialize to the case where the risk measures and regularizers are entropic. That is, in the special case where the *policy-space* risk penalty is entropic (KL) and the player policy is entropy-regularized, there are number of methods one can employ including no-regret learning methods such as extra-gradient or mirror-prox, multiplicative weights, or iterative best response. In

this case, we can obtain more explicit bounds which reveal how the risk sensitivity and bounded rationality parameters influence different performance criteria.

Set constants

$$A_{\max} := \max_{i\in[n]} |\mathcal{A}_i|, \qquad \epsilon_{\min} := \min_{i\in[n]} \epsilon_i, \qquad \tau_{\min} := \min_{i\in[n]} \tau_i.$$

Assume stage rewards satisfy $r_{i,h}(x,a) \in [0,1]$ and $V_{i,H+1}^{\epsilon_i,\tau_i} \equiv 0$. Then the cumulative reward-to-go is bounded by H. Moreover the entropy regularization satisfies

$$0 \le \Phi(\pi_i(\cdot|x)) \le \log|\mathcal{A}_i|.$$

Hence the additive regularization term satisfies

$$0 \le \Phi(\pi_i(\cdot|x))/\epsilon_i \le \log(|\mathcal{A}_i|)/\epsilon_i.$$

Since the entropic policy-value operator (7) is a log-sum-exp aggregation of bounded Q values, it preserves the same scale as the underlying payoffs. Consequently, clipping Q to $[0,B]$ with

$$B_i := H\left(1 + \frac{\log|\mathcal{A}_i|}{\epsilon_i}\right), \qquad B := \max_{i\in[n]} B_i,$$

ensures

$$0 \le V_{i,h}^{\epsilon_i,\tau_i}(\pi; x; Q) \le B \quad \text{for all } i,h,x,\pi.$$

Using $A_{\max}$ and $\epsilon_{\min}$ we obtain the simpler bound

$$B \le H\left(1 + \frac{\log A_{\max}}{\epsilon_{\min}}\right). \tag{40}$$

Substituting this value range into Theorem 2 and expanding the logarithmic factors yields the following explicit regret bound.

Corollary 3 (Explicit regret bound with $(n, |\mathcal{A}_i|)$ dependence)**.** Under the assumptions of Theorem 2 with entropic policy regularization $\nu_i(\mu_i) = \Phi(\mu_i)$, let $A_{\max} = \max_i |\mathcal{A}_i|$ and $\epsilon_{\min} = \min_i \epsilon_i$. Then for an absolute constant $C_1 > 0$, with probability at least $1-\delta$, the bound holds:

$$\begin{aligned}\mathrm{reg}(K) \le{}& C_1 L_{\mathsf{env}} H\Big(1 + \tfrac{\log A_{\max}}{\epsilon_{\min}}\Big)\sqrt{K d^3 H^3 \log\left(\tfrac{nH}{\delta}\right)} \\ &+ KH\Big(\varepsilon_{\mathsf{env}} + L_{\mathsf{env}}\varepsilon_{\mathsf{pol}} + L_{\mathsf{env}}\, c_{\mathsf{eq}}\, B\sqrt{\tfrac{\Delta_{\mathsf{eq}}}{\tau_{\min}}}\Big).\end{aligned} \tag{41}$$

If the stage solver runs a no-regret algorithm for T iterations, then the duality gap satisfies

$$\Delta_{\mathsf{eq}} \;\lesssim\; \sum_{i=1}^{n} \sqrt{\frac{\log|\mathcal{A}_i|}{T}}.$$

Alternatively, if the equilibrium variational inequality is strongly monotone with modulus $\mu(\tau)$ and Lipschitz constant $L(\tau)$, then an extragradient or Mirror-Prox solver yields

$$\varepsilon_{\mathsf{eq}}(T) \le C\exp(-T\mu/L).$$

Suppose the environment risk operator is the entropic risk measure

$$\rho_\tau(Z) = \frac{1}{\tau}\log\mathbb{E}[\exp(\tau Z)].$$

Then the operator is 1-Lipschitz in the sup norm—i.e.,

$$|\rho_\tau(Z) - \rho_\tau(Z')| \le \|Z - Z'\|_\infty,$$

so we may take $L_{\mathsf{env}} = 1$.

For the environment risk estimation error, let $Z \in [0, B]$ denote the continuation value. Given m samples $Z_1, \ldots, Z_m$, define the estimator

$$\hat{\rho}_\tau = \frac{1}{\tau} \log \left(\frac{1}{m} \sum_{j=1}^{m} \exp(\tau Z_j) \right).$$

Applying a union bound over all players $i \in [n]$ and stages $h \in [H]$ and setting the per-event failure probability to $\delta/(nH)$ yields

$$\varepsilon_{\mathsf{env}} \lesssim \frac{1}{\tau_{\min}} \exp(\tau_{\max} B) \sqrt{\frac{\log(nH/\delta)}{m}}. \tag{42}$$

The exponential factor arises because the entropic risk depends on the moment generating function $\mathbb{E}[e^{\tau Z}]$, and without additional distributional assumptions the range bound $Z \in [0, B]$ implies $e^{\tau Z} \in [1, e^{\tau B}]$, which necessarily introduces $\exp(\tau B)$.

If the policy-side entropic operator (7) is evaluated exactly (finite action spaces), then $\varepsilon_{\mathsf{pol}} = 0$. If it is approximated via sampling m actions, a typical bound is

$$\varepsilon_{\mathsf{pol}} \lesssim \frac{1}{\tau_{\min}} \exp(\tau_{\max} B) \sqrt{\frac{\log(nH/\delta)}{m}}.$$

Substituting these expressions yields the regret bound

$$\begin{aligned} \mathrm{reg}(K) \leq \tilde{\mathcal{O}} \left(H \left(1 + \tfrac{\log A_{\max}}{\epsilon_{\min}} \right) \sqrt{K d^3 H^3} \right) \\ + KH \left(\frac{1}{\tau_{\min}} \exp(\tau_{\max} B) \sqrt{\tfrac{\log(nH/\delta)}{m}} + c_{\mathsf{eq}} B \sqrt{\tfrac{\Delta_{\mathsf{eq}}}{\tau_{\min}}} \right). \end{aligned} \tag{43}$$

Making the (τ, ϵ) interaction explicit. Under entropic policy regularization with rewards in $[0, 1]$ and horizon H, clipping ensures the continuation values satisfy $Z \in [0, B]$ with

$$B = \max_{i \in [n]} B_i, \qquad B_i := H \Big(1 + \frac{\log |\mathcal{A}_i|}{\epsilon_i} \Big) \leq H \Big(1 + \frac{\log A_{\max}}{\epsilon_{\min}} \Big),$$

where $A_{\max} = \max_i |\mathcal{A}_i|$ and $\epsilon_{\min} = \min_i \epsilon_i$. Plugging this into (42) gives

$$\varepsilon_{\mathsf{env}} \lesssim \frac{1}{\tau_{\min}} \exp \left(\tau_{\max} H \Big(1 + \frac{\log A_{\max}}{\epsilon_{\min}} \Big) \right) \sqrt{\frac{\log(nH/\delta)}{m}}. \tag{44}$$

Note that in the dual representation the ambiguity penalty scales as $\frac{1}{\tau} \mathrm{KL}(\cdot \| \cdot)$, so increasing τ reduces the penalty weight and therefore moves the objective *toward the risk-neutral limit*. In this sense, larger τ corresponds to *less* robustness, while smaller τ enforces stronger regularization. Consequently, the statistical error bound (44) worsens with $\tau_{\max}$ through the factor $\exp(\tau_{\max} B)$ even though the underlying objective becomes closer to risk-neutral as τ increases.

Risk-sensitivity regimes. The generic estimation bound

$$\varepsilon_{\mathsf{env}} \lesssim \frac{1}{\tau_{\min}} \exp(\tau_{\max} B) \sqrt{\frac{\log(nH/\delta)}{m}}$$

exhibits two opposing effects: increasing τ decreases the prefactor $1/\tau_{\min}$ but increases the exponential $\exp(\tau_{\max} B)$.

One can interpret the bound in terms of the "effective scale"

$$g(\tau) := \frac{1}{\tau} \exp(\tau B),$$

which is minimized at $\tau^\star = 1/B$ (treating B as fixed). Consequently, the estimate is most favorable when $\tau_{\max} B$ is $O(1)$, and becomes large both when τ is too small (large $1/\tau$) and when τ is too large (large $\exp(\tau B)$).

In particular, for $\tau_{\max} B \lesssim 1$ we may expand $\exp(\tau_{\max} B) = 1 + \tau_{\max} B + O((\tau_{\max} B)^2)$ to obtain

$$\varepsilon_{\mathsf{env}} \lesssim \frac{1}{\tau_{\min}} \sqrt{\frac{\log(nH/\delta)}{m}} + \frac{\tau_{\max}}{\tau_{\min}} B \sqrt{\frac{\log(nH/\delta)}{m}} + O\Big(\frac{\tau_{\max}^2 B^2}{\tau_{\min}} \sqrt{\frac{\log(nH/\delta)}{m}}\Big). \tag{45}$$

If $\tau_{\max} \asymp \tau_{\min} = \tau$ and $\tau B \ll 1$, this simplifies to

$$\varepsilon_{\mathsf{env}} \lesssim (1 + B + O(\tau B^2)) \sqrt{\frac{\log(nH/\delta)}{m}},$$

up to constants, recovering the familiar bounded-value estimation rate. Thus, in the strongly robust regime the leading dependence is typically $\varepsilon_{\mathsf{env}} \asymp \frac{1}{\tau}\sqrt{\log(nH/\delta)/m}$ (up to lower-order additive terms involving B). When τ is large, the exponential factor dominates and the bound behaves as $\varepsilon_{\mathsf{env}} \asymp B\sqrt{\log(nH/\delta)/m}$. Decreasing $\epsilon_{\min}$ increases $\varepsilon_{\mathsf{env}}$ through the value range B, while decreasing $\tau_{\min}$ increases $\varepsilon_{\mathsf{env}}$ through the prefactor $1/\tau_{\min}$. In addition, $\tau_{\max}$ and $\epsilon_{\min}$ interact multiplicatively in the exponent $\exp(\tau_{\max} B)$ via B.

B.5 QRE-OVI: Quantal Response Optimistic Value Iteration

The risk-neutral, boundedly rational setting—in which agents employ *quantal responses*—is a special case of our framework. Specifically, QRE-OVI is obtained from Algorithm 1 by setting the environment and policy risk operators to their risk-neutral limits, i.e., replacing $\rho^{\mathsf{e}}_{i,h}$ with the standard expectation $\mathbb{E}_{x' \sim \mathcal{P}_h(\cdot|x,a)}[\cdot]$ and removing the policy-risk adversary by setting $\varphi_{\mathsf{p},i} \equiv 0$. The resulting algorithm retains the entropy regularization $(1/\epsilon_i)\nu_i(\pi_i)$ that defines the quantal response, and at each stage game computes a QRE rather than a Nash equilibrium or an RQRE.

This yields a direct replacement of the Nash oracle in NQOVI (Cisneros-Velarde & Koyejo, 2023): the only modification is that the stage-game solver `Nash` is replaced by $\texttt{QRE}_\varepsilon$, while the optimistic value iteration structure, linear function approximation, and confidence bonus construction remain unchanged. The advantage of this substitution is that QRE is unique for any finite game and any $\epsilon_i > 0$ (McKelvey & Palfrey, 1995), eliminating the equilibrium selection problem inherent to Nash.

Corollary 4 (Regret bound for QRE-OVI)**.** Consider Algorithm 1 in the risk-neutral specialization described above, so that $L_{\mathsf{env}} = 1$, $\varepsilon_{\mathsf{env}} = 0$, and $\varepsilon_{\mathsf{pol}} = 0$. Under Assumption 2 with $B = \max_{i \in [n]} H(1 + \log|\mathcal{A}_i|/\epsilon_i)$, with probability at least $1 - \delta$,

$$\mathrm{reg}(K) \le \tilde{\mathcal{O}}\left(B\sqrt{K\, d^3\, H^3}\right) + KH\, \varepsilon_{\mathsf{eq}},$$

where $\varepsilon_{\mathsf{eq}}$ is the stage-game QRE approximation error.

Proof. Set $L_{\mathsf{env}} = 1$, $\varepsilon_{\mathsf{env}} = 0$, and $\varepsilon_{\mathsf{pol}} = 0$ in Theorem 2. The bound follows immediately from (6). □

Compared to the NQOVI regret bound of Cisneros-Velarde & Koyejo (2023), the leading statistical term acquires the factor $B(\epsilon) = H(1 + \log|\mathcal{A}|/\epsilon)$ in place of H, reflecting the enlarged value range due to entropy regularization. This is the price of uniqueness and Lipschitz stability (Corollary 2): as $\epsilon \to \infty$, $B \to H$ and the bound approaches that of NQOVI, while the equilibrium computation becomes increasingly ill-conditioned. Conversely, moderate ϵ yields a modest increase in the regret bound while guaranteeing a unique, stable equilibrium at every stage game.

C Risk Estimation

In Algorithm 1, the true risk operator ρ_i cannot be computed exactly because the transition kernel $\mathcal{P}_h$ is unknown to the agents. Only samples $(x_h^t, a_h^t, x_{h+1}^t)$ are observed, so the true distribution of the continuation value $Z := V_{h+1}^i(X_{h+1})$, where $X_{h+1} \sim \mathcal{P}_h(\cdot \mid x, a)$, is unavailable. In this section, we formalize the distinction between ρ and its computable proxy $\hat{\rho}$, quantify the resulting approximation error, and provide concrete instantiations for common risk measures.

C.1 Empirical Environment Risk Operator

Recall from (2) that the true environment risk operator at stage h takes the form

$$\rho_{i,h}^{\mathsf{e}}(Z \mid \mathcal{P}_h(\cdot \mid x, a)) = \inf_{\widetilde{\mathcal{P}} \in \Delta(\mathcal{X})} \left\{ \int Z(x')\, \widetilde{\mathcal{P}}(dx') + \varphi_{\mathsf{e},i}(\widetilde{\mathcal{P}} \| \mathcal{P}_h(\cdot \mid x, a)) \right\}.$$

Since $\mathcal{P}_h$ is unknown, Algorithm 1 replaces this with an empirical approximation. Two strategies are available depending on the structure of the penalty $\varphi_{\mathsf{e},i}$.

Finite dual discretization for general penalties. For a general convex penalty $\varphi_{\mathsf{e},i}$, we approximate the infimum over $\Delta(\mathcal{X})$ by restricting to a finite set of candidate transition kernels $\widehat{\mathcal{Q}}_{i,h}(x,a) = \{\widetilde{P}_h^{(1)}(\cdot \mid x, a), \ldots, \widetilde{P}_h^{(M)}(\cdot \mid x, a)\}$, yielding

$$\widehat{\rho}_{i,h}^{\mathsf{e},k}(Z \mid x, a) := \min_{m \in [M]} \left\{ \widehat{\mathbb{E}}_{x' \sim \widetilde{P}_h^{(m)}(\cdot \mid x, a)}[Z(x')] + \varphi_{\mathsf{e},i}\left(\widetilde{P}_h^{(m)}(\cdot \mid x, a) \,\middle\|\, \widehat{P}_h^k(\cdot \mid x, a) \right) \right\},$$

where $\widehat{P}_h^k$ is the empirical estimate of $\mathcal{P}_h$ constructed from the first $k-1$ episodes. The expectations under each candidate kernel may themselves be approximated via Monte Carlo sampling:

$$\int Z(x')\, \widetilde{P}_h^{(m)}(dx') \approx \frac{1}{N} \sum_{j=1}^{N} Z(x'^{(j)}), \qquad x'^{(j)} \sim \widetilde{P}_h^{(m)}(\cdot \mid x, a).$$

The approximation error $\varepsilon_{\mathsf{env}}$ then depends on both the covering quality of $\widehat{\mathcal{Q}}_{i,h}$ and the Monte Carlo sample size N.

Closed-form evaluation under entropic penalties. When the penalty is the scaled KL divergence $\varphi_{\mathsf{e},i}(\widetilde{\mathcal{P}} \| \mathcal{P}_h) = \frac{1}{\tau_i} \mathrm{KL}(\widetilde{\mathcal{P}} \| \mathcal{P}_h)$, the environment risk operator admits the closed form

$$\rho_{i,h}^{\mathsf{e}}(Z \mid \mathcal{P}_h) = \frac{1}{\tau_i} \log \mathop{\mathbb{E}}_{x' \sim \mathcal{P}_h(\cdot \mid x, a)} \left[\exp(\tau_i\, Z(x')) \right].$$

In this case, no finite dual discretization is needed; the operator is approximated by replacing $\mathcal{P}_h$ with the empirical transition estimate $\widehat{P}_h^k$ and evaluating the resulting log-sum-exp expression from samples. Given N independent next-state samples $x'^{(1)}, \ldots, x'^{(N)} \sim \widehat{P}_h^k(\cdot \mid x, a)$, the empirical estimator is

$$\widehat{\rho}_{i,h}^{\mathsf{e},k}(Z \mid x, a) = \frac{1}{\tau_i} \log \left(\frac{1}{N} \sum_{j=1}^{N} \exp(\tau_i\, Z(x'^{(j)})) \right).$$

C.2 True versus empirical risk operators

For a fixed policy profile π, the true risk-adjusted Q-function is defined via the Bellman recursion

$$Q_h^{i,\star}(x,a) \;:=\; r_{i,h}(x,a) + \rho_i \left(V_{h+1}^{i,\star}(X_{h+1}) \right), \quad X_{h+1} \sim \mathcal{P}_h(\cdot \mid x, a),$$

where ρ_i is applied to the distribution of the random continuation value induced by the true transition kernel. Since ρ_i integrates out all randomness, $Q_h^{i,\star}$ is a deterministic function of (x, a).

The learning algorithm replaces ρ_i with a computable proxy $\hat{\rho}_i$, as appears in the regression targets of Algorithm 1. Two sources of approximation arise. First, if ρ_i admits a dual representation (Theorem 1),

$$\rho_i(Z) \;=\; \sup_{p\in\mathcal{P}_i} \{\mathbb{E}_p[Z] - D_i(p)\}\,,$$

the algorithm may use a finite subset $\widehat{\mathcal{P}}_i = \{p^{(1)},\dots,p^{(M)}\} \subset \mathcal{P}_i$, yielding the finite-dual approximation

$$\hat{\rho}_i(Z) \;:=\; \max_{j\in[M]} \left\{\mathbb{E}_{p^{(j)}}[Z] - D_i\left(p^{(j)}\right)\right\}.$$

Second, the expectations under each $p^{(j)}$ may themselves be estimated from samples. We define the worst-case approximation error as

$$\varepsilon_{\mathcal{P}} \;:=\; \max_{i\in[n]} \sup_{Z:\,\|Z\|_\infty\le H} |\rho_i(Z) - \hat{\rho}_i(Z)|\,. \tag{46}$$

This quantity enters the regret bounds as an additive bias of $\varepsilon_{\mathcal{P}}$ per Bellman backup, accumulating to $O(KH\varepsilon_{\mathcal{P}})$ over K episodes. When $\widehat{\mathcal{P}}_i = \mathcal{P}_i$ and expectations are computed exactly, $\varepsilon_{\mathcal{P}} = 0$ and the bound in Theorem 2 applies without modification.

C.3 Examples

We instantiate the approximation error for three common risk measures.

Example 3 (Entropic risk)**.** For parameter $\tau_i > 0$, the entropic risk measure is

$$\rho_i(Z) \;:=\; \frac{1}{\tau_i}\,\log\left(\mathbb{E}\left[e^{\tau_i Z}\right]\right).$$

Given m independent samples $Z_1,\dots,Z_m$ from the distribution of Z, the empirical entropic risk is

$$\hat{\rho}_i(Z) \;:=\; \frac{1}{\tau_i}\,\log\left(\frac{1}{m}\sum_{\ell=1}^{m} e^{\tau_i Z_\ell}\right).$$

Under sub-Gaussian assumptions on Z, the approximation error satisfies $|\hat{\rho}_i(Z) - \rho_i(Z)| = O_{\mathbb{P}}\left(\sqrt{\log(1/\delta)/m}\right)$.

Example 4 (Conditional Value-at-Risk)**.** For confidence level $\alpha \in (0,1)$, the CVaR of Z is

$$\rho_i(Z) \;:=\; \mathrm{CVaR}_\alpha(Z) \;=\; \frac{1}{1-\alpha}\int_\alpha^1 F_Z^{-1}(u)\,du,$$

where F_Z^{-1} is the quantile function. Given samples $Z_1,\dots,Z_m$ with order statistics $Z_{(1)} \le \cdots \le Z_{(m)}$, the empirical CVaR is

$$\hat{\rho}_i(Z) \;:=\; \frac{1}{(1-\alpha)m}\sum_{\ell=\lceil \alpha m\rceil}^{m} Z_{(\ell)}.$$

Standard results give $|\hat{\rho}_i(Z) - \mathrm{CVaR}_\alpha(Z)| = O_{\mathbb{P}}\left(\sqrt{1/m}\right)$.

Example 5 (Coherent risk with finite dual set)**.** Let $\mathcal{P}_i$ be a convex set of distributions over the support of Z and $D_i : \mathcal{P}_i \to \mathbb{R}_+$ a convex penalty. The coherent risk measure

$$\rho_i(Z) \;:=\; \sup_{p\in\mathcal{P}_i} \{\mathbb{E}_p[Z] - D_i(p)\}$$

is approximated using a finite subset $\widehat{\mathcal{P}}_i = \{p^{(1)},\dots,p^{(M)}\}$ via $\hat{\rho}_i(Z) := \max_{j\in[M]}\{\mathbb{E}_{p^{(j)}}[Z] - D_i(p^{(j)})\}$. The approximation error $\varepsilon_{\mathcal{P}} = \sup_Z |\rho_i(Z) - \hat{\rho}_i(Z)|$ depends on the covering quality of $\widehat{\mathcal{P}}_i$ and enters the regret bound as an additive term.

D Stability Analysis: Stage Games to Finite-Horizon Value Robustness

In this section, we provide the complete stability analysis for RQRE under linear function approximation. We first establish stagewise robustness (Section D.1), then lift these bounds to finite-horizon value functions via backward induction (Section D.2).

D.1 Stagewise Robustness under Linear Function Approximation

Consider the finite-horizon Markov game with players $i \in [n]$, state space $\mathcal{X}$, joint action space $\mathcal{A} = \prod_{i=1}^{n} \mathcal{A}_i$, and stages $h \in [H]$. Rewards are bounded i.e., $|r_{i,h}(x,a)| \le R_{\max}$.

Under linear function approximation (Assumption 2), each player's Q-function takes the form

$$Q_h^i(x,a;w_h^i) = \phi(x,a,h)^\top w_h^i, \qquad \|\phi(x,a,h)\|_2 \le 1 \quad \forall (x,a,h).$$

For fixed (x,h), define the induced stage-game payoff table for player i as

$$U_h^i(x)_a := Q_h^i(x,a;w_h^i).$$

Let $U_h(x) := (U_h^1(x), \dots, U_h^n(x))$ denote the collection of payoff tables.

Assumption 3 (RQRE Lipschitz Continuity)**.** At each (x,h), the stage-game RQRE is defined by a μ-strongly concave objective with $\mu := \frac{1}{\epsilon} + \tau \cdot \alpha_R > 0$, yielding a unique equilibrium profile $\pi_h^{\texttt{RQRE}}(x; U_h(x)) \in \prod_{i=1}^{n} \Delta(\mathcal{A}_i)$.

A consequence of this assumption is that the equilibrium map is Lipschitz in payoff tables:

$$\left\| \pi_h^{\texttt{RQRE}}(x;U) - \pi_h^{\texttt{RQRE}}(x;\widetilde{U}) \right\|_1 \le L_{\texttt{RQRE}} \|U - \widetilde{U}\|_\infty, \qquad L_{\texttt{RQRE}} \le \frac{c}{\mu}, \tag{47}$$

where $c > 0$ is a universal constant.

Theorem 3 (Stagewise Robustness under Linear Function Approximation)**.** Fix state x and stage h. Let $w_h := \{w_h^i\}_{i\in[n]}$ and $\widetilde{w}_h := \{\widetilde{w}_h^i\}_{i\in[n]}$ be two parameter collections inducing payoff tables $U_h(x)$ and $\widetilde{U}_h(x)$. Then $\|U_h(x) - \widetilde{U}_h(x)\|_\infty \le \max_{i\in[n]} \|w_h^i - \widetilde{w}_h^i\|_2$ so that the RQRE policies satisfy

$$\left\| \pi_h^{\texttt{RQRE}}(x;w_h) - \pi_h^{\texttt{RQRE}}(x;\widetilde{w}_h) \right\|_1 \le \frac{c}{\mu} \max_{i\in[n]} \|w_h^i - \widetilde{w}_h^i\|_2. \tag{48}$$

Proof. For any player i and joint action $a \in \mathcal{A}$, we have that

$$|U_h^i(x)_a - \widetilde{U}_h^i(x)_a| = |\phi(x,a,h)^\top (w_h^i - \widetilde{w}_h^i)| \le \|\phi(x,a,h)\|_2 \|w_h^i - \widetilde{w}_h^i\|_2 \le \|w_h^i - \widetilde{w}_h^i\|_2.$$

Taking the maximum over i and a yields the payoff bound. The policy bound follows from Assumption 3. □

Remark 3. Theorem 3 is agnostic to how the parameter estimates are obtained. It applies to any estimation procedure (OVI, least-squares regression, etc.) that provides parameter-error bounds.

D.2 Finite-Horizon Value Robustness via Backward Induction

We now lift stagewise policy perturbations to value function perturbations over the full horizon.

For a Markov policy profile $\pi = \{\pi_h\}_{h=1}^H$, recall the recursive definition of the risk-adjusted value function:

$$Q_h^{i,\pi}(x,a) := \mathcal{R}_{i,h}^{\mathsf{e}}\left(r_{i,h}, \mathcal{P}_h, V_{h+1}^{i,\pi}; x, a\right),$$
$$V_h^{i,\pi}(x) := \mathcal{R}_{i,h}^{\mathsf{p}}(Q_h^{i,\pi}, \pi_h; x) + \tfrac{1}{\epsilon_i}\nu_i(\pi_{i,h}(\cdot \mid x)).$$

For two joint-action distributions $\pi_h(\cdot \mid x)$ and $\widetilde{\pi}_h(\cdot \mid x)$, define the pointwise deviation

$$\Delta_h(x) := \|\pi_h(\cdot \mid x) - \widetilde{\pi}_h(\cdot \mid x)\|_1.$$

Lemma 4 (One-Step Value Sensitivity)**.** Let $f : \mathcal{A} \to \mathbb{R}$ satisfy $\|f\|_\infty \le M$. For any distributions $p, q \in \Delta(\mathcal{A})$,

$$|\mathbb{E}_{a\sim p}[f(a)] - \mathbb{E}_{a\sim q}[f(a)]| \le M\|p - q\|_1.$$

Theorem 4 (Risk-Sensitive Value Perturbation)**.** Let $\pi = \{\pi_h\}_{h=1}^H$ and $\widetilde{\pi} = \{\widetilde{\pi}_h\}_{h=1}^H$ be two Markov policy profiles. Let L_{env} be the Lipschitz constant of the environment risk operator as in Theorem 2. Assume $Q_{\max,h}$ is an upper bound incorporating both the payoff and regularization ranges, i.e., $Q_{\max,h} \ge \|Q_h^{i,\pi}\|_\infty + L_\nu/\epsilon_i$ for all h, i, π, where L_ν is the Lipschitz constant of ν_i. Then for all $h \in [H]$ and $x \in \mathcal{X}$,

$$\left|V_h^{i,\pi}(x) - V_h^{i,\widetilde{\pi}}(x)\right| \le \sum_{t=h}^{H} (L_{\mathsf{env}})^{t-h} Q_{\max,t} \cdot \mathbb{E}_{\widetilde{\pi}}\left[\Delta_t(X_t) \mid X_h = x\right]. \tag{49}$$

Taking the supremum over the state space, we have that

$$\sup_{x\in\mathcal{X}} \left|V_h^{i,\pi}(x) - V_h^{i,\widetilde{\pi}}(x)\right| \le \sum_{t=h}^{H} (L_{\mathsf{env}})^{t-h} Q_{\max,t} \cdot \sup_{x'\in\mathcal{X}} \Delta_t(x'). \tag{50}$$

Proof. We proceed by backward induction on h.

Base case ($h = H$): At the terminal stage there is no continuation value, so the policy-risk operator reduces to evaluation of the immediate payoff:

$$V_H^{i,\pi}(x) = \mathcal{R}_{i,H}^{\mathsf{p}}(r_{i,H}, \pi_H; x) + \tfrac{1}{\epsilon_i}\nu_i(\pi_{i,H}(\cdot \mid x)),$$

which has sensitivity to π_H bounded by $Q_{\max,H}$ since $r_{i,h}(x, a) \in [0, 1]$ and $Q_{\max,H}$ incorporates the regularization range. By Lemma 4,

$$\left|V_H^{i,\pi}(x) - V_H^{i,\widetilde{\pi}}(x)\right| \le Q_{\max,H}\|\pi_H(\cdot|x) - \widetilde{\pi}_H(\cdot|x)\|_1 = (L_{\mathsf{env}})^0 Q_{\max,H}\Delta_H(x).$$

This satisfies (49) for $t = H$.

Inductive step: Assume (49) holds for stage $h + 1$. Consider stage h. We decompose the value difference into two terms: deviation due to the policy change (Term I) and deviation due to the continuation value propagation (Term II) as follows:

$$\left|V_h^{i,\pi}(x) - V_h^{i,\widetilde{\pi}}(x)\right| \le \underbrace{\left|\mathcal{R}_{i,h}^{\mathsf{p}}(Q_h^{i,\pi}, \pi_h; x) - \mathcal{R}_{i,h}^{\mathsf{p}}(Q_h^{i,\pi}, \widetilde{\pi}_h; x)\right| + \tfrac{1}{\epsilon_i}\left|\nu_i(\pi_{i,h}(\cdot \mid x)) - \nu_i(\widetilde{\pi}_{i,h}(\cdot \mid x))\right|}_{\text{(I) Policy Error}}$$
$$+ \underbrace{\left|\mathcal{R}_{i,h}^{\mathsf{p}}(Q_h^{i,\pi}, \widetilde{\pi}_h; x) - \mathcal{R}_{i,h}^{\mathsf{p}}(Q_h^{i,\widetilde{\pi}}, \widetilde{\pi}_h; x)\right|}_{\text{(II) Value Propagation}}.$$

Bounding (I): By Lemma 4, the policy-risk operator term satisfies $|\mathcal{R}_{i,h}^{\mathsf{p}}(Q_h^{i,\pi}, \pi_h; x) - \mathcal{R}_{i,h}^{\mathsf{p}}(Q_h^{i,\pi}, \widetilde{\pi}_h; x)| \le \|Q_h^{i,\pi}\|_\infty \Delta_h(x)$, and the Lipschitz continuity of ν_i gives $\frac{1}{\epsilon_i}|\nu_i(\pi_{i,h}(\cdot|x)) - \nu_i(\widetilde{\pi}_{i,h}(\cdot|x))| \le (L_\nu/\epsilon_i)\Delta_h(x)$. Since $Q_{\max,h} \ge \|Q_h^{i,\pi}\|_\infty + L_\nu/\epsilon_i$ by assumption, we obtain

$$\text{(I)} \le Q_{\max,h}\, \Delta_h(x).$$

Bounding (II): Since $\mathcal{R}_{i,h}^{\mathsf{p}}$ is 1-Lipschitz in the Q-argument and $Q_h^{i,\pi}$ depends on $V_{h+1}^{i,\pi}$ through $\mathcal{R}_{i,h}^{\mathsf{e}}$, which is L_{env}-Lipschitz in the continuation value, we obtain

$$\text{(II)} \le L_{\mathsf{env}}\, \mathbb{E}_{\widetilde{\pi}}\left[\left|V_{h+1}^{i,\pi}(X_{h+1}) - V_{h+1}^{i,\widetilde{\pi}}(X_{h+1})\right| \;\middle|\; X_h = x\right].$$

Combining the bounds on (I) and (II): Applying the inductive hypothesis for the term inside the expectation, we have that

$$\begin{aligned}\text{(II)} &\leq L_{\mathsf{env}}\mathbb{E}_{\widetilde{\pi}}\left[\sum_{t=h+1}^{H}(L_{\mathsf{env}})^{t-(h+1)}Q_{\max,t}\mathbb{E}_{\widetilde{\pi}}[\Delta_t(X_t)\mid X_{h+1}]\;\middle|\; X_h=x\right]\\ &=\sum_{t=h+1}^{H}(L_{\mathsf{env}})^{t-h}Q_{\max,t}\mathbb{E}_{\widetilde{\pi}}[\Delta_t(X_t)\mid X_h=x].\end{aligned}$$

Adding Term (I) ($Q_{\max,h}\Delta_h(x)$) to Term (II) completes the sum from $t=h$ to H, proving the theorem. □

Corollary 5 (End-to-End Value Robustness for RQRE-OVI)**.** Suppose Algorithm 1 produces parameter estimates $\widehat{w}_h$ satisfying

$$\sup_{x\in\mathcal{X}}\max_{i\in[n]}\|\widehat{w}_h^i - w_h^{i,\star}\|_2 \leq \delta_h \quad \forall h\in[H].$$

Let $\pi_h(x) := \pi_h^{\texttt{RQRE}}(x;\widehat{w}_h)$ and $\pi_h^\star(x) := \pi_h^{\texttt{RQRE}}(x;w_h^\star)$, where both are exact stage-game RQRE solutions for the respective payoff tables. By Theorem 3, the policy deviation is bounded by $\sup_x \Delta_h(x) \leq \frac{c}{\mu}\delta_h$. When the stage solver is approximate with error $\varepsilon_{\mathsf{eq}}$, the policy deviation bound becomes $\sup_x \Delta_h(x) \leq \frac{c}{\mu}\delta_h + \varepsilon_{\mathsf{eq}}$, and the value bound (51) acquires an additional additive term of $H\,\varepsilon_{\mathsf{eq}}$. Consequently, by Theorem 4, the value error bound holds:

$$\sup_{x\in\mathcal{X}}\left|V_1^{i,\pi}(x) - V_1^{i,\pi^\star}(x)\right| \leq \frac{c}{\mu}\sum_{t=1}^{H}(L_{\mathsf{env}})^{t-1}Q_{\max,t}\delta_t. \tag{51}$$

If $\delta_t \leq \bar{\delta}$ and $Q_{\max,t} \leq H$ for all t, then the following bound also holds:

$$\sup_{x\in\mathcal{X}}\left|V_1^{i,\pi}(x) - V_1^{i,\pi^\star}(x)\right| \leq \frac{cH\bar{\delta}}{\mu}\left(\sum_{t=0}^{H-1}(L_{\mathsf{env}})^t\right). \tag{52}$$

Remark 4 (Risk Sensitivity and Comparison with Nash)**.** The bound in (52) highlights the stability benefits of RQRE. When the risk measure is coherent or risk-neutral we have $L_{\mathsf{env}} = 1$ (Artzner et al., 1999), and recover a polynomial dependence $O(H^2\bar{\delta}/\mu)$. For general convex risk measures where $L_{\mathsf{env}} > 1$, the error bound scales exponentially with the horizon, reflecting the inherent difficulty of risk-sensitive planning. However, crucially, this error remains *Lipschitz continuous* with respect to the estimation error $\bar{\delta}$. Example 2 shows that Nash equilibria can exhibit $O(1)$ policy jumps from arbitrarily small payoff perturbations at even a single stage, making the analogous bound for Nash-based algorithms potentially unbounded.

E Proofs for Distributional Robustness of RQRE

We begin by showing a result which demonstrates the equivalence of the penalty and constraint formulations of the convex risk measure. What follows is a simplified version of results in (Föllmer & Schied, 2002).

Proposition 4 (Penalty–Constraint Duality for Convex Risk Measures)**.** Let $\rho : \mathcal{Z} \to \mathbb{R}$ be a convex risk measure with dual representation $\rho(Z) = \sup_{p\in\Delta(\Omega)}\{\mathbb{E}_p[-Z] - \varphi(p)\}$ as in Theorem 1, where $\varphi : \Delta(\Omega) \to (-\infty,\infty]$ is a convex, lower-semicontinuous penalty function. Then for any bounded random variable Z, the penalized problem is equivalent to the constrained problem

$$\sup_{p\in\Delta(\Omega)}\left\{\mathbb{E}_p[-Z] - \varphi(p)\right\} = \sup_{p\in\Delta(\Omega)}\left\{\mathbb{E}_p[-Z] \text{ s.t. } \varphi(p)\leq\delta\right\}$$

for a radius $\delta = \delta(\rho, Z)$ uniquely determined by the optimality conditions, and always larger than the minimum value of φ. The risk parameter governing ρ acts as the Lagrange multiplier of the constraint.

Proof. The equivalence follows by Lagrangian duality. Introduce the multiplier $\lambda \geq 0$ for the constraint $\varphi(p) \leq \delta$ and write

$$\begin{aligned}\sup_{p\in\Delta(\Omega)} \left\{ \mathop{\mathbb{E}}_{p}[-Z] \text{ s.t. } \varphi(p) \leq \delta \right\} &= \sup_{p\in\Delta(\Omega)} \left\{ \mathop{\mathbb{E}}_{p}[-Z] + \min_{\lambda\geq 0} \lambda(\delta - \varphi(p)) \right\} \\ &= \min_{\lambda\geq 0} \sup_{p\in\Delta(\Omega)} \left\{ \mathop{\mathbb{E}}_{p}[-Z] - \lambda\varphi(p) \right\} + \lambda\delta \\ &= \min_{\lambda\geq 0} \phi(\lambda) + \lambda\delta,\end{aligned}$$

where the second equality is due to strong duality, which holds by Slater's condition as the minimum value of φ is always a feasible point. Note that the above implies that for any fixed δ, we can recover the equivalent λ by minimizing $\phi(\lambda) + \lambda\delta$. Thus the penalized version of the problem is equivalent to the hard constraint $\varphi(p) \leq \delta$. □

We now prove Proposition 2 by showing how each component of the RQRE objective corresponds to a distributional robustness guarantee, using Proposition 4 as the key tool.

Bounded rationality as policy robustness. The bounded rationality component of the RQRE objective assigns each player a regularized best response of the form

$$\pi_i \in \arg\max_{\mu_i\in\Delta(\mathcal{A}_i)} \left\{ \mathop{\mathbb{E}}_{a\sim(\mu_i,\pi_{-i})}[u_i(a)] + \tfrac{1}{\epsilon_i}\nu_i(\mu_i) \right\}.$$

Applying Proposition 4 with $\varphi = \nu_i$ and $\lambda = 1/\epsilon_i$, this is equivalent to

$$\pi_i \in \arg\max_{\mu_i\in\Delta(\mathcal{A}_i)} \left\{ \mathop{\mathbb{E}}_{a\sim(\mu_i,\pi_{-i})}[u_i(a)] \text{ s.t. } \nu_i(\mu_i) \leq \delta_{\mathsf{pol}} \right\},$$

where the radius $\delta_{\mathsf{pol}} = \delta_{\mathsf{pol}}(\epsilon_i)$ is uniquely determined by the optimality conditions, with $1/\epsilon_i$ serving as the Lagrange multiplier of the constraint. Thus bounded rationality corresponds to distributionally robust policy selection where each agent maximizes expected payoff over all policies with a ν_i-ball of radius δ_{pol}. As $\epsilon_i \to \infty$, the multiplier vanishes, the constraint tightens, and the Nash best response is recovered. Conversely, as $\epsilon_i \to 0$, the ball expands and the policy approaches the maximizer of ν_i (e.g., the uniform distribution when ν_i is the negative entropy).

Strategic robustness from risk aversion. In risk averse Markov games, the policy risk operator can be written as $\mathcal{R}^{\mathrm{pol}}_{i,h}\big(Q^i_h, \pi_h; x\big) := \rho^{\mathsf{p}}_{i,h}\big(Z^{\mathsf{p}}_{i,h} \,\big|\, \pi_{-i,h}(\cdot \mid x)\big)$, where

$$\rho^{\mathsf{p}}_{i,h}(Z \mid \pi_{-i}) = \sup_{p_i\in\mathcal{P}^{\mathsf{p}}_i} \{-\langle Z, p_i\rangle - \varphi_{\mathsf{p},i}(p_i, \pi_{-i})\},$$

is the risk measure given in dual form. Applying Proposition 4 to this penalized problem yields the constrained formulation

$$\rho^{\mathsf{p}}_{i,h}(Z \mid \pi_{-i}) = \sup_{p_i\in\mathcal{P}^{\mathsf{p}}_i} \{-\langle Z, p_i\rangle \;\mid\; \varphi_{\mathsf{p},i}(p_i, \pi_{-i}) \leq \delta_{\mathsf{opp}}\}.$$

Environment robustness from risk aversion. Similarly, we write the environment risk operator in terms of the environment risk measure

$$\rho^{\mathsf{e}}_{i,h}(Z|\mathcal{P}) = \inf_{\widetilde{\mathcal{P}}\in\mathcal{P}(\mathcal{X})} \left\{ \int Z(x')\,\widetilde{\mathcal{P}}(dx') + \varphi_{\mathsf{e},i}(\widetilde{\mathcal{P}} \,\|\, \mathcal{P}) \right\}$$

which, due to Proposition 4, has the constrained formulation

$$\rho^{\mathsf{e}}_{i,h}(Z|\mathcal{P}) = \inf_{\widetilde{\mathcal{P}}\in\mathcal{P}(\mathcal{X})} \left\{ \int Z(x')\,\widetilde{\mathcal{P}}(dx') \text{ s.t. } \varphi_{\mathsf{e},i}(\widetilde{\mathcal{P}} \,\|\, \mathcal{P}) < \delta_{\mathsf{env}} \right\}.$$

Here, we replace the $\sup$ with $\inf$ and flip the sign on the penalty function.

In both these settings, the radii $\delta_{\text{opp}}, \delta_{\text{env}}$ is determined by by the optimality conditions of the penalized-constrained duality, with the Lagrange multiplier governed by the structure of $\varphi_{\text{p},i}$ and $\varphi_{\text{e},i}$. The first, δ_{opp}, can be viewed as defining a radius of opponent actions we are risk averse to. Similarly, δ_{env} can be viewed as defining a radius of environment transitions. We usually set φ to be scaled by a risk aversion parameter $1/\tau$, and thus as $\tau \to \infty$ the radius δ will grow to cover the entire space of measures (either over the opponents actions or the environment transition). As $\tau \to 0$ we recover the risk neutral Markov game setting, where risk averse utilities become utilities and risk averse value functions become value functions.

F Additional Experimental Details

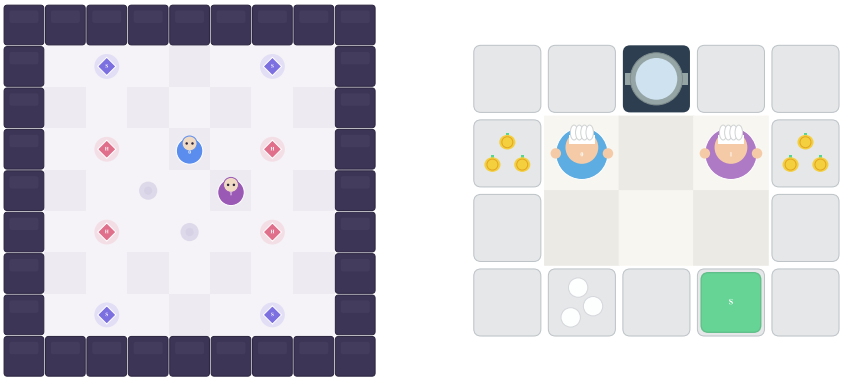

Figure 4: Dynamic Stag Hunt (left) and Overcooked (right) environments used in experiments.

F.1 Dynamic Stag Hunt

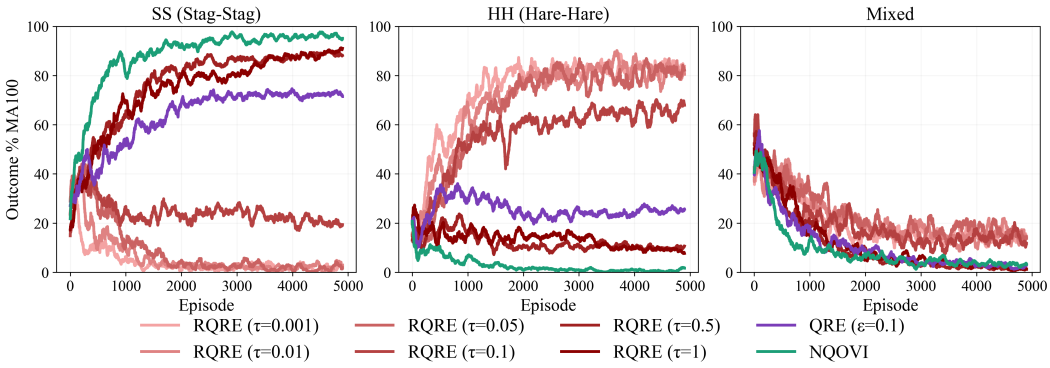

Figure 5: **Stag Hunt outcome distributions during training.** Each panel shows the fraction of stag–stag, hare–hare, and mixed interaction outcomes (rolling average) for a given algorithm and risk-aversion level. `NQ-OVI`, `QRE-OVI`, and low risk averse `RQRE-OVI` agents converge to payoff-dominant stag–stag outcomes, while highly risk averse agents converge to risk dominant hare–hare, confirming the expected equilibrium selection.

Environment 1: Dynamic Stag Hunt. We evaluate on a Dynamic Stag Hunt environment based on the Melting Pot suite (Agapiou et al., 2022) implemented as a 9×9 grid-world with episodes of 75 timesteps. At each step, agents choose from six actions: four cardinal movements, stay, and interact. The grid contains four *stag* resources placed in the corners and four *hare* resources adjacent

to the spawn points; agents pick up a resource by walking over it and may swap by stepping onto a different type. An interaction resolves when at least one agent chooses the interact action while both agents are adjacent and carrying a resource. The resulting payoffs follow the classic Stag Hunt matrix: mutual stag yields $(4, 4)$, mutual hare yields $(2, 2)$, and stag-hare mismatch gives $(0, 2)$ with 0 to the stag-holder and 2 to the hare-holder. Upon interaction, both agents are respawned randomly at one of the spawn points with empty inventories, allowing multiple interactions per episode. The spatial structure introduces a coordination challenge beyond the matrix game: hare resources are close to spawn, providing a safe default, while stag resources require navigating to the corners, demanding both spatial and strategic coordination.

Each state is represented by a 15-dimensional observation vector capturing both players' normalized grid positions, Manhattan distances to the nearest stag and hare resources, inventory indicators (one-hot encoding of whether each agent carries a stag, hare, or nothing), normalized inter-agent distance, a timing feature (fraction of episode elapsed), and bias term. The joint state-action feature vector $\phi(x, a_1, a_2) \in \mathbb{R}^{200}$ is constructed by concatenating the observation, one-hot action encoding for both players, observation-action cross products for each player, and an action-action cross products.

F.1.1 Dynamic Stag Hunt Training details

We train all agents for $K = 5000$ episodes using optimistic value iteration with ridge regression ($\lambda = 1.0$) and exploration bonus coefficient $\beta = 0.1$. The experience buffer stores the most recent 1000 transitions and parameters are updated every 20 episodes via batch least-squares regression over the buffered data. The reward scale is set to 4.0, matching the maximum single-interaction payoff. For QRE and RQRE agents, the stage-game equilibrium solver runs for up to $T = 100$ fixed-point iterations with early stopping at tolerance 10^{-6}, using bounded rationality $\epsilon = 0.05$ for both players. RQRE agents use symmetric risk-aversion parameters $\tau_1 = \tau_2 = \tau$, swept over $\{0.005, 0.01, 0.05, 0.1, 0.5, 1.0\}$. All experiments use a single seed with results averaged over 200 evaluation rollouts for cross-play and perturbation experiments.

F.2 Overcooked

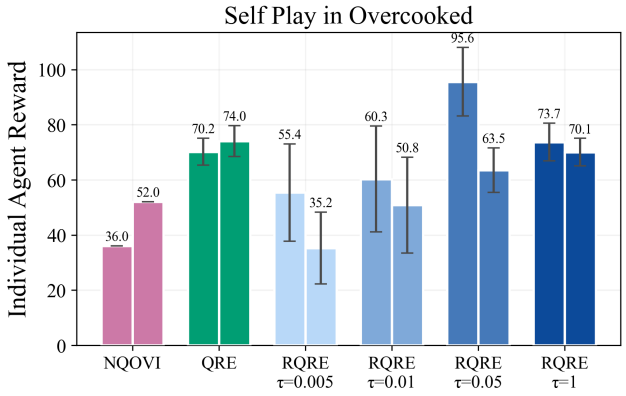

Figure 6: Self play performance for each algorithm in Overcooked over 200 evaluation rollouts.

Environment 2: Overcooked. We also evaluate on the Overcooked environment (Gessler et al., 2025); specifically, we use the version introduced in JaxMARL (Rutherford et al., 2024) with custom visualizations. Two agents operate in a small kitchen and must cooperate to prepare and deliver onion soups: pick up onions from piles, place them in a pot (which requires three onions to begin cooking), retrieve the finished soup with a plate, and deliver it to a serving location for a sparse reward of 20. Each agent chooses from six actions (four cardinal movements, stay, and interact) over a horizon of 100 timesteps. The environment provides both sparse delivery rewards and shaped rewards for intermediate subtask completion (e.g., placing an onion in the pot, picking up a plate when

soup is cooking). The tight layout and sequential subtask structure make coordination essential, as agents must avoid blocking each other while efficiently dividing labor.

The observation is a 49-dimensional vector encoding both agents' positions, orientations, and inventory states, pot status (number of onions, cooking progress, readiness), per-agent proximity features to key locations (pot, onion pile, plate pile, serving goal), a time fraction, and hand-crafted goal-potential features that capture task-relevant interactions. The joint state-action feature vector $\phi(x, a_1, a_2) \in \mathbb{R}^{685}$ concatenates the observation, one-hot action encodings, and cross products between a selected subset of a core observation features and each agents' action.

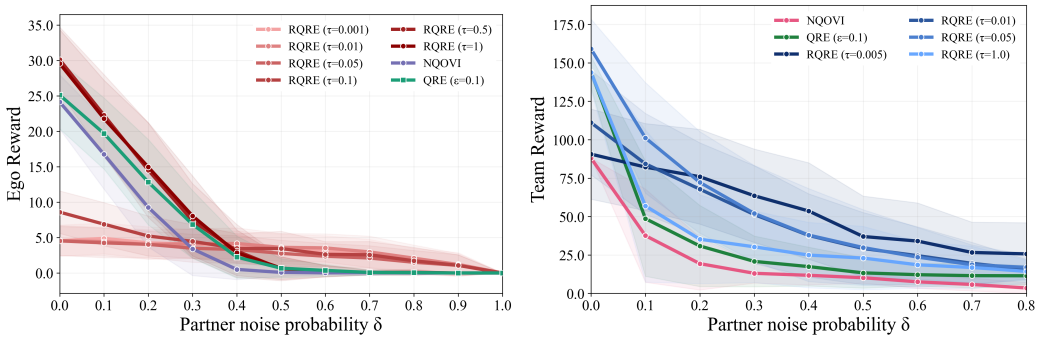

Figure 7: **Cross-play with a perturbed partner.** Team reward as a function of the partner noise for Stag Hunt (left) and Overcooked (right). At each evaluation step, the perturbed partner's action is replaced with a fixed deterministic action (e.g., always move in a single direction) with probability δ, and the trained policy is executed otherwise. This deterministic perturbation ensures that deviations are high-signal rather than averaging out under uniform noise. The results are averaged over 200 evaluation rollouts at each noise level.

F.2.1 Overcooked Training details

Agents are trained for $K = 5000$ episodes with ridge parameter $\lambda = 5.0$ (larger than Stag Hunt due to the higher feature dimension $d = 685$) and exploration bonus $\beta = 0.1$. The buffer stores 500 transitions and updates occur every 10 episodes. Training rewards combine sparse delivery rewards and shaped intermediate rewards with equal weighting (both coefficients set to 1.0), and the reward scale is set to 20.0 to match the delivery reward. The QRE solver runs for up to $T = 100$ iterations with early stopping at tolerance 10^{-6}, with $\epsilon_1 = \epsilon_2 = 0.1$. RQRE agents use symmetric $\tau_1 = \tau_2 = \tau$ swept over $\{0.005, 0.01, 0.05, 1.0\}$. The environment runs in cooperative mode, where both agents receive identical sparse rewards upon delivery.

F.3 Hyperparameters

We provide the hyperparameters used in each experiment below.

Table 1: Hyperparameters for Stag Hunt experiments

Category	Hyperparameter	Value
Setup	Episodes (K)	5000
	Feature dimension (d)	200
	Ridge parameter (λ)	1.0
	Exploration bonus (β)	0.1
	Buffer size	1000
	Update frequency	10
	Seed	42
Environment	Horizon (H)	75
	Actions per agent	6
	Grid size	9×9
Rewards	Stag–Stag payoff	$(4, 4)$
	Hare–Hare payoff	$(2, 2)$
	Stag–Hare payoff	$(0, 2)$
	Step cost	0
RQRE	Bounded rationality (ϵ)	0.1
	Risk aversion (τ)	$\{0.005, 0.01, 0.05, 1.0\}$
	Solver tolerance	10^{-6}
	Max solver iterations (T)	100

Table 2: Hyperparameters for Overcooked experiments

Category	Hyperparameter	Value
Setup	Episodes (K)	5000
	Feature dimension (d)	685
	Ridge parameter (λ)	5.0
	Exploration bonus (β)	0.1
	Buffer size	500
	Update frequency	10
	Seed	42
Environment	Horizon (H)	100
	Actions per agent	6
	Layout	Cramped Room
Rewards	Placement in pot	3
	Plate pickup	4
	Soup pickup	8
	Delivery completion	6
	Soup drop penalty	-1
RQRE	Bounded rationality (ϵ)	0.1
	Risk aversion (τ)	$\{0.005, 0.01, 0.05, 1.0\}$
	Solver tolerance	10^{-6}
	Max solver iterations (T)	100