Training Generalizable Collaborative Agents via Strategic Risk Aversion

Chengrui Qu
California Institute of Technology
cqu@caltech.edu

Yizhou Zhang
California Institute of Technology
yzhang8@caltech.edu

Nicolas Lanzetti
California Institute of Technology
lnicolas@caltech.edu

Eric Mazumdar
California Institute of Technology
mazumdar@caltech.edu

Abstract

Many emerging agentic paradigms require agents to collaborate with one another (or people) to achieve shared goals. However, existing approaches to learning policies for such collaborative problems produce brittle solutions that fail when paired with new partners. We attribute these failures to a combination of *free-riding* during training and a lack of *strategic robustness*. To address these problems, we study the concept of strategic risk aversion and interpret it as a principled inductive bias for generalizable cooperation with unseen partners. While strategically risk-averse players are robust to deviations in their partner's behavior by design, we show that, in collaborative games, they also (1) can have better equilibrium outcomes than those at classical game-theoretic concepts like Nash, and (2) exhibit less or no free-riding. Inspired by these insights, we develop a multi-agent reinforcement learning algorithm that integrates strategic risk aversion into standard policy optimization methods. Our empirical results across collaborative benchmarks (including an LLM collaboration task) validate our theory and demonstrate that our approach consistently achieves reliable collaboration with heterogeneous and previously unseen partners across collaborative tasks.

1 Introduction

AI systems increasingly operate in multi-agent environments, where success depends on effective interaction with other agents. An increasingly important subset of these problems involves agents solving *collaborative tasks*—i.e., tasks where agents must work together towards a shared goal. Examples of such problems range from robots coordinating with people in shared physical spaces like warehouses [25], to emerging "agentic AI" problems where multiple models collaborate to write code [67] or solve math problems [44].

A common approach to such problems is to view them as collaborative *games* and use concepts from game theory to design algorithms to learn collaborative policies [10]. Viewed through this lens, a central challenge in collaborative games is *partner generalization*—where one seeks to maintain effective interaction across new sets of partners [6, 23, 57, 49]. Indeed, in many real-world settings, agents must navigate situations in which they interact with different partners (either algorithmic or human) who may have slightly different goals, heuristics, or levels of competence. Agents should learn strategies that work well across a broad range of partners without sacrificing performance. Unfortunately, current approaches tend to struggle with this—with learned policies often failing to generalize to new partners [56], or *overfitting* to other agents' eccentricities and becoming overly reliant on specific conventions [6, 32].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We view the problem of partner generalization as ultimately a question of *robustness* of the learned policy and of *alignment* between agents [31]. The need for robustness emerges from the requirement that learned strategies be insensitive to changes in a partner's strategy and degrade gracefully in performance as the partners become less collaborative. The need for alignment emerges from the desire to find agents that contribute proportionally to the task. If agents learn to under-contribute or delegate costly actions [58, 35] to their partners—effectively *free-riding* on their partner's effort—they cannot generalize to new environments and partners. While free-riding is a widely acknowledged phenomenon in game theory, it is rarely studied as a problem in collaborative multi-agent learning despite recent empirical evidence that it arises even with large AI models [76, 34].

Prior work—which we review in Section 2 and then in depth in Appendix A—has studied partner generalization in specific settings that fail to scale to large-scale problems like the post-training of LLM agents [23] or via largely heuristic approaches [17]. In contrast, in this paper we advance *strategic* risk aversion as a *principled* and *scalable* approach to addressing this issue in a broad class of collaborative games. Recently proposed in multi-agent reinforcement learning (MARL) to derive human-like strategies in general games [38], strategic risk aversion mirrors behaviors observed in people in experimental economics [20]. Unlike many approaches to risk aversion or robustness in MARL which primarily focus on robustness to uncertainty in the underlying environment, *strategic* risk aversion (robustness) posits that agents should be risk-averse (robust) to uncertainty stemming from their opponents' decisions [38, 29]. This perspective aligns with the goal of partner generalization, as it forces agents to be robust to deviations in their partner's play.

Formalizing this principle leads to a new equilibrium concept: (Strategically) Risk-Averse Quantal Response Equilibria (RQE) [38]. While RQE has been shown to better capture human play in simple experiments and offer better computational tractability [74] than classic game theoretic concepts, its potential for collaborative games and partner generalization remains unknown. We argue that strategic risk aversion provides a suitable inductive bias for robust cooperation in collaborative games: at an RQE, an agent is trained against a structured set of *plausible* partner deviations, governed by the agent's degree of risk aversion. This ensures strategies are robust without being overly conservative. Remarkably, although strategic risk aversion is introduced for *robustness*, we prove that it can also lead to better *alignment* between agents.

Contributions: Through both a theoretical analysis of RQE in structured collaborative games and extensive experiments across a range of diverse multi-agent benchmarks, we validate that RQE are particularly suited for collaborative tasks. In particular, our contributions are as follows:

- **Incentivizing collaboration (Theorem 4.1):** We prove that in continuous quadratic aggregative games, strategic risk aversion can encourage collaboration, with higher risk aversion leading to a greater focus on the shared goal. This implies a counterintuitive result: unlike in classic robust optimization or RL, *robustness does not necessarily require sacrificing performance*.
- **Alleviating free-riding (Theorem 4.4):** We prove in finite-action collaborative games with private costs that strategic risk aversion can mitigate free-riding at equilibrium. *Thus, risk aversion can reduce the alignment problem in collaborative MARL.*
- **Scalable risk-averse MARL algorithms:** We develop *Strategically Risk-Averse Policy Optimization* (SRPO), a MARL algorithm that optimizes an RQE-derived objective which integrates naturally with policy-optimization algorithms like independent proximal policy optimization (IPPO).
- **Empirical validation:** Across collaborative MARL benchmarks, SRPO *consistently* achieves more reliable coordination with heterogeneous and unseen partners than IPPO (the current scalable MARL baseline). We observe that IPPO consistently learns free-riding equilibria that limit its generalization. In contrast and in line with our theory, SRPO converges to RQE-like behavior, eliminates free-riding, and improves partner generalization. We also provide preliminary small-scale experiments showing that these findings can even extend to the fine-tuning of collaborative agentic AI teams made up of large language models, demonstrating the scalability of our approach.

2 Related Work

Our work studies strategic risk aversion in collaborative MARL. We cover the most relevant related work here and defer a more in depth discussion to Appendix A.

The most related literature in collaborative MARL studies partner generalization: learning collaborative policies that remain effective when paired with previously unseen partners [3, 6, 12]. This

problem is also studied as zero-shot coordination [23, 24], or ad-hoc teamwork problem [56]. A major class of approaches is population-based training [65, 77, 72, 66, 47, 37, 57, 64, 9], where agents train against a diverse set of learned partners, essentially performing domain randomization for partners, with methods differing mainly in how they construct the population and how they train against it. Although simple, these methods require nontrivial partner-population design, often rely on task-specific heuristics, and can be costly in high-dimensional settings such as LLM fine-tuning. They also offer limited guarantees of improved generalization. Empirically, we find that fictitious co-play [57], a representative population-based method, requires substantially more compute yet still performs poorly in our setting. To avoid these costs, Forkel et al. [17] increase entropy regularization to induce robustness. While scalable, this approach does not eliminate free-riding and therefore does not fully address collaborative generalization.

In contrast, our approach based on strategic risk aversion scales as a simple modification to existing policy-optimization techniques, is grounded in a principled game-theoretic formulation, and empirically consistently delivers more stable cross-play performance across collaborative benchmarks. Rather than approximating robustness through brute-force partner diversification, our approach directly optimizes a risk-averse robustness objective. Although risk aversion has long been studied in experimental economics [21, 19, 20], strategic risk aversion has only recently been studied in game theory [29, 38], control [63, 15, 45], and MARL [74, 75, 55]. Existing work mainly focuses on tractability and Q-learning-based computation of risk-averse equilibria [74, 75], whose scalability remains unclear. In contrast, we study the effects of risk qualitatively in *collaborative* problems. We prove two free-lunch theorems (Theorems 4.1 and 4.4) characterizing the effect of risk on welfare and payoffs and propose a policy optimization-based approach that can be readily incorporated into existing code-bases.

3 Problem Setup

We consider a general-sum collaborative game with N players. Each player $i \in [N]$ is endowed with actions $a_i \in \mathcal{A}_i$, where the action spaces can be discrete or continuous, and we use the notation $\mathcal{A}_{-i} = \prod_{j \neq i} \mathcal{A}_j$ for the action space of all players except i. Each player aims to maximize a utility function $u_i : \mathcal{A}_i \times \mathcal{A}_{-i} \to \mathbb{R}$. The utility function consists of a shared reward R, which coincides for all players, and a private cost c_i, so that

$$u_i(a_i, a_{-i}) = R(a_1, \dots, a_N) - c_i(a_i).$$

These games are collaborative since all players seek to maximize the reward, but pay for their own effort. They have been well studied in economics under the guise of public goods games and naturally model many real-world scenarios where players aim to collaboratively accomplish a task but are penalized for their own effort. The players select a mixed strategy $x_i \in \Delta(\mathcal{A}_i)$, which is a probability over the available actions $\mathcal{A}_i$ (here, $\Delta(\mathcal{A}_i)$ is the set of probability distributions over $\mathcal{A}_i$). In the standard game-theoretic setting, the utility of each player is the standard risk-neutral expected utility

$$U_i(x_i, x_{-i}) = \mathbb{E}_{a_1 \sim x_1, \cdots, a_N \sim x_N}[u_i(a_i, a_{-i})],$$

where x_{-i} is the collection of the mixed strategies of all other players $j \neq i$.

In this work, we follow Mazumdar et al. [38] and introduce risk aversion and entropy regularization.

Risk aversion To model risk aversion, we consider players that seek to optimize a risk-adjusted utility based on the *entropic risk measure* with risk aversion parameter $\tau_i > 0$ (e.g., see Föllmer and Schied [16]), instead of their expected utility. Players' resulting objectives are given by

$$U_i^{\tau_i}(x_i, x_{-i}) = \inf_{p \in \Delta(\mathcal{A}_{-i})} U_i(x_i, p) + \frac{1}{\tau_i} \mathrm{KL}(p, x_{-i}). \tag{1}$$

In words, the utility results from a worst-case approach, in which a fictitious adversary tries to inflict maximum damage at the (risk-neutral) expected utility, while not deviating "too much" from the strategies of all other players. For the sake of this paper, we will measure this deviation with the KL divergence, defined as $\mathrm{KL}(p, q) = \sum_j p_j \log\left(\frac{p_j}{q_j}\right)$ (when $\mathcal{A}_{-i}$ is discrete) and $\mathrm{KL}(p, q) = \int_{\mathbb{R}^n} \log\left(\frac{\mathrm{d}p}{\mathrm{d}q}(x)\right) \mathrm{d}p(x)$ (when $\mathcal{A}_{-i} = \mathbb{R}^n$ is continuous), where $\frac{\mathrm{d}p}{\mathrm{d}q}$ is the density of p with respect to q (provided it exists, else $\mathrm{KL}(p, q) = +\infty$). Our choice of the entropic risk measure is rooted

in operations research, but also motivated by ubiquitous use of the KL divergence across machine learning and in particular in policy optimization in reinforcement learning. Using the KL divergence, we derive algorithms that can be implemented with minimal changes to existing codebases.

Entropy regularization We add entropy to each player's utility, so that their utility is

$$U_i^{\tau_i,\epsilon_i}(x_i, x_{-i}) = U_i^{\tau_i}(x_i, x_{-i}) - \epsilon_i H(x_i), \tag{2}$$

where $\epsilon_i > 0$ and $H(x_i)$ is the (negative) entropy of the mixed strategy x_i; i.e., $H(p) = \sum_j p_j \log(p_j)$ (if $\mathcal{A}_i$ is discrete) and $H(p) = \int_{\mathbb{R}^n} \rho_p(x) \log(\rho_p(x)) \mathrm{d}x$ (when $\mathcal{A}_i = \mathbb{R}^n$ is continuous), where ρ_p is again the density of p (provided it exists, else $H(p) = +\infty$). Entropy regularization of the utility leads to the celebrated quantal response model in behavioral economics [39] and is widely employed in machine learning for exploration purposes (e.g., see Ahmed et al. [1]). For more general regularization schemes, we refer to Mazumdar et al. [38].

Risk-averse quantal response equilibrium (RQE). Given our utility $U_i^{\tau_i,\epsilon_i}(x_i, x_{-i})$, we define an RQE as a set of mixed strategies from which no player has an incentive to unilaterally deviate:

Definition 3.1 (Mazumdar et al. [38], Definition 5)**.** A tuple of mixed strategies $(x_1^\star, \ldots, x_N^\star)$ is a risk-averse quantal response equilibrium (RQE) with degrees of risk aversion $\tau_1, \ldots, \tau_N$ and entropy regularization $\epsilon_1, \ldots, \epsilon_N$ if for all players i we have

$$U_i^{\tau_i,\epsilon_i}(x_i, x_{-i}^\star) \leq U_i^{\tau_i,\epsilon_i}(x_i^\star, x_{-i}^\star) \, \forall \, x_i \in \Delta(\mathcal{A}_i). \tag{3}$$

When $\tau_i, \epsilon_i \to 0$ we recover Nash equilibria, when $\tau_i \to 0$ we recover quantal response equilibria, and when $\epsilon_i \to 0$ and $\tau_i \to \infty$ we recover security strategies. In the next section, we investigate the benefits of risk aversion, together with entropy regularization, in two classes of collaborative games: aggregative continuous quadratic games and finite symmetric collaborative games. These insights prompt us, in the subsequent section, to use RQE to design algorithms to train collaborative agents.

4 "Free-Lunch" Theorems for Strategic Risk Aversion

We first study strategic risk aversion in simple, *structured* classes of collaborative games which are more tractable to analyze than complex MARL problems we ultimately design algorithms for. These simple games act as a theoretical laboratory, allowing us to rigorously prove two "free-lunch" theorems that illustrate how strategic risk aversion, besides ensuring robustness, can (1) increase collaboration and (2) mitigate free-riding—both of which are essential to the task of partner generalization. These results serve as a principled rationale for incorporating strategic risk aversion into MARL. We empirically observe both these properties across our experiments in Section 6, validating that our insights extend beyond structured games.

4.1 Strategic Risk Aversion Can Induce Collaboration

Our first setting lies in continuous ($\mathcal{A}_i = \mathbb{R}^n$) quadratic aggregative games, popular in economics [8] and control [42]. In line with our collaborative setting, we write the utility of each player as the difference between a shared reward R that depends on the aggregate action $a_1 + \ldots + a_N$, and a private cost c_i that depends on the player's own action a_i:

$$R(a_1, \ldots, a_N) = \frac{1}{2} \left\langle \sum_{i=1}^N a_i, H \sum_{i=1}^N a_i \right\rangle + \left\langle h, \sum_{i=1}^N a_i \right\rangle, \; c_i(a_i) = \frac{\rho_i}{2} \|a_i\|^2, \; \forall i \in [N] \tag{4}$$

for some H negative definite, $h \neq 0$ of appropriate dimensions, and $\rho_i > 0$. For simplicity of exposition, we assume that all players have the same degree of risk aversion τ_i, bounded rationality ϵ_i, and ρ_i. As we show in Section B.1, all results extend to the case where these are player-dependent.

Being a continuous game, the computation of an RQE is infinite dimensional, as it involves searching over mixed strategies over continuous action spaces. In Proposition B.3 in the appendix, however, we bypass this complexity by showing that there is a unique RQE (at which players have finite risk-adjusted utility) and that the equilibrium mixed strategies are Gaussian distributions, whose mean and variance can be computed efficiently. This effectively allows us to study the effect of risk aversion on the expected shared reward $J(\tau) \coloneqq \mathbb{E}_{a_i \sim x_i^\star(\tau)} \left[R(a_1, \ldots, a_N) \right]$.

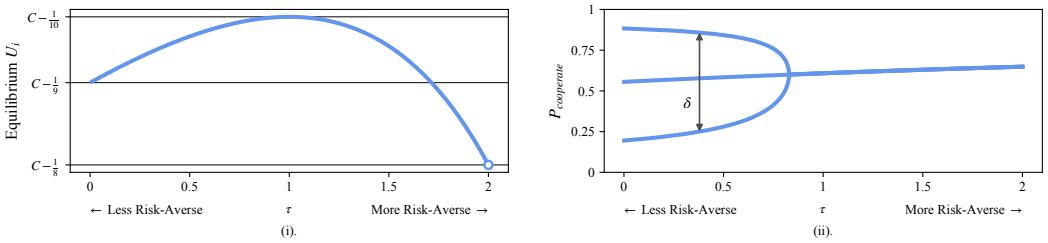


Figure 1: (i). Expected utility of each player as a function of the degree of risk aversion τ at equilibrium, when $\epsilon = 1$ and $\bar{a} = 1$ in Example 4.2. $C = -3\epsilon/4$ is a τ-independent constant. Player's utilities can first *increase* with risk aversion before decreasing due to over-conservatism, meaning that strategic risk aversion can yield better performing RQE. (ii). Probability that a player collaborates at an RQE as a function of the level of risk aversion, for $\epsilon = 0.2$ for the game in Example 4.5. The arrow shows the gap in collaboration probability between the two players at a free-riding equilibrium, which is proportional to the degree of free-riding. Strategic risk aversion alleviates free-riding entirely after a given threshold (i.e., $\delta \to 0$) as our theory predicts.

Remarkably, as we show in the next theorem (proven in Section B.1), this function is strictly increasing. That is, risk aversion *monotonically* increases the shared reward and thus induces collaboration.

Theorem 4.1 (risk induces collaboration)**.** *Let $x_i^\star(\tau)$ be the mixed strategy of player i at the unique RQE of the game, as a function of the degree of risk aversion τ. Then, the expected shared reward $\tau \mapsto J(\tau)$ is strictly increasing.*[1] *That is, players contribute more to the shared reward as they become more risk-averse.*

This increasing contribution to the shared reward generally entails an increase in the players' private cost. Thus, in practice, we observe a *tradeoff* between being risk neutral ($\tau \to 0$) and more risk-averse (τ large) with large personal cost but large shared reward. In a one-dimensional example, we can study this tradeoff analytically.

Example 4.2. Consider two robots that move an object from the origin to a target location $\bar{a}$. Each robot exerts a force a_i, so that the total force is $a_1 + a_2$. Players incur a penalty for their own actuation effort. Thus, the shared reward is $-\frac{1}{2}(a_1 + a_2 - \bar{a})^2$ and the personal cost is $\frac{1}{2}a_i^2$. In this case, for $\tau < 2/\epsilon$, the RQE consists of the Gaussian mixed strategies with mean $m_i^\star(\tau) = \frac{\bar{a}}{3-\tau\epsilon/2}$ and variance $\Sigma_i = \frac{\epsilon}{2}$. The equilibrium expected utility is concave in τ as shown in Figure 1, highlighting that risk aversion can yield higher utilities.

We refer to Theorem 4.1 as a "free-lunch" theorem because it suggests that one can add risk aversion into a game and *not* incur a loss in performance. This can be observed in Figure 1(i) in which equilibrium utility increases from a risk-neutral baseline before decreasing again, and is further validated in the practical settings with aggregative structure in Section 6.1. This is in stark contrast to single-agent RL or robust optimization where robustness sacrifices performance. We validate our theory in broader classes of games in Section 6 and observe that, in some games, some degree of strategic risk aversion can increase collaboration and performance over a risk-neutral baseline.

4.2 Strategic Risk Aversion Alleviates Free-riding

We now prove a second "free-lunch" theorem for strategic risk aversion in collaborative games: strategically risk-averse players free-ride less at equilibrium. We establish this result in the context of 2-player collaborative *finite* symmetric games played over the probability simplex. The reward R is symmetric, i.e., $R(a_1, a_2) = R(a_2, a_1)$, and both players share the same private cost function $c(a_i)$.

Although the game is symmetric, its equilibria can be asymmetric and exhibit *free-riding*: one player may exert significantly less effort while still enjoying high utility. As our experiments show, such equilibria are ubiquitous across many MARL benchmarks and pose a significant challenge to the generalization of agents' strategies. Concretely, we define free-riding in our game as follows.

[1] Here, $\tau \in (0, \bar{\tau})$ with $\bar{\tau} = \frac{\rho + \lambda_{\max}(-H)}{(N-1)\epsilon\lambda_{\max}(-H)}$, and uniqueness is among RQE at which all players have finite risk-adjusted utility. Degenerate RQE can exist: if all players play Gaussian strategies with sufficiently large covariances, every player's risk-adjusted utility is $-\infty$ regardless of their own strategy; see Section B.1.

Definition 4.3 (free-riding)**.** An RQE $(x_1, x_2) \in \Delta(\mathcal{A}) \times \Delta(\mathcal{A})$ exhibits free-riding with degree $\delta \geq 0$ if the difference in costs paid by each player at the RQE is δ; i.e., $|\mathbb{E}_{a_1 \sim x_1}[c(a_1)] - \mathbb{E}_{a_2 \sim x_2}[c(a_2)]| = \delta$. If $\delta = 0$, the RQE does not exhibit free-riding.

We now show that as the degree of risk aversion increases, free-riding becomes less prevalent.

Theorem 4.4 (risk removes free-riding)**.** *There exists a game-dependent constant C depending only on R, c, and the degree of bounded rationality ϵ such that if players' degrees of risk aversion τ in a two-player collaborative game satisfy $\tau > \frac{C}{\delta^2}$, then the game admits no RQE with degree of free-riding greater than δ.*

This theorem (proven in Section B.2) shows that as strategic risk aversion increases, RQE must exhibit less free-riding. The intuition for this result is as follows: suppose that player i free-rides at an RQE. If they become risk-averse, the worst-case deviation for their opponent is simply to stop putting any effort into the game. Since player i is free-riding and not putting their own effort, this deviation causes a large drop in performance. Thus, at a risk-averse equilibrium, a player must contribute some of their own effort. We validate this intuition in a simple coordination game.

Example 4.5. Consider the game with $\mathcal{A} = \{\text{C} = \text{collaborate}, \text{D} = \text{defect}\}$, $R(a_1, a_2) = 1$ if $a_1 = \text{C}$ or $a_2 = \text{C}$ (at least one player collaborates) and $R(a_1, a_2) = 0$ if $a_1 = a_2 = \text{D}$ (both defect), and $c(a_i) = 0.4$ if $a_i = \text{C}$ (cost of collaboration) and $c(a_i) = 0.0$ otherwise. We fix $\epsilon_i = 0.2$ and study the effect of the degree of risk aversion on the game in Figure 1. Without risk aversion, the game has one symmetric and two free-riding RQE. As the degree of risk aversion increases, players become increasingly collaborative and the degree of free-riding diminishes. Importantly, there is a threshold where the two free-riding RQE disappear and "merge" into the symmetric one.

We validate this theorem in both the Overcooked Gridworld and the Tag environment. In both settings, Nash agents tend to learn free-riding behaviors, which severely impair generalization to new partners. By contrast, our strategically risk-averse agents contribute more actively to the task and achieve better partner generalization. Ablation studies further reveal a threshold-like phenomenon under which free-riding disappears as risk aversion increases past a certain threshold. This validates that the implications of Theorem 4.4 hold beyond the simple classes of games covered by the theorem.

5 MARL Algorithm Design

The theoretical benefits of strategic risk aversion—robustness to other players' behaviors, increased collaboration and no free-riding—prompt us to design scalable *strategically risk-averse* training algorithms for MARL. We do so in the general setting of policy optimization algorithms. For a more detailed mathematical background of risk aversion in MARL, we refer to Appendix D.

Meta-algorithm for strategically risk-averse policy optimization. In standard policy optimization, each agent seeks a policy π_{θ_i}, parametrized by parameters θ_i, that maximizes an objective function $\mathcal{L}_i(\theta_i, \theta_{-i})$, which is a function of the policies of all agents. Inspired by our strategically risk-averse utility (1), it is tempting to replace this objective function with

$$\inf_{\phi_i} \mathcal{L}_i(\theta_i, \phi_i) + \frac{1}{\tau_i} \mathrm{KL}(\phi_i, \theta_{-i}) - \epsilon_i H(\theta_i), \tag{5}$$

where, to ease exposition, we slightly overload the notation of $\mathrm{KL}(\phi_i, \theta_{-i})$ to be the (expected) KL-divergence between the induced policies across states (if the problem is one of MARL) . However, directly optimizing this objective is computationally prohibitive, since both evaluation and gradient computation require solving a high-dimensional non-convex inner optimization over ϕ_i.

To resolve this complexity, we follow Mazumdar et al. [38] and resort to an auxiliary game in which we augment the number of players. For each agent i, we bypass the direct computation of the infimum in (5) by introducing a *fictitious adversary* that directly interacts with this agent during training. This adversary aims to minimize agent i's performance while remaining close to the evolving policies of the other agents, thereby evaluating the infimum in (5). As such, the objective of agent i is now a function of their parameters θ_i and those of their adversary:

$$\mathcal{L}_i(\theta_i, \phi_i) + \frac{1}{\tau_i} \mathrm{KL}(\phi_i, \theta_{-i}) - \epsilon_i H(\theta_i), \tag{6}$$

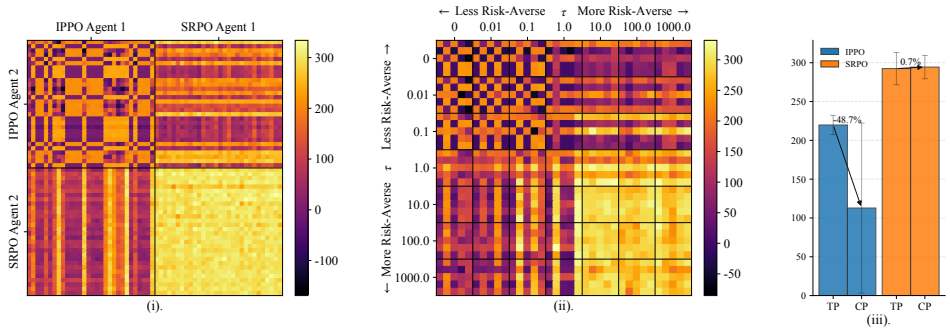


Figure 2: Overcooked cross-play and ablations. Each square is the average return over 10 episodes of length 128 for each pair of agents. Diagonal blocks represent the training performance of the agents. (i) We directly observe that IPPO ($\epsilon = 0.1$) learns to free-ride while SRPO ($\tau = 10, \epsilon = 0.1$) does not, and SRPO yields higher utility (cf. Theorem 4.1). (ii) Ablation over τ ($\epsilon = 0.1$). We empirically observe that free-riding vanishes as τ increases, matching Theorem 4.4. (iii) Difference between Training Performance (TP) and Cross-play Performance (CP) (mean and standard deviation): the performance of IPPO drastically decreases, with lower average and larger standard deviation in cross-play, while the performance of SRPO is unaffected.

where the KL term can now be dropped as it does not affect the agent's parameters. The objective of the fictitious adversary agent, who controls the adversarial policy π_{ϕ_i} and whose goal is to minimize the objective of agent i, depends on the parameters of all agents and is therefore given by

$$-\mathcal{L}_i(\theta_i, \phi_i) - \frac{1}{\tau_i} \mathrm{KL}(\phi_i, \theta_{-i}) + \epsilon_i H(\theta_i), \tag{7}$$

where the entropy term can be dropped. While adversarial training has been proposed as a robustness technique, such methods are often unstable to train and result in over-conservative policies [4, 30]. The key insight resulting from strategic risk aversion is the idea to constrain the adversary to not deviate from the evolving opponent policies. This extra step stabilizes training as the adversary is constrained to a set of plausible deviations and not arbitrary attacks.

This framing yields a meta-algorithm for strategically risk-averse MARL. In Theorem C.4, we prove that gradient descent on this meta-algorithm converges to an RQE in matrix games, which provides theoretical motivation for instantiating it with practical policy optimization methods; see Section C for details. We then instantiate the framework with Proximal Policy Optimization (PPO [51]), also called independent PPO (IPPO) when deployed in multi-agent settings [11], by setting $\mathcal{L}_i(\theta_i, \phi_i)$ to the standard clipped surrogate objective. We refer to the resulting algorithm as *Strategically Risk-averse Policy Optimization* (SRPO). See Section D.2 for the pseudo-code and details. In Section 6, we show that policy sharing can be incorporated to improve scalability.

6 Experiments

We evaluate the proposed SRPO algorithm against IPPO, the dominant scalable algorithm for collaborative MARL [71, 17, 11], on four environments: (1) a modified grid-world inspired by Overcooked AI [6], with both shared team rewards and private costs; (2) Tag [36], a control coordination task; (3) the multi-robot warehouse (RWARE) environment [43], and (4) an LLM-based multi-agent debate setting [13, 44] on GSM8K [7]. We focus on IPPO because PPO-based methods have been shown to outperform several population-based alternatives while remaining scalable, and this also allows us to isolate the effect of adding strategic risk aversion to the same policy optimization method. Nevertheless, to ensure comparison with a representative partner-generalization baseline, we also implement FCP [57] in Overcooked Gridworld; results in Section E.2 show that SRPO substantially outperforms FCP. We further provide additional results on Hanabi [2], a coordination card game, in Section E.5. Our experimental design closely follows the theoretical motivations developed earlier and evaluates the following claims:

- **Partner generalization.** SRPO achieves higher and more stable cross-play performance than IPPO when paired with previously unseen partners, and even varying number of partners.

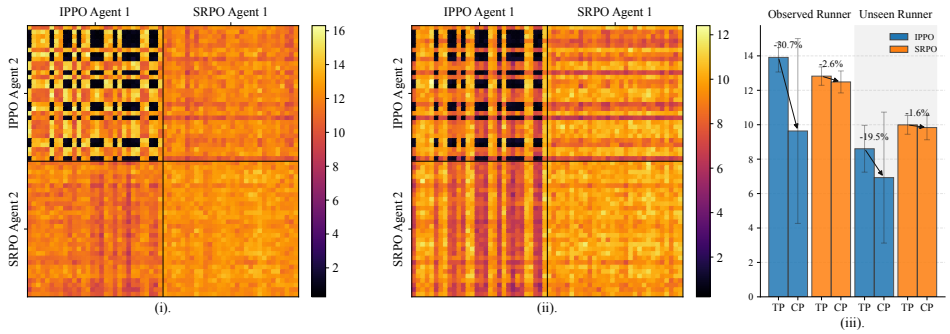


Figure 3: Tag cross-play with SRPO ($\tau = 10, \epsilon = 0.01$) and IPPO ($\epsilon = 0.01$) against a runner seen during training (i) and an unseen runner (ii). Each square averages over 100 episodes. IPPO does well in training (yet still learns free-riding like policies), but degrades drastically against an unseen runner. SRPO trades a small amount of training performance for substantially smaller TP-CP gap (iii).

- **Shared reward and free-riding.** In cooperative tasks with private costs, SRPO converges to equilibria with higher shared reward and reduced free-riding behavior.
- **Scalability.** SRPO remains practical in complex multi-agent settings when implemented with policy sharing and adversary sampling.

Training and evaluation. Across all environments, the number of interactions with the environment for both SRPO and IPPO are kept the same. We evaluate partner generalization using **cross-play performance** (CP), where a trained agent is paired with *held-out* partners trained from different seeds and not encountered during training, corresponding to off-diagonal entries in the evaluation matrix. We also report training performance (TP), where the agent is paired with itself, corresponding to diagonal entries. This evaluation makes free-riding policies especially visible: they produce a distinctive *checkerboard* pattern when free-riding agents are paired with each other and fail to achieve high performance. We consistently observe this phenomenon among IPPO-trained agents across environments.

Ablation studies. Across environments, we also report results of ablation studies on both entropy and strategic risk aversion to further validate our theory, show how our results are robust to hyper-parameter selection, and highlight how entropy alone cannot guarantee generalization in general collaborative games despite positive empirical evidence in special cases [17]. The code and visualizations of the trained policies are available here and the detailed setup is in Section E.

6.1 Overcooked Gridworld

We construct an environment based on the Overcooked [6] multi-agent benchmark to serve as a simple laboratory to verify our theory. In our Overcooked Gridworld, agents receive a shared *team reward* when an onion is picked up ($+1$) and placed into a pot ($+10$) and incur *private costs* for movement (-0.2) and collision (-2). This induces a canonical social dilemma: although the team objective is collaborative, each agent has an incentive to avoid costly effort and rely on its teammate to complete the task. Results are shown in Figure 2.

SRPO reduces free-riding and exhibits higher performance. Across all 30 seeds, IPPO converges to a free-riding equilibrium where one teammate avoids movement but collects high reward due to the partner's effort, as reflected by the checkerboard pattern in the top-left block of Figure 2(i). In contrast, SRPO agents learn to coordinate and *both* contribute to the task. This matches SRPO's objective: because training explicitly optimizes performance against *plausible adversarial partner deviations* (Section 5), non-contribution becomes risky. As summarized in Figure 2(iii), SRPO attains both higher performances and improved cross-play reliability, consistent with our theoretical predictions. Ablations in Figure 2(ii) show that IPPO ($\tau = 0$) and small values of τ lead to free-riding and thus poor cross-play performance, reflecting fragile coordination conventions and lack of effort on the part of agents. As τ increases, collaboration becomes more stable: policies converge to a consistent solution, free-riding disappears, and cross-play performance improves, validating that Theorem 4.4 holds more broadly. SRPO is also robust to choices of τ, with ablations showing consistent gains over IPPO across orders of magnitude of τ. Similar trends are observed in Tag and RWARE in Section E.7.

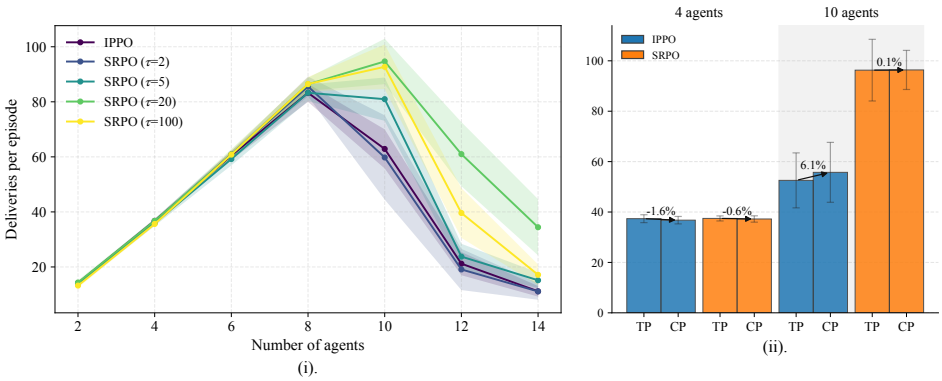


Figure 4: RWARE performances of SRPO and IPPO ($\epsilon = 0.05$). (i) Episodic delivery rate for both IPPO and SRPO, with 95% confidence intervals, as the number of agents varies; (ii) Difference between Training Performance (TP) and Cross-play Performance (CP) (mean and standard deviation): SRPO ($\tau = 20$) does not sacrifice performance in the training (4 agents TP), but achieves much better CP and TP performances in the ten-agent cases.

6.2 Tag

In Tag, two chasers must coordinate to catch a runner. The chasers get a shared positive reward (+1) if a chaser collides with the runner. We first train runner policies independently via adversarial training, and then train chaser policies using either IPPO or SRPO against a fixed runner to keep the game collaborative. Cross-play performance is evaluated along two axes: (i) pairing trained chasers with an unseen chaser partner (teammate shift), and (ii) evaluating chasers against a runner policy not observed during training (opponent shift). The results are shown in Figure 3.

SRPO learns more generalizable policies. In Tag, IPPO-trained agents can achieve strong training performance, but their policies often encode free-riding and overfit to both the runner encountered during training and specific coordination conventions it develops with its training partner. This yields strong in-distribution performance, but leads to sharp degradation under partner or opponent shift (Figure 3(iii)). In contrast, SRPO sacrifices a small amount of training performance—unsurprising here since Tag lacks the aggregative structure required for Theorem 4.1—but attains consistently higher and more stable cross-play performance against different partners (including free-riding partners) and runners, showing their potential to generalize more widely.

6.3 Multi-robot Warehouse

In the multi-robot warehouse (RWARE) environment, robots must coordinate to deliver requested goods without collision. We train SRPO and IPPO independently in `rware-tiny-4ag-easy-v2` setting with four robots, and evaluate them with varying number of partners.

Scalability of SRPO. To scale SRPO to larger agent populations, we use policy sharing: all agents share a single risk-averse policy, and training maintains only one adversary policy, with the adversarial role assigned to a uniformly-random agent at each rollout. For cross-play evaluation, half of the agents use one policy and the remaining half use the other. As shown in Figure 4(i), IPPO performance drops quickly as the number of agents increases due to more frequent robot collisions. In contrast, SRPO can achieve higher delivery rates through more robust coordination. Moreover, Figure 4(ii) shows that SRPO does not sacrifice training performance in the four-agent setting, while achieving substantially stronger training (all agents share the same policy) and cross-play performance in the ten-agent setting.

6.4 Multi-LLM-Agent Debate on GSM8K

To conclude, in a language-based cooperative reasoning setting, we train SRPO and IPPO agents to solve GSM8K math problems [7] through structured debate, where each agent must reinforce correct reasoning while remaining robust to potentially misleading or incorrect proposals from its teammate. We use Qwen2.5-{**0.5B**, **3B**}-Instruct [68] and Qwen3-{**0.6B**, **4B**-Instruct-2507} [69] as base models,

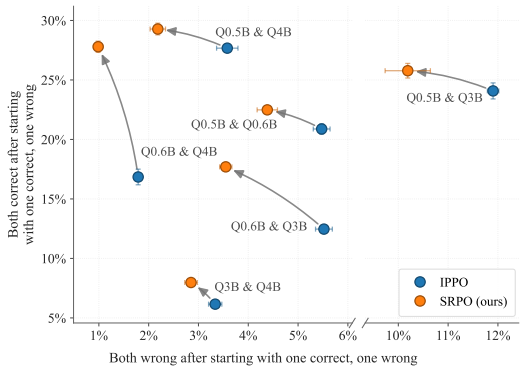


Figure 5: **How strategic risk aversion helps debate.** For each method and cross-play pair, we report the fraction of all evaluated debates in which exactly one agent is correct after the first round and both agents are correct at the end **(y-axis)** or both are wrong at the end **(x-axis)**. SRPO consistently increases the fraction ending with both agents correct and decreases the fraction ending with both agents wrong relative to IPPO.

Table 1: GSM8K debate robustness. Panel (a) reports joint accuracy for pairs of Qwen models. Panel (b) reports the Qwen agent's individual accuracy with an untuned Llama 3.2-1B-Instruct partner. Values are means with 95% confidence intervals over five evaluation rollouts. Gains are in percentage points. We set $\epsilon = 0$ for both methods and $\tau = 10$ for SRPO.

(a) Qwen cross-play joint accuracy

Qwen pair	IPPO	SRPO	Gain (pp)
0.5B / 0.6B	0.603 ± 0.004	0.621 ± 0.001	$+\mathbf{1.71} \pm 0.44$
0.5B / 3B	0.649 ± 0.005	0.684 ± 0.008	$+\mathbf{3.49} \pm 0.69$
0.5B / 4B	0.845 ± 0.005	0.872 ± 0.002	$+\mathbf{2.67} \pm 0.50$
0.6B / 3B	0.572 ± 0.003	0.615 ± 0.002	$+\mathbf{4.25} \pm 0.41$
0.6B / 4B	0.710 ± 0.007	0.848 ± 0.001	$+\mathbf{13.81} \pm 0.62$
3B / 4B	0.706 ± 0.005	0.732 ± 0.004	$+\mathbf{2.67} \pm 0.35$

(b) Accuracy with untuned Llama

Qwen	IPPO	SRPO	Gain (pp)
0.5B	0.379 ± 0.003	0.405 ± 0.004	$+\mathbf{2.56} \pm 0.48$
0.6B	0.588 ± 0.004	0.673 ± 0.009	$+\mathbf{8.49} \pm 0.58$
3B	0.552 ± 0.006	0.634 ± 0.007	$+\mathbf{8.26} \pm 0.95$
4B	0.900 ± 0.004	0.917 ± 0.004	$+\mathbf{1.65} \pm 0.14$

trained with verl [53]. See Section E.6 for full setup. We evaluate two robustness dimensions: (i) cross-play **joint accuracy** (problem correct only if *both* agents output the correct final answer) when agents with different model scales are paired, and (ii) **individual accuracy** when paired with a trained Qwen agent with an *untuned* Llama 3.2-1B-Instruct model [22], isolating robustness to drastic partner shift. Results are summarized in Table 1.

SRPO extends to LLM-based multi-agent systems. Table 1 shows that SRPO consistently improves cross-play joint accuracy, with gains up to 13.81 percentage points, as well as individual accuracy when paired with an untuned partner, with gains up to 8.49 percentage points. Across all evaluated pairings, Figure 5 further shows that SRPO increases the fraction of debates that start with one correct and one incorrect answer and end with both agents correct, and decreases the fraction that start in the same state and end with both agents wrong, relative to IPPO. This pattern suggests that SRPO agents are harder to mislead and more effective at persuading an incorrect partner to adopt the correct answer. These behaviors illustrate how strategic risk aversion can improve robustness to partner behavior during debate. Overall, SRPO promotes more robust and generalizable cooperative reasoning in multi-agent language environments.

7 Conclusion

We studied strategic risk aversion as an inductive bias for training collaborative agents. In structured games, we showed that it can increase shared rewards and reduce free-riding at equilibrium. These results motivated SRPO, which improves collaboration with unseen partners across several benchmarks, including multi-agent LLM debate. More broadly, this paper opens the door for integrating more aspects of human decision-making from behavioral economics, such as mutual trust, altruism, and fairness, into multi-agent reinforcement learning and AI training to improve the general awareness and behavior of autonomous agents.

Acknowledgments and Disclosure of Funding

Nicolas Lanzetti was supported by the PIMCO Fellows Program, by the National Science Foundation under Grant #CCF-2326609, and by the Resnick Sustainability Institute. Eric Mazumdar acknowledges support from NSF Award 2240110.

References

[1] Z. Ahmed, N. Le Roux, M. Norouzi, and D. Schuurmans. Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning*, pages 151–160. PMLR, 2019.

[2] N. Bard, J. N. Foerster, S. Chandar, N. Burch, M. Lanctot, H. F. Song, E. Parisotto, V. Dumoulin, S. Moitra, E. Hughes, I. Dunning, S. Mourad, H. Larochelle, M. G. Bellemare, and M. Bowling. The Hanabi challenge: A new frontier for AI research. *Artificial Intelligence*, 280(C), Mar. 2020.

[3] S. Barrett, N. Agmon, N. Hazon, S. Kraus, and P. Stone. Communicating with unknown teammates. In *International Conference on Autonomous Agents and Multi-Agent Systems*, page 1433–1434. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[4] A. Bukharin, Y. Li, Y. Yu, Q. Zhang, Z. Chen, S. Zuo, C. Zhang, S. Zhang, and T. Zhao. Robust multi-agent reinforcement learning via adversarial regularization: Theoretical foundation and stable algorithms. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2023.

[5] Y. Cao, W. Yu, W. Ren, and G. Chen. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial Informatics*, 9(1):427–438, 2013.

[6] M. Carroll, R. Shah, M. K. Ho, T. Griffiths, S. Seshia, P. Abbeel, and A. Dragan. On the utility of learning about humans for human-ai coordination. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[7] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[8] L. C. Corchón. Comparative statics for aggregative games the strong concavity case. *Mathematical Social Sciences*, 28(3):151–165, 1994.

[9] B. Cui, A. Lupu, S. Sokota, H. Hu, D. J. Wu, and J. N. Foerster. Adversarial diversity in hanabi. In *The Eleventh International Conference on Learning Representations*, 2023.

[10] A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel. Cooperative AI: Machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021.

[11] C. S. de Witt, T. Gupta, D. Makoviichuk, V. Makoviychuk, P. H. S. Torr, M. Sun, and S. Whiteson. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.

[12] T. Dizdarević, R. Hammond, T. Gessler, A. Calinescu, J. Cook, M. Gallici, A. Lupu, and J. N. Foerster. Ad-hoc human-AI coordination challenge. In *International Conference on Machine Learning*, volume 267, pages 13900–13937. PMLR, 2025.

[13] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*. PMLR, 2024.

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

[15] A. Feik, N. Lanzetti, S. Bolognani, F. Dörfler, and D. Paccagnan. Strategically robust aggregative games. In *2026 European Control Conference (ECC)*, pages 2014–2019, 2026.

[16] H. Föllmer and A. Schied. Convex measures of risk and trading constraints. *Finance and stochastics*, 6:429–447, 2002.

[17] J. Forkel, C. Ruhdorfer, A. Bulling, and J. Foerster. High entropy leads to symmetry equivariant policies in dec-pomdps. *arXiv preprint arXiv:2511.22581*, 2026.

[18] S. Ganesh, N. Vadori, M. Xu, H. Zheng, P. Reddy, and M. Veloso. Reinforcement learning for market making in a multi-agent dealer market. *arXiv preprint arXiv:1911.05892*, 2019.

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

[20] J. K. Goeree, C. A. Holt, and T. R. Palfrey. Risk averse behavior in generalized matching pennies games. *Games and Economic Behavior*, 45(1):97–113, 2003.

[21] C. Gollier. *The economics of risk and time*. MIT press, 2001.

[22] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[23] H. Hu, A. Lerer, A. Peysakhovich, and J. Foerster. "other-play" for zero-shot coordination. In *International Conference on Machine Learning*, pages 4399–4410. PMLR, 2020.

[24] H. Hu, A. Lerer, B. Cui, L. Pineda, N. Brown, and J. Foerster. Off-belief learning. In *International Conference on Machine Learning*, pages 4369–4379. PMLR, 2021.

[25] T. Kruse, A. K. Pandey, R. Alami, and A. Kirsch. Human-aware robot navigation: A survey. *Robot. Auton. Syst.*, 61(12):1726–1743, Dec. 2013.

[26] G. Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1):1059–1106, 2023.

[27] N. Lanzetti, A. Terpin, and F. Dörfler. Variational analysis in the Wasserstein space. *arXiv preprint arXiv:2406.10676*, 2024.

[28] N. Lanzetti, S. Bolognani, and F. Dörfler. First-order conditions for optimization in the Wasserstein space. *SIAM Journal on Mathematics of Data Science*, 7(1):274–300, 2025.

[29] N. Lanzetti, S. Fricker, S. Bolognani, F. Dörfler, and D. Paccagnan. Strategically robust game theory via optimal transport. *arXiv preprint arXiv:2507.15325*, 2025.

[30] N. Lauffer, A. Shah, M. Carroll, S. A. Seshia, S. Russell, and M. D. Dennis. Robust and diverse multi-agent learning via rational policy gradient. In *Advances in Neural Information Processing Systems*, 2025.

[31] J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Conference on Autonomous Agents and MultiAgent Systems*, page 464–473, Richland, SC, 2017. International Foundation for Autonomous Agents and Multiagent Systems.

[32] A. Lerer and A. Peysakhovich. Learning existing social conventions via observationally augmented self-play. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 107–114, New York, NY, USA, 2019. Association for Computing Machinery.

[33] Y. Liang, D. Chen, A. Gupta, S. S. Du, and N. Jaques. Learning to cooperate with humans using generative agents. In *Advances in Neural Information Processing Systems*, 2024.

[34] Y. Liang, Y. Qu, J. Yang, S. Lin, and Z.-J. M. Shen. Everyone contributes! Incentivizing strategic cooperation in multi-LLM systems via sequential public goods games. *arXiv preprint arXiv:2508.02076*, 2025.

[35] B. Liu, Z. Pu, Y. Pan, J. Yi, Y. Liang, and D. Zhang. Lazy agents: A new perspective on solving sparse reward problem in multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 21937–21950. PMLR, 2023.

[36] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, NIPS'17, page 6382–6393, Red Hook, NY, USA, 2017. Curran Associates Inc.

[37] A. Lupu, B. Cui, H. Hu, and J. Foerster. Trajectory diversity for zero-shot coordination. In *International Conference on Machine Learning*, pages 7204–7213. PMLR, 2021.

[38] E. Mazumdar, K. Panaganti, and L. Shi. Tractable multi-agent reinforcement learning through behavioral economics. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

[40] Meta Fundamental AI Research Diplomacy Team (FAIR)†, A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624): 1067–1074, 2022.

[41] D. Muglich, C. Schroeder de Witt, E. van der Pol, S. Whiteson, and J. Foerster. Equivariant networks for zero-shot coordination. In *Advances in Neural Information Processing Systems*, volume 35, pages 6410–6423. Curran Associates, Inc., 2022.

[42] D. Paccagnan, B. Gentile, F. Parise, M. Kamgarpour, and J. Lygeros. Nash and Wardrop equilibria in aggregative games with coupling constraints. *IEEE Transactions on Automatic Control*, 64(4):1373–1388, 2018.

[43] G. Papoudakis, F. Christianos, L. Schäfer, and S. V. Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS)*, 2021.

[44] C. Park, S. Han, X. Guo, A. E. Ozdaglar, K. Zhang, and J.-K. Kim. MAPoRL: Multi-agent post-co-training for collaborative large language models with reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30215–30248. Association for Computational Linguistics, 2025.

[45] V. L. Qin, N. Lanzetti, S. Bolognani, and H. Balakrishnan. Strategically robust game-theoretic multi-agent trajectory optimization. *arXiv preprint arXiv:2609.35142*, 2026.

[46] W. Qiu, X. Wang, R. Yu, R. Wang, X. He, B. An, S. Obraztsova, and Z. Rabinovich. RMIX: Learning risk-sensitive policies for cooperative reinforcement learning agents. In *Advances in Neural Information Processing Systems*, volume 34, pages 23049–23062. Curran Associates, Inc., 2021.

[47] A. Rahman, E. Fosong, I. Carlucho, and S. V. Albrecht. Generating teammates for training robust ad hoc teamwork agents via best-response diversity. *Transactions on Machine Learning Research*, 2023.

[48] J. B. Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, pages 520–534, 1965.

[49] C. Ruhdorfer, M. Bortoletto, A. Penzkofer, and A. Bulling. The overcooked generalisation challenge: Evaluating cooperation with novel partners in unknown environments using unsupervised environment design. *Transactions on Machine Learning Research*, 2025.

[50] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897. PMLR, 2015.

[51] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[52] S. Shen, C. Ma, C. Li, W. Liu, Y. Fu, S. Mei, X. Liu, and C. Wang. Riskq: risk-sensitive multi-agent reinforcement learning value factorization. In *Advances in Neural Information Processing Systems*, volume 36, pages 34791–34825. Curran Associates, Inc., 2023.

[53] G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin, and C. Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, page 1279–1297, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400711961. doi: 10.1145/3689031.3696075. URL https://doi.org/10.1145/3689031.3696075.

[54] L. Shi, E. Mazumdar, Y. Chi, and A. Wierman. Sample-efficient robust multi-agent reinforcement learning in the face of environmental uncertainty. In *International Conference on Machine Learning*. PMLR, 2024.

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

[56] P. Stone, G. Kaminka, S. Kraus, and J. Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proceedings of the AAAI Conference On Artificial Intelligence*, volume 24, pages 1504–1509, 2010.

[57] D. Strouse, K. R. McKee, M. Botvinick, E. Hughes, and R. Everett. Collaborating with humans without human data. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

[58] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, and T. Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, page 2085–2087. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

[59] J. Terry, B. Black, N. Grammel, M. Jayakumar, A. Hari, R. Sullivan, L. S. Santos, C. Dieffendahl, C. Horsch, R. Perez-Vicente, et al. Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:15032–15043, 2021.

[60] K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O'Sullivan, and H. D. Nguyen. Multi-agent collaboration mechanisms: A survey of LLMs. *arXiv preprint arXiv:2501.06322*, 2025.

[61] J. Treutlein, M. Dennis, C. Oesterheld, and J. Foerster. A new formalism, method and open issues for zero-shot coordination. In *International Conference on Machine Learning*, pages 10413–10423. PMLR, 2021.

[62] J. Tsitsiklis. *Problems in Decentralized Decision Making and Computation*. Massachusetts Institute of Technology, 1984.

[63] B. Velasevic, N. Lanzetti, and E. Mazumdar. Strategically robust linear quadratic dynamic games. *arXiv preprint arXiv:2604.22318*, 2026.

[64] V. Villin, T. K. Buening, and C. Dimitrakakis. A minimax approach to ad hoc teamwork. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '25, page 2105–2114, Richland, SC, 2025. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400714269.

[65] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[66] X. Wang, S. Zhang, W. Zhang, W. Dong, J. Chen, Y. Wen, and W. Zhang. ZSC-eval: An evaluation toolkit and benchmark for multi-agent zero-shot coordination. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[67] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

[68] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[69] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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

[71] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, volume 35, pages 24611–24624. Curran Associates, Inc., 2022.

[72] C. Yu, J. Gao, W. Liu, B. Xu, H. Tang, J. Yang, Y. Wang, and Y. Wu. Learning zero-shot cooperation with humans, assuming humans are biased. In *International Conference on Learning Representations*, 2023.

[73] K. Zhang, T. SUN, Y. Tao, S. Genc, S. Mallya, and T. Basar. Robust multi-agent reinforcement learning with model uncertainty. In *Advances in Neural Information Processing Systems*, volume 33, pages 10571–10583, 2020.

[74] Y. Zhang and E. Mazumdar. Convergent Q-learning for infinite-horizon general-sum Markov games through behavioral economics. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pages 5899–5904. IEEE, 2025.

[75] Y. Zhang and E. Mazumdar. Provably convergent actor-critic for MARL through risk-aversion. In *Forty-third International Conference on Machine Learning*, 2026.

[76] Z. Zhang, X. Li, Y. Lin, H. Liu, R. Chandradevan, L. Wu, M. Lin, F. Wang, X. Tang, Q. He, and S. Wang. Unlocking the power of multi-agent LLM for reasoning: From lazy agents to deliberation. *arXiv preprint arXiv:2511.02303*, 2025.

[77] R. Zhao, J. Song, Y. Yuan, H. Hu, Y. Gao, Y. Wu, Z. Sun, and W. Yang. Maximum entropy population-based training for zero-shot human-AI coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6145–6153, 2023.

A Further Discussion on Related Work

In this section, we discuss a broader set of related work to better position our work in the landscape of collaborative multi-agent learning and risk-averse/robust MARL.

Collaborative MARL and Partner Generalization Research into solving collaborative decision-making problems using MARL has a long history [5], originating from a desire for solving decentralized decision-making problems [62]. Far from fading, these issues remain at the fore of research interest with the advent of powerful AI systems and a desire to build large teams of AI agents [60] for solving collaborative tasks. Emerging from this literature, is a general consensus that classic single-agent policy optimization algorithms like PPO [51], used in a decentralized way, can yield state-of-the-art or near state-of-the-art performance without needing to be overly specialized to the multi-agent nature of the problem [71, 11, 17]. Despite such results, learned policies resulting from such approaches can fail to generalize to unseen agents [17] and new environments. We validate this through our experiments and further observe that these state-of-the-art algorithms are consistently prone to learning free-riding strategies.

Beyond the lines of work mentioned in Section 2, other important lines of work on addressing the partner generalization problem seek to avoid brittle coordination conventions by learning symmetry-invariant or partner-agnostic policies [23, 61, 41]. These often require strong structural assumptions and may not scale to complex collaborative environments. Another set of approaches attempt to achieve robustness by training against partner policies derived from human data [6, 40, 33], but their effectiveness is constrained by data availability and coverage. Neither of these approaches are fully agnostic to the structure of the underlying problem and are thus not general principles that can be broadly applied across different tasks.

Consequently, it remains unclear how to systematically achieve partner generalization without relying on human data or ad-hoc constructions. In this paper we showed that strategic risk aversion gives us one potential approach. Furthermore, we show that it can readily be combined with policy optimization algorithms like PPO, resulting in scalable and performant algorithms for learning generalizable policies.

Robustness and Risk Aversion in MARL A second related line of work is the emerging literature on robust and risk-averse MARL [73, 54, 70, 55, 46]. While most of the papers study robustness or risk aversion to changes in the underlying environment [14, 18, 46, 52, 54, 73] or to tail risks in large populations of agents [70], in this work we study *strategic* risk aversion. Strategic risk aversion and robustness yield solutions that are more amenable to computation than classic game-theoretic solution concepts [38], ensure robustness to unseen behaviors of the other agents, and sometimes even induce a coordination effect between the agents [29]—an empirical observation also seen in [15, 63] that we make rigorous in Section 4. Despite the prior work in this area, the broader benefits of the concept in collaborative games are unknown. This is the problem that we study in this paper.

B Proofs for Section 4

In this section, we provide the proofs of our two main theorems, Theorem 4.1 and Theorem 4.4, as well as the derivation of Example 4.2 used in Figure 1.

B.1 Proof of Theorem 4.1

We begin with the proof of Theorem 4.1. We first show that there exists an RQE in the aggregative games studied in this section. We then provide the proof of a more general version of Theorem 4.1.

B.1.1 Preliminaries

Before proving Proposition B.3, we study RQE for quadratic games. We start with the risk-averse quantal best response. We conduct our analysis in slightly more general settings where the players' utilities are

$$u_i(a_i, a_{-i}) = \frac{1}{2}\left\langle \begin{bmatrix} a_i \\ a_{-i} \end{bmatrix}, \begin{bmatrix} H_{i,i} & H_{i,-i} \\ H_{-i,i} & H_{-i,-i} \end{bmatrix} \begin{bmatrix} a_i \\ a_{-i} \end{bmatrix} \right\rangle + \left\langle \begin{bmatrix} h_i \\ h_{-i} \end{bmatrix}, \begin{bmatrix} a_i \\ a_{-i} \end{bmatrix} \right\rangle, \tag{8}$$

where the matrix

$$\begin{bmatrix} H_{i,i} & H_{i,-i} \\ H_{-i,i} & H_{-i,-i} \end{bmatrix}$$

is assumed to be symmetric negative semidefinite, and $H_{i,i}$ is assumed to be symmetric negative definite for all i. Also note that, for now, we do not assume that $h \neq 0$. Throughout this section, with slight abuse of notation, we denote by x_{-i} the product measure of the mixed strategies of all players except i. Moreover, in this quadratic setting, we set $U_i^{\tau_i,\epsilon_i}(x_i, x_{-i}) \coloneqq -\infty$ whenever $x_i \notin \mathcal{P}_2(\mathbb{R}^n)$, i.e., whenever x_i has infinite second moment. This is consistent with the fact that, since $H_{i,i} \prec 0$, the expected utility $\mathbb{E}_{a_i \sim x_i}[u_i(a_i, a_{-i})]$ equals $-\infty$ for all a_{-i} for such x_i, and it avoids undefined expressions in the entropy term. Similarly, we restrict the adversary in (1) to $p \in \mathcal{P}_2(\mathbb{R}^{n(N-1)})$, so that all expectations are well defined; as the proof of Lemma B.1 shows, this does not change the value of the infimum.

Lemma B.1 (risk-averse quantal best response)**.** *Let x_{-i} be a probability measure on $\mathbb{R}^{n(N-1)}$ with finite second moment, and let*

$$\varphi_i(m) \coloneqq \frac{1}{\tau_i} \log \mathbb{E}_{a_{-i} \sim x_{-i}} \left[\exp\left(-\tau_i q_i(m, a_{-i})\right) \right] \in (-\infty, +\infty],$$

where $q_i(m, a_{-i}) \coloneqq \frac{1}{2}\langle a_{-i}, H_{-i,-i} a_{-i} \rangle + \langle m, H_{i,-i} a_{-i} \rangle + \langle h_{-i}, a_{-i} \rangle$. If $\varphi_i(m) < +\infty$ for all $m \in \mathbb{R}^n$, then the risk-averse quantal best response, i.e., the maximizer of $U_i^{\tau_i,\epsilon_i}(\cdot, x_{-i})$ over $\Delta(\mathbb{R}^n)$, is uniquely given by the Gaussian mixed strategy with covariance matrix $\Sigma_i = -\epsilon_i H_{i,i}^{-1}$ and mean

$$m_i = \operatorname*{arg\,max}_{m \in \mathbb{R}^n} \ \frac{1}{2}\langle m, H_{i,i} m \rangle + \langle h_i, m \rangle - \varphi_i(m), \tag{9}$$

where the maximizer exists and is unique. In particular, let x_{-i} be Gaussian with mean m_{-i} and covariance matrix $\Sigma_{-i} \succ 0$, and let $P_{-i} \coloneqq \frac{1}{\tau_i}\Sigma_{-i}^{-1} + H_{-i,-i}$. If P_{-i} is positive definite, then $\varphi_i < +\infty$ on $\mathbb{R}^n$ and the best response has

$$\begin{aligned} \Sigma_i &= -\epsilon_i H_{i,i}^{-1} \\ m_i &= \left(-H_{i,i} + H_{i,-i} P_{-i}^{-1} H_{i,-i}^{\top}\right)^{-1} \left(h_i + H_{i,-i} P_{-i}^{-1} \left(\frac{1}{\tau_i} \Sigma_{-i}^{-1} m_{-i} - h_{-i} \right) \right). \end{aligned}$$

If instead P_{-i} is not positive definite, then $\varphi_i \equiv +\infty$ and the risk is infinite for every mixed strategy x_i.

Proof. Recall that $U_i^{\tau_i,\epsilon_i}(x_i, x_{-i}) = -\infty$ whenever $x_i \notin \mathcal{P}_2(\mathbb{R}^n)$, and let $x_i \in \mathcal{P}_2(\mathbb{R}^n)$ have mean $m \coloneqq \int_{\mathbb{R}^n} a_i \mathrm{d}x_i(a_i)$. With $V_i(a_i) \coloneqq -\frac{1}{2}\langle a_i, H_{i,i} a_i \rangle - \langle h_i, a_i \rangle$, we have $-u_i(a_i, a_{-i}) = V_i(a_i) - q_i(a_i, a_{-i})$, and, since q_i is affine in its first argument, $f(a_{-i}) \coloneqq \mathbb{E}_{a_i \sim x_i}[-u_i(a_i, a_{-i})] = \int_{\mathbb{R}^n} V_i \mathrm{d}x_i - q_i(m, a_{-i})$. By the duality representation of the entropic risk measure (e.g., see Föllmer and Schied [16]),

$$\begin{aligned} -U_i^{\tau_i,\epsilon_i}(x_i, x_{-i}) &= \sup_{p \in \mathcal{P}_2(\mathbb{R}^{n(N-1)})} \left(\mathbb{E}_{a_{-i} \sim p}[f(a_{-i})] - \frac{1}{\tau_i} \mathrm{KL}(p, x_{-i}) \right) + \epsilon_i H(x_i) \\ &= \frac{1}{\tau_i} \log \mathbb{E}_{a_{-i} \sim x_{-i}} \left[e^{\tau_i f(a_{-i})} \right] + \epsilon_i H(x_i) = \varphi_i(m) + \int_{\mathbb{R}^n} V_i \mathrm{d}x_i + \epsilon_i H(x_i) =: J_i(x_i). \end{aligned}$$

Föllmer and Schied [16] state the second equality for bounded f; it extends, with both sides possibly $+\infty$, to our f, which is bounded below by an affine function since $H_{-i,-i} \preceq 0$ (in particular, "$\leq$" follows from Jensen's inequality, and "$\geq$" from $p_k \propto e^{\tau_i \min\{f,k\}} x_{-i}$, which has finite second moment since its density with respect to x_{-i} is bounded, with $k \to \infty$). Moreover, $J_i(x_i) \in (-\infty, +\infty]$, since $H(x_i) > -\infty$ for $x_i \in \mathcal{P}_2(\mathbb{R}^n)$ and, by Jensen's inequality, $\varphi_i(m) \geq -\mathbb{E}_{a_{-i} \sim x_{-i}}[q_i(m, a_{-i})]$, which is an affine function of m, since x_{-i} has finite second moment.

Now let $\varphi_i < +\infty$ on $\mathbb{R}^n$. Then, φ_i is convex, by Hölder's inequality, hence continuous, and differentiable, with

$$\nabla \varphi_i(m) = -\frac{\mathbb{E}_{a_{-i} \sim x_{-i}} \left[H_{i,-i} a_{-i} \exp\left(-\tau_i q_i(m, a_{-i})\right) \right]}{\mathbb{E}_{a_{-i} \sim x_{-i}} \left[\exp\left(-\tau_i q_i(m, a_{-i})\right) \right]},$$

by differentiation under the integral sign, which is justified since φ_i is finite on all of $\mathbb{R}^n$. Since $H_{i,i} \prec 0$ and φ_i is convex and bounded below by an affine function, the function $\psi_i(m) \coloneqq \frac{1}{2}\langle m, H_{i,i}m\rangle + \langle h_i, m\rangle - \varphi_i(m)$ maximized in (9) is strictly concave and continuous, and $\psi_i(m) \to -\infty$ as $\|m\| \to \infty$. Hence, its maximizer m_i exists, is unique, and is the unique solution of $\nabla\psi_i(m_i) = H_{i,i}m_i + h_i - \nabla\varphi_i(m_i) = 0$. It remains to show that $x_i^\star \coloneqq \mathcal{N}(m_i, \Sigma_i)$ is the unique minimizer of J_i over $\mathcal{P}_2(\mathbb{R}^n)$; since $J_i(x_i^\star) < +\infty$, $x_i^\star$ is then the unique best response. We use the calculus in the Wasserstein space of Lanzetti et al. [27, 28], and denote by $\partial^- J_i(x_i)$ the Wasserstein subdifferential of J_i at x_i.

Step 1: Motivation of the candidate. If x_i has a smooth positive density ρ, then, by the computations in Step 2, the condition $0 \in \partial^- J_i(x_i)$ reads

$$\nabla\varphi_i(m) - H_{i,i}a_i - h_i + \epsilon_i\frac{\nabla\rho(a_i)}{\rho(a_i)} = 0 \qquad \text{for all } a_i \in \mathbb{R}^n. \tag{10}$$

Thus, $\nabla\log\rho(a_i) = \frac{1}{\epsilon_i}(H_{i,i}a_i + h_i - \nabla\varphi_i(m))$ is affine and, since $H_{i,i} \prec 0$, ρ is the density of a Gaussian distribution with covariance $\Sigma_i = -\epsilon_i H_{i,i}^{-1}$ and mean $-H_{i,i}^{-1}(h_i - \nabla\varphi_i(m))$, which coincides with m if and only if $\nabla\psi_i(m) = 0$, i.e., $m = m_i$. This motivates the candidate $x_i^\star$.

Step 2: Optimality and uniqueness. First, $0 \in \partial^- J_i(x_i^\star)$. Indeed, the first term of J_i is Wasserstein differentiable at $x_i^\star$ with constant gradient $\nabla\varphi_i(m_i)$, as in the chain rule [28, Proposition 2.16]: for any optimal coupling γ of $x_i^\star$ and ν, the difference of the means of ν and $x_i^\star$ is $\int(y-x)\mathrm{d}\gamma(x,y)$, whose norm is at most $W_2(x_i^\star, \nu)$, and φ_i is differentiable. The second term is Wasserstein differentiable with gradient ∇V_i [28, Proposition 2.20]. Since the density ρ of $x_i^\star$ is smooth and positive with $\nabla\log\rho(a_i) = -\Sigma_i^{-1}(a_i - m_i)$ square integrable with respect to $x_i^\star$, $\nabla\log\rho$ is a Wasserstein subgradient of the negative entropy at $x_i^\star$ [28, Example 2.27]. By the sum rule [28, Proposition 2.15], the left-hand side of (10) with $m = m_i$ is thus a Wasserstein subgradient of J_i at $x_i^\star$, and it vanishes for all $a_i \in \mathbb{R}^n$ since $\epsilon_i\Sigma_i^{-1} = -H_{i,i}$ and $\nabla\psi_i(m_i) = 0$. Second, J_i is lower semicontinuous on $\mathcal{P}_2(\mathbb{R}^n)$, since its first two terms are continuous (φ_i is continuous and V_i has quadratic growth) and the negative entropy is lower semicontinuous [28, Proposition 2.25]. Moreover, J_i is λ-geodesically convex with $\lambda \coloneqq \lambda_{\min}(-H_{i,i}) > 0$, since each of its terms is (λ-)convex along every geodesic: (i) the mean is affine along geodesics and φ_i is convex; (ii) the potential energy $\int_{\mathbb{R}^n} V_i\mathrm{d}x_i$ is λ-convex, since $\nabla^2 V_i = -H_{i,i} \succeq \lambda I$ [28, Proposition 2.20]; (iii) the negative entropy is displacement convex [28, Proposition 2.25]. Thus, by [28, Theorem 3.3], $x_i^\star$ is the unique minimizer of J_i over $\mathcal{P}_2(\mathbb{R}^n)$.

Step 3: Gaussian opponents. Let x_{-i} be Gaussian with mean m_{-i} and covariance $\Sigma_{-i} \succ 0$. Collecting the exponents, and since $\Sigma_{-i}^{-1} + \tau_i H_{-i,-i} = \tau_i P_{-i}$, the density of x_{-i} multiplied by $\exp(-\tau_i q_i(m,\cdot))$ is proportional to

$$\exp\left(-\frac{1}{2}\langle a_{-i}, \tau_i P_{-i}a_{-i}\rangle + \left\langle \Sigma_{-i}^{-1}m_{-i} - \tau_i\left(H_{i,-i}^\top m + h_{-i}\right), a_{-i}\right\rangle\right).$$

For every $m \in \mathbb{R}^n$, this function is integrable if and only if $P_{-i} \succ 0$. Hence, if P_{-i} is not positive definite, then $\varphi_i \equiv +\infty$, so that $U_i^{\tau_i,\epsilon_i}(x_i, x_{-i}) = -J_i(x_i) = -\infty$ for every $x_i \in \mathcal{P}_2(\mathbb{R}^n)$, and, by convention, for every other x_i; i.e., the risk is infinite. If instead $P_{-i} \succ 0$, then $\varphi_i < +\infty$ on $\mathbb{R}^n$, and this function is, up to normalization, the density of a Gaussian distribution with mean

$$\bar{m}_{-i}(m) \coloneqq \frac{1}{\tau_i}P_{-i}^{-1}\left(\Sigma_{-i}^{-1}m_{-i} - \tau_i\left(H_{i,-i}^\top m + h_{-i}\right)\right),$$

so that, by the expression for $\nabla\varphi_i$ above, $\nabla\varphi_i(m) = -H_{i,-i}\bar{m}_{-i}(m)$. Thus, the stationarity condition $H_{i,i}m_i + h_i - \nabla\varphi_i(m_i) = 0$ becomes the linear equation

$$\left(-H_{i,i} + H_{i,-i}P_{-i}^{-1}H_{i,-i}^\top\right)m_i = h_i + H_{i,-i}P_{-i}^{-1}\left(\frac{1}{\tau_i}\Sigma_{-i}^{-1}m_{-i} - h_{-i}\right),$$

whose matrix is positive definite, since $H_{i,i} \prec 0$ and $P_{-i} \succ 0$. Solving for m_i gives the expression in the statement. □

With this lemma, we obtain a way to compute RQE in continuous games:

Lemma B.2. *For each player i, let $\Sigma_i \coloneqq -\epsilon_i H_{i,i}^{-1} \succ 0$, let $\Sigma_{-i} \coloneqq \mathrm{blkdiag}(\Sigma_j)_{j\neq i}$, and let $P_{-i} \coloneqq \frac{1}{\tau_i}\Sigma_{-i}^{-1} + H_{-i,-i}$. Assume that $P_{-i} \succ 0$ for all i. Consider a surrogate "game of the means", where each player selects a vector $m_i \in \mathbb{R}^n$ to maximize the surrogate concave quadratic utility*

$$\begin{aligned}\bar{U}_i(m_i, m_{-i}) = \frac{1}{2}\langle m_i, (H_{i,i} - H_{i,-i}P_{-i}^{-1}H_{i,-i}^\top)m_i\rangle \\ + \frac{1}{\tau_i}\langle m_i, H_{i,-i}P_{-i}^{-1}\Sigma_{-i}^{-1}m_{-i}\rangle + \langle m_i, h_i - H_{i,-i}P_{-i}^{-1}h_{-i}\rangle,\end{aligned} \tag{11}$$

and, for $m = (m_1, \dots, m_N)$, let $x(m) \coloneqq (x_1(m_1), \dots, x_N(m_N))$ with $x_i(m_i) \coloneqq \mathcal{N}(m_i, \Sigma_i)$. Then:

- *(i) $m^\star$ is a Nash equilibrium of the surrogate game if and only if $x(m^\star)$ is an RQE of the game with utilities* (8)*;*
- *(ii) if $(x_1^\star, \dots, x_N^\star)$ is an RQE with $U_i^{\tau_i,\epsilon_i}(x_i^\star, x_{-i}^\star) > -\infty$ for all i, then $(x_1^\star, \dots, x_N^\star) = x(m^\star)$ for some Nash equilibrium $m^\star$ of the surrogate game.*

Proof. Fix a player i and $m_{-i} \in \mathbb{R}^{n(N-1)}$, and let $x_{-i}(m_{-i}) \coloneqq \bigotimes_{j\neq i} x_j(m_j)$, i.e., the Gaussian distribution with mean m_{-i} and covariance $\Sigma_{-i} \succ 0$. Since $P_{-i} \succ 0$, Lemma B.1 implies that $x_{-i}(m_{-i})$ satisfies the assumptions of its first part, and that the unique best response to $x_{-i}(m_{-i})$ is $\mathcal{N}(b_i(m_{-i}), \Sigma_i) = x_i(b_i(m_{-i}))$, where

$$b_i(m_{-i}) \coloneqq \left(-H_{i,i} + H_{i,-i}P_{-i}^{-1}H_{i,-i}^\top\right)^{-1}\left(h_i + H_{i,-i}P_{-i}^{-1}\left(\frac{1}{\tau_i}\Sigma_{-i}^{-1}m_{-i} - h_{-i}\right)\right).$$

On the other hand, since $H_{i,i} \prec 0$ and $P_{-i} \succ 0$, we have $H_{i,i} - H_{i,-i}P_{-i}^{-1}H_{i,-i}^\top \prec 0$. Thus, $\bar{U}_i(\cdot, m_{-i})$ is strictly concave, and its unique maximizer is the unique solution of

$$\left(H_{i,i} - H_{i,-i}P_{-i}^{-1}H_{i,-i}^\top\right)m_i + \frac{1}{\tau_i}H_{i,-i}P_{-i}^{-1}\Sigma_{-i}^{-1}m_{-i} + h_i - H_{i,-i}P_{-i}^{-1}h_{-i} = 0,$$

namely $m_i = b_i(m_{-i})$. Hence, for all i and m_{-i}, the unique best response to $x_{-i}(m_{-i})$ in the original game is $x_i(b_i(m_{-i}))$, where $b_i(m_{-i})$ is the unique best response to m_{-i} in the surrogate game.

Proof of (i). If $m^\star$ is a Nash equilibrium of the surrogate game, then $m_i^\star = b_i(m_{-i}^\star)$ for all i, so $x_i(m_i^\star)$ is the best response to $x_{-i}(m_{-i}^\star)$ for all i; i.e., $x(m^\star)$ is an RQE. Conversely, if $x(m^\star)$ is an RQE, then, for all i, $x_i(m_i^\star)$ is a best response to $x_{-i}(m_{-i}^\star)$, and, by uniqueness of the best response, $x_i(m_i^\star) = x_i(b_i(m_{-i}^\star))$, i.e., $m_i^\star = b_i(m_{-i}^\star)$. Thus, $m^\star$ is a Nash equilibrium of the surrogate game.

Proof of (ii). Let $(x_1^\star, \dots, x_N^\star)$ be an RQE with $U_j^{\tau_j,\epsilon_j}(x_j^\star, x_{-j}^\star) > -\infty$ for all j, and fix a player i. By the proof of Lemma B.1, each $x_j^\star$ has finite second moment, and, for every x_i with finite second moment and mean m,

$$U_i^{\tau_i,\epsilon_i}(x_i, x_{-i}^\star) = -\varphi_i(m) - \int_{\mathbb{R}^n} V_i(a_i)\mathrm{d}x_i(a_i) - \epsilon_i H(x_i),$$

with φ_i defined with $x_{-i} = x_{-i}^\star$. Let $m_i^\star$ be the mean of $x_i^\star$ and $G_i \coloneqq \mathcal{N}(m_i^\star, \Sigma_i)$. Since $U_i^{\tau_i,\epsilon_i}(x_i^\star, x_{-i}^\star) > -\infty$, we have $\varphi_i(m_i^\star) < +\infty$ and $H(x_i^\star) < +\infty$, so $U_i^{\tau_i,\epsilon_i}(G_i, x_{-i}^\star)$ is finite as well. Since $\Sigma_i^{-1} = -H_{i,i}/\epsilon_i$, the log-density of G_i is $-V_i/\epsilon_i$ plus an affine function of a_i; as $x_i^\star$ and G_i have the same mean, this yields $\epsilon_i \int_{\mathbb{R}^n} \log \frac{\mathrm{d}G_i}{\mathrm{d}a_i}\mathrm{d}(x_i^\star - G_i) = -\int_{\mathbb{R}^n} V_i \mathrm{d}(x_i^\star - G_i)$, and thus

$$U_i^{\tau_i,\epsilon_i}(G_i, x_{-i}^\star) - U_i^{\tau_i,\epsilon_i}(x_i^\star, x_{-i}^\star) = \int_{\mathbb{R}^n} V_i \mathrm{d}(x_i^\star - G_i) + \epsilon_i\left(H(x_i^\star) - H(G_i)\right) = \epsilon_i \,\mathrm{KL}(x_i^\star, G_i).$$

Since $x_i^\star$ is a best response to $x_{-i}^\star$, the left-hand side is nonpositive, so $\mathrm{KL}(x_i^\star, G_i) = 0$, i.e., $x_i^\star = G_i = x_i(m_i^\star)$. As i was arbitrary, $(x_1^\star, \dots, x_N^\star) = x(m^\star)$, and, by (i), $m^\star$ is a Nash equilibrium of the surrogate game. □

At this point, we can study RQE for games in which

$$R(a_1, \dots, a_N) = -\frac{1}{2}\left\|\sum_{i=1}^N a_i - \bar{a}\right\|_H^2 = -\frac{1}{2}\left\langle \sum_{i=1}^N a_i - \bar{a}, H\left(\sum_{i=1}^N a_i - \bar{a}\right)\right\rangle,\ c_i(a_i) = \frac{\rho_i}{2}\|a_i\|^2,$$

where we assume that the matrix H, used in the weighted norm, is symmetric positive definite and $\rho_i > 0$. As we will show below, considering this class is sufficient.

Proposition B.3 (computation of RQE)**.** *Let $\epsilon_i > 0$, let $\Sigma_i \coloneqq \epsilon_i(\rho_i I + H)^{-1} \succ 0$, and let $\tau_i > 0$ be such that*

$$\frac{1}{\tau_i}\Big(\sum_{j\neq i}\Sigma_j\Big)^{-1} - H \succ 0 \qquad \forall i \in [N], \tag{12}$$

or, equivalently, $\tau_i \sum_{j\neq i}\Sigma_j \prec H^{-1}$. Let $W_i \coloneqq \big(H^{-1} - \tau_i \sum_{j\neq i}\Sigma_j\big)^{-1} \succ 0$, and consider the surrogate "game of the means", where each player selects $m_i \in \mathbb{R}^n$ to maximize the strongly concave quadratic utility

$$\bar{U}_i(m_i, m_{-i}) = -\frac{1}{2}\left\langle m_i, (\rho_i I + W_i)\, m_i\right\rangle - \left\langle m_i, W_i\Big(\sum_{j\neq i} m_j - \bar{a}\Big)\right\rangle. \tag{13}$$

Then, the surrogate game has a unique Nash equilibrium $(m_1^\star, \dots, m_N^\star)$, given by

$$m_i^\star = \frac{1}{\rho_i} W_i\Big(I + \sum_{j=1}^{N}\frac{1}{\rho_j}W_j\Big)^{-1}\bar{a}, \qquad i \in [N], \tag{14}$$

and the Gaussian mixed strategies $x_i^\star \coloneqq \mathcal{N}(m_i^\star, \Sigma_i)$ form an RQE of the game at which all players have finite risk-adjusted utility. Moreover, $(x_1^\star, \dots, x_N^\star)$ is the unique RQE at which all players have finite risk-adjusted utility.

Proof. We derive the statement from Lemma B.2. Fix a player i, let $S_i \coloneqq \sum_{j\neq i}\Sigma_j$, and let $E \coloneqq \mathbf{1}_{N-1} \otimes I \in \mathbb{R}^{n(N-1)\times n}$, where $\mathbf{1}_{N-1} \in \mathbb{R}^{N-1}$ is the vector of all ones, so that $E^\top a_{-i} = \sum_{j\neq i} a_j$.

Step 1: Quadratic form of the utilities. Expanding the square in $u_i(a_i, a_{-i}) = -\frac{1}{2}\|a_i + E^\top a_{-i} - \bar{a}\|_H^2 - \frac{\rho_i}{2}\|a_i\|^2$ shows that, up to the additive constant $-\frac{1}{2}\|\bar{a}\|_H^2$, the utilities are of the form (8) with

$$H_{i,i} = -(\rho_i I + H), \quad H_{i,-i} = -HE^\top, \quad H_{-i,-i} = -EHE^\top, \quad h_i = H\bar{a}, \quad h_{-i} = EH\bar{a}.$$

In particular, every $n \times n$ block of $H_{-i,-i}$ equals $-H$, since the opponents are coupled through the aggregate $\sum_{j\neq i} a_j$. The additive constant shifts $U_i^{\tau_i,\epsilon_i}$ by the same constant and thus affects neither best responses nor the finiteness of the risk-adjusted utilities. Moreover, $H_{i,i} \prec 0$, the block matrix in (8) equals $-\left[\begin{smallmatrix} I \\ E \end{smallmatrix}\right] H \left[\begin{smallmatrix} I & E^\top \end{smallmatrix}\right] - \left[\begin{smallmatrix} \rho_i I & 0 \\ 0 & 0 \end{smallmatrix}\right] \preceq 0$, and $-\epsilon_i H_{i,i}^{-1} = \Sigma_i$.

Step 2: Positive definiteness of P_{-i}. Let Σ_{-i} and $P_{-i} = \frac{1}{\tau_i}\Sigma_{-i}^{-1} - EHE^\top$ be as in Lemma B.2, and note that $E^\top\Sigma_{-i}E = S_i$. With $K \coloneqq \Sigma_{-i}^{1/2}EH^{1/2}$, we have $P_{-i} = \frac{1}{\tau_i}\Sigma_{-i}^{-1/2}(I - \tau_i KK^\top)\Sigma_{-i}^{-1/2}$. Since $KK^\top$ and $K^\top K = H^{1/2}S_iH^{1/2}$ have the same nonzero eigenvalues, $P_{-i} \succ 0$ if and only if $\tau_i H^{1/2}S_iH^{1/2} \prec I$, i.e., $\tau_i S_i \prec H^{-1}$, which is (12).

Step 3: Identification of the surrogate game. By the Woodbury identity and since $E^\top\Sigma_{-i}E = S_i$ and $W_i^{-1} = H^{-1} - \tau_i S_i$, we have $P_{-i}^{-1} = \tau_i\Sigma_{-i} + \tau_i^2\Sigma_{-i}EW_iE^\top\Sigma_{-i}$, and thus

$$\frac{1}{\tau_i}E^\top P_{-i}^{-1}\Sigma_{-i}^{-1} = (I + \tau_i S_i W_i)E^\top = H^{-1}W_iE^\top.$$

Multiplying from the left by H and from the right by $\tau_i\Sigma_{-i}EH$ gives $HE^\top P_{-i}^{-1}EH = \tau_i W_i S_i H = W_i - H$. Therefore, the coefficients of the surrogate utility (11) are

$$\begin{aligned} H_{i,i} - H_{i,-i}P_{-i}^{-1}H_{i,-i}^\top &= -(\rho_i I + H) - (W_i - H) = -(\rho_i I + W_i), \\ \frac{1}{\tau_i}H_{i,-i}P_{-i}^{-1}\Sigma_{-i}^{-1}m_{-i} &= -W_iE^\top m_{-i} = -W_i\sum_{j\neq i} m_j, \\ h_i - H_{i,-i}P_{-i}^{-1}h_{-i} &= H\bar{a} + (W_i - H)\bar{a} = W_i\bar{a}, \end{aligned}$$

so that (11) coincides with (13).

Step 4: Closed form. Since $\rho_i I + W_i \succ 0$, each $\bar{U}_i(\cdot, m_{-i})$ is strongly concave, so $(m_1, \dots, m_N)$ is a Nash equilibrium of the surrogate game if and only if $(\rho_i I + W_i)m_i + W_i(\sum_{j\neq i} m_j - \bar{a}) = 0$ for all i, i.e.,

$$\rho_i m_i + W_i v = 0 \quad \forall i \in [N], \qquad \text{where } v \coloneqq \sum_{j=1}^{N} m_j - \bar{a}.$$

Summing $m_i = -\frac{1}{\rho_i} W_i v$ over i gives $(I + \sum_j \frac{1}{\rho_j} W_j)v = -\bar{a}$. Since $W_j \succ 0$ and $\rho_j > 0$, the matrix $I + \sum_j \frac{1}{\rho_j} W_j$ is positive definite and hence invertible, so v, and with it each $m_i = -\frac{1}{\rho_i} W_i v$, is uniquely determined, which yields (14). Conversely, (14) satisfies these conditions, since it gives $\sum_j m_j - \bar{a} = v$ for $v \coloneqq -(I + \sum_j \frac{1}{\rho_j} W_j)^{-1}\bar{a}$.

Step 5: Conclusion. By Steps 1 and 2, Lemma B.2 applies, and, by Step 3, its surrogate game is the one in (13), whose unique Nash equilibrium is (14) by Step 4. Hence, by Lemma B.2(i), $(x_1^\star, \dots, x_N^\star)$ is an RQE. At this RQE, all players have finite risk-adjusted utility: since $P_{-i} \succ 0$, Lemma B.1 gives $\varphi_i < +\infty$ on $\mathbb{R}^n$, and $x_i^\star$ is Gaussian, so that it has finite second moment and finite entropy. Finally, by Lemma B.2(ii), every RQE at which all players have finite risk-adjusted utility is of the form $(\mathcal{N}(m_1, \Sigma_1), \dots, \mathcal{N}(m_N, \Sigma_N))$ for some Nash equilibrium $(m_1, \dots, m_N)$ of the surrogate game, i.e., by Step 4, for $(m_1, \dots, m_N) = (m_1^\star, \dots, m_N^\star)$. Thus, $(x_1^\star, \dots, x_N^\star)$ is the unique such RQE. □

Discussion We make two observations on this result. First, the equilibrium is available in closed form for any number of players and for player-dependent parameters ρ_i, τ_i, and ϵ_i; moreover, the covariances Σ_i do not depend on the degrees of risk aversion. Second, uniqueness holds among all RQE at which all players have finite risk-adjusted utility. This qualification cannot be dropped: if all players play Gaussian strategies with sufficiently large covariances, then, by Lemma B.1, every player has infinite risk regardless of their own mixed strategy, so that any such profile is trivially a (degenerate) RQE.

B.1.2 Proof of Theorem 4.1

We prove a more general version of Theorem 4.1, valid for player-dependent degrees of risk aversion τ_i, bounded rationality ϵ_i, and effort costs ρ_i.

Theorem B.4 (risk induces collaboration)**.** *Consider the game of Section 4.1 and let $\epsilon_i > 0$ for all i. Let $\Sigma_i \coloneqq \epsilon_i(\rho_i I - H)^{-1} \succ 0$, and define the open box of admissible degrees of risk aversion*

$$\mathcal{T} \coloneqq (0, \bar{\tau}_1) \times \cdots \times (0, \bar{\tau}_N), \qquad \bar{\tau}_i \coloneqq \frac{1}{\lambda_{\max}\big(-H \sum_{j\neq i} \Sigma_j\big)}. \tag{15}$$

For $\tau = (\tau_1, \dots, \tau_N) \in \mathcal{T}$, let $(x_1^\star(\tau), \dots, x_N^\star(\tau))$ be the unique RQE of the game at which all players have finite risk-adjusted utility, which exists and is Gaussian by Proposition B.3. Then, the expected shared reward

$$J(\tau) \coloneqq \mathbb{E}_{a_i \sim x_i^\star(\tau)}\left[R(a_1, \dots, a_N)\right]$$

is increasing on $\mathcal{T}$: if $\tau, \tau' \in \mathcal{T}$ with $\tau_i' \geq \tau_i$ for all $i \in [N]$, then $J(\tau') \geq J(\tau)$, with strict inequality whenever $\tau' \neq \tau$. That is, players contribute more to the shared reward as they become more risk-averse.

Proof. **Step 1: Reduction.** Let $\bar{H} \coloneqq -H \succ 0$ and $\bar{a} \coloneqq -H^{-1}h$. Completing the square in the shared reward of Section 4.1 gives

$$R(a_1, \dots, a_N) = -\frac{1}{2}\Big\| \sum\nolimits_{i=1}^{N} a_i - \bar{a}\Big\|_{\bar{H}}^2 - \frac{1}{2}\langle h, H^{-1}h\rangle.$$

Thus, up to the additive constant $-\frac{1}{2}\langle h, H^{-1}h\rangle$, the utilities are those of the game studied in Proposition B.3, with $\bar{H}$ in place of H. This constant shifts every $U_i^{\tau_i,\epsilon_i}$ by the same constant and thus affects neither best responses nor the finiteness of the risk-adjusted utilities; hence, the two games have the same RQE at which all players have finite risk-adjusted utility. Moreover, $\Sigma_i = \epsilon_i(\rho_i I + \bar{H})^{-1}$ is the covariance in Proposition B.3.

Step 2: Diagonalization. Let $\bar{H} = \sum_{k=1}^n \lambda_k e_k e_k^\top$ be a spectral decomposition of $\bar{H}$, with $\lambda_k > 0$ and $e_1, \dots, e_n$ orthonormal. For $y \in \mathbb{R}^n$, let $y_k \coloneqq \langle e_k, y \rangle$, so that $\|y\|_{\bar{H}}^2 = \sum_{k=1}^n \lambda_k y_k^2$. Linear combinations and inverses of matrices that are diagonal in the basis $e_1, \dots, e_n$ are again diagonal in this basis, with the corresponding eigenvalues. Hence, Σ_i and $S_i \coloneqq \sum_{j \neq i} \Sigma_j$ are diagonal in this basis, with eigenvalues $\frac{\epsilon_i}{\rho_i + \lambda_k}$ and

$$s_{i,k} \coloneqq \sum_{j \neq i} \frac{\epsilon_j}{\rho_j + \lambda_k} > 0,$$

respectively. In particular, $-H \sum_{j \neq i} \Sigma_j = \bar{H} S_i$ has eigenvalues $\lambda_k s_{i,k} > 0$, so that $\bar{\tau}_i = (\max_k \lambda_k s_{i,k})^{-1} \in (0, +\infty)$ and, for $\tau_i > 0$,

$$\tau_i < \bar{\tau}_i \quad \iff \quad \tau_i s_{i,k} < \frac{1}{\lambda_k} \ \forall k \in [n] \quad \iff \quad \tau_i S_i \prec \bar{H}^{-1}.$$

Therefore, $\tau \in \mathcal{T}$ if and only if $\tau_i > 0$ for all i and assumption (12) of Proposition B.3 holds with $\bar{H}$ in place of H.

Step 3: Shared reward at equilibrium. Fix $\tau \in \mathcal{T}$. By Steps 1 and 2, Proposition B.3 applies, so that $x_i^\star(\tau) = \mathcal{N}(m_i^\star(\tau), \Sigma_i)$, where Σ_i does not depend on τ and $m_i^\star(\tau)$ is given by (14) with $W_i = W_i(\tau_i) \coloneqq (\bar{H}^{-1} - \tau_i S_i)^{-1}$. By Step 2, $W_i(\tau_i)$ and $M(\tau) \coloneqq I + \sum_{j=1}^N \frac{1}{\rho_j} W_j(\tau_j)$ are diagonal in the basis $e_1, \dots, e_n$, with eigenvalues

$$w_{i,k}(\tau_i) \coloneqq \Big(\frac{1}{\lambda_k} - \tau_i s_{i,k}\Big)^{-1} > 0 \qquad \text{and} \qquad \mu_k(\tau) \coloneqq 1 + \sum_{j=1}^N \frac{w_{j,k}(\tau_j)}{\rho_j} > 1,$$

respectively. Hence, by (14), $\big(m_i^\star(\tau)\big)_k = \frac{w_{i,k}(\tau_i)}{\rho_i \mu_k(\tau)} \bar{a}_k$, and, by definition of $\mu_k(\tau)$,

$$\Big(\sum\nolimits_{i=1}^N m_i^\star(\tau) - \bar{a}\Big)_k = \frac{\mu_k(\tau) - 1}{\mu_k(\tau)} \, \bar{a}_k - \bar{a}_k = -\frac{\bar{a}_k}{\mu_k(\tau)}.$$

Since $a_1, \dots, a_N$ are independent in the definition of J, the vector $\sum_{i=1}^N a_i - \bar{a}$ is Gaussian with mean $\sum_{i=1}^N m_i^\star(\tau) - \bar{a}$ and covariance $\sum_{i=1}^N \Sigma_i$. Using $\mathbb{E}[\|X\|_{\bar{H}}^2] = \|\mathbb{E}[X]\|_{\bar{H}}^2 + \operatorname{tr}(\bar{H} \operatorname{Cov}(X))$ for every random vector X with finite second moment, we obtain

$$J(\tau) = -\frac{1}{2} \sum_{k=1}^n \frac{\lambda_k \bar{a}_k^2}{\mu_k(\tau)^2} + C, \qquad C \coloneqq -\frac{1}{2} \operatorname{tr}\Big(\bar{H} \sum\nolimits_{i=1}^N \Sigma_i\Big) - \frac{1}{2} \langle h, H^{-1} h \rangle,$$

where C does not depend on τ.

Step 4: Monotonicity. Let $\tau, \tau' \in \mathcal{T}$ with $\tau_i' \geq \tau_i$ for all i. Since $s_{i,k} > 0$, each $w_{i,k}$ is strictly increasing on $(0, \bar{\tau}_i)$. Hence, $\mu_k(\tau') \geq \mu_k(\tau) > 0$ for all k, with strict inequality for all k if $\tau' \neq \tau$. Therefore,

$$J(\tau') - J(\tau) = \frac{1}{2} \sum_{k=1}^n \lambda_k \bar{a}_k^2 \left(\frac{1}{\mu_k(\tau)^2} - \frac{1}{\mu_k(\tau')^2} \right) \geq 0.$$

If, moreover, $\tau' \neq \tau$, then the inequality is strict, since $h \neq 0$ gives $\bar{a} = -H^{-1} h \neq 0$, so that $\bar{a}_k \neq 0$ for some k. □

The proof of Theorem 4.1 is now a direct consequence:

Proof of Theorem 4.1. The proof follows from Theorem B.4 and Proposition B.3, using $\rho_i = \rho$, $\tau_i = \tau$, and $\epsilon_i = \epsilon$ for all agents. □

To conclude, we provide more details on Example 4.2:

Example B.5 (Details on Example 4.2)**.** For the game in Example 4.2, we have $H = 1$ and $\rho_i = 1$ in the notation of Proposition B.3, so that $\Sigma_i = \frac{\epsilon_i}{2}$ and $W_i = \big(1 - \frac{\tau_i \epsilon_j}{2}\big)^{-1} = \frac{2}{2 - \tau_i \epsilon_j}$, with $j \neq i$, so that

$$\bar{U}_i(m_i, m_{-i}) = -\frac{1}{2} \left(1 + \frac{2}{2 - \tau_i \epsilon_j}\right) m_i^2 - \frac{2}{2 - \tau_i \epsilon_j} m_i \left(m_{-i} - \bar{a}\right).$$

With $\tau_1 = \tau_2 = \tau$ and $\epsilon_1 = \epsilon_2 = \epsilon$ (so that (12) reads $\tau\epsilon < 2$), (14) gives $m_i^\star = \frac{\bar{a}}{3-\tau\epsilon/2}$, and we obtain the expected utility of each player from the expected shared reward $J(\tau)$ minus the expected personal cost and it can be shown to be

$$-\frac{\bar{a}^2}{2}\frac{2-\tau\epsilon+(\frac{\tau\epsilon}{2})^2}{(3-\frac{\tau\epsilon}{2})^2}-\frac{3\epsilon}{4},$$

where the constant $-\frac{3\epsilon}{4}$ results from the variances $\Sigma_i = \frac{\epsilon}{2}$ of the mixed strategies and is therefore not dependent on τ but only on ϵ.

B.2 Proof of Theorem 4.4

In this section, we provide the proof of Theorem 4.4. To do so, we first derive several helper-lemmas which guarantee that (i) players' strategies in an RQE are on the interior of the simplex, (ii) the KL-divergence between any strategy and an RQE strategy is uniformly bounded, and (iii) we can relate the degree of free-riding to a *lower bound* on the distance between equilibrium strategies. To begin, we define some useful quantities that capture measures of the spread of players' costs and reward functions $c_{\max} \coloneqq \max_a c(a)$, $c_{\min} \coloneqq \min_a c(a)$, $v_{\min} \coloneqq \min_{a_1,a_2} R(a_1,a_2) - c_{\max}$, $v_{\max} \coloneqq \max_{a_1,a_2} R(a_1,a_2) - c_{\min}$, and $\bar{v} \coloneqq v_{\max} - v_{\min}$. Without loss of generality, we assume $c_{\max} > c_{\min}$; otherwise, both players always pay the same cost, so that no RQE exhibits free-riding of any degree $\delta > 0$ and Theorem B.9 holds trivially. Throughout the proof, we let $n \coloneqq |\mathcal{A}|$ be the number of actions and $\Delta_n = \Delta(\mathcal{A})$ be the probability simplex. For a mixed strategy $x \in \Delta_n$, we denote by $x(a) \in [0,1]$ the probability mass assigned to the action a.

Lemma B.6. *Suppose that players have degree of entropy regularization ϵ. At an RQE, a player's strategy x_i^* has support on all actions, with probability at least $m = \frac{\exp(-\bar{v}/\epsilon)}{n}$.*

Proof. In an RQE, a player's strategy x_i^* must satisfy

$$x_i^* = \arg\max_{x_i \in \Delta_n} \ \langle x_i, Rp^* \rangle - \langle c, x_i \rangle - \epsilon H(x_i),$$

where $p^* = \arg\min_{p\in\Delta_n} \langle x_i^*, Rp\rangle + \frac{1}{\tau}\,\mathrm{KL}(p, x_{-i}^*)$. Indeed, since p^* is unique by strict convexity of $\mathrm{KL}(\cdot, x_{-i}^*)$, by Danskin's theorem the risk-adjusted utility $U^\tau(\cdot, x_{-i}^*)$ has gradient $Rp^* - c$ at x_i^*, so x_i^* satisfies the first-order optimality conditions of this strictly concave problem. Letting $v \coloneqq Rp^* - c$, the classic results show that x_i^* is a Boltzmann distribution of the form:

$$x_i^*(a) = \frac{\exp(v(a)/\epsilon)}{\sum_{a'\in\mathcal{A}} \exp(v(a')/\epsilon)},$$

where $x_i^*(a)$ is the probability mass assigned to action a. By minimizing the numerator and maximizing the denominator of the right hand side we find that:

$$x_i^*(a) \geq \frac{\exp(-\bar{v}/\epsilon)}{n},$$

which completes the proof. □

Lemma B.7. *Suppose $x_2 \in \sigma(\Delta_n)$ where $\sigma(\Delta_n) \coloneqq \{x \in \Delta_n : x(a) \geq m \text{ for all } a \in \mathcal{A}\}$ where $m \in (0, \frac{1}{n}]$. Then the KL divergence between any distribution $q \in \Delta_n$ and x_2 is bounded by*

$$\sup_{q\in\Delta_n} \mathrm{KL}(q, x_2) \leq \log\left(\frac{1}{m}\right).$$

Proof. We directly have

$$\mathrm{KL}(q, x_2) = \sum_{a\in\mathcal{A}} q(a)\log\left(\frac{q(a)}{x_2(a)}\right) \leq \sum_a q(a)\log\left(\frac{1}{m}\right) = \log\left(\frac{1}{m}\right)$$

since $x_2(a) \geq m$. □

Lemma B.8. *Let $x_1, x_2 \in \Delta_n$. If $|\langle c, x_1 - x_2\rangle| \geq \delta$, then*

$$\|x_1 - x_2\|_1 \geq \frac{2\delta}{c_{\max} - c_{\min}}.$$

Proof. Let $d \coloneqq x_1 - x_2$. Since $x_1, x_2 \in \Delta_n$, $\sum_a d(a) = 0$. We use the fact that, for any $d \in \mathbb{R}^n$ with $\sum_a d(a) = 0$, $|\langle c, d\rangle| \leq \frac{c_{\max} - c_{\min}}{2} \|d\|_1$ (indeed, with $\bar{c} \coloneqq (c_{\max} + c_{\min})/2$ and $\mathbf{1}$ the vector of all ones, $|\langle c, d\rangle| = |\langle c - \bar{c}\mathbf{1}, d\rangle| \leq \|c - \bar{c}\mathbf{1}\|_\infty \|d\|_1 = \frac{c_{\max} - c_{\min}}{2} \|d\|_1$). This directly yields

$$\delta \leq |\langle c, d\rangle| \leq \frac{c_{\max} - c_{\min}}{2} \|x_1 - x_2\|_1.$$

Rearranging yields the lower bound. □

We can now prove Theorem 4.4, restated below with all constants made explicit.

Theorem B.9. *Consider a two-player game as in Section 4.2, let $\delta > 0$, and let both players have entropy regularization $\epsilon > 0$ and risk aversion $\tau > 0$. If*

$$\tau > \frac{2(\epsilon \log n + \bar{v})(c_{\max} - c_{\min})^2}{\epsilon^2 \delta^2},$$

then the game admits no RQE with degree of free-riding greater than δ.

Proof. We prove this by contradiction. We assume that $\tau > \frac{2(\epsilon \log n + \bar{v})(c_{\max} - c_{\min})^2}{\epsilon^2 \delta^2}$ and that x_1 and x_2 constitute a resulting RQE with a level of free-riding larger than δ; i.e., $|\langle c, x_1 - x_2\rangle| \geq \delta$. Define the worst-case utility $W(x)$ and the robust-regularized objective $\Phi(x)$:

$$\begin{aligned} U(x, q) &\coloneqq \langle x, Rq\rangle - \langle c, x\rangle \\ W(x) &\coloneqq \inf_{q \in \Delta_n} U(x, q) \\ \Phi(x) &\coloneqq W(x) - \epsilon H(x) \end{aligned}$$

Since W is concave (as the infimum of affine functions) and H is 1-strongly convex on Δ_n with respect to the ℓ_1 norm (by Pinsker's inequality), Φ is ϵ-strongly concave on Δ_n with respect to the ℓ_1 norm. In particular, Φ has a unique maximizer $x^\star \coloneqq \arg\max_{x \in \Delta_n} \Phi(x)$.

Let us consider player 1's risk-adjusted utility. Without loss of generality, we develop the following for x_1; since R is symmetric and both players share the cost c, player 2's objective has the same form, so the same arguments apply to x_2 with the same Φ and, hence, the same maximizer $x^\star$. We first show that we can lower bound the risk-adjusted objective by our worst case objective for any $x_1, x_2 \in \Delta_n$:

$$U^\tau(x_1, x_2) = \min_{q \in \Delta_n} U(x_1, q) + \frac{1}{\tau} \mathrm{KL}(q, x_2) \geq \min_{q \in \Delta_n} U(x_1, q) = W(x_1).$$

Now consider an RQE made up of strategies (x_1, x_2). We upper bound the risk-adjusted objective by using the fact that x_1 and x_2 are on the interior of the simplex since they are entropy-regularized best responses (Lemma B.6), and thus the KL term is bounded by Lemma B.7. Letting $K \coloneqq \log n + \frac{\bar{v}}{\epsilon}$ be the resulting upper bound (which depends solely on R, c, ϵ and the dimension n) we conclude that

$$U^\tau(x_1, x_2) = \min_{q \in \Delta_n} U(x_1, q) + \frac{1}{\tau} \mathrm{KL}(q, x_2) \leq W(x_1) + \frac{K}{\tau}.$$

Putting these together, we have

$$\Phi(x_1) = W(x_1) - \epsilon H(x_1) \leq U^\tau(x_1, x_2) - \epsilon H(x_1) \leq W(x_1) + \frac{K}{\tau} - \epsilon H(x_1) = \Phi(x_1) + \frac{K}{\tau}.$$

Using the definition of an RQE, we can further develop this to find that, for all $x \in \Delta_n$:

$$\Phi(x_1) + \frac{K}{\tau} \geq U^\tau(x_1, x_2) - \epsilon H(x_1) \geq U^\tau(x, x_2) - \epsilon H(x) \geq \Phi(x)$$

Moreover, since Φ is ϵ-strongly concave and $x^\star$ maximizes it over Δ_n, we have

$$\Phi(x_1) \leq \Phi(x^\star) - \frac{\epsilon}{2} \|x_1 - x^\star\|_1^2$$

Thus,

$$\frac{\epsilon}{2} \|x_1 - x^\star\|_1^2 \leq \Phi(x^\star) - \Phi(x_1) \leq \frac{K}{\tau} \implies \|x_1 - x^\star\|_1 \leq \sqrt{\frac{2K}{\epsilon\tau}}.$$

The same bound holds for x_2, and thus via the triangle inequality we obtain

$$\|x_1 - x_2\|_1 \leq \|x_1 - x^\star\|_1 + \|x_2 - x^\star\|_1 \leq 2\sqrt{\frac{2K}{\epsilon\tau}}.$$

Via Lemma B.8, however we have that

$$\frac{2\delta}{c_{\max} - c_{\min}} \leq \|x_1 - x_2\|_1 \leq \|x_1 - x^\star\|_1 + \|x_2 - x^\star\|_1 \leq 2\sqrt{\frac{2K}{\epsilon\tau}}.$$

Re-arranging, we find that

$$\tau \leq \frac{2K(c_{\max} - c_{\min})^2}{\epsilon\delta^2}.$$

Plugging in our value for $K = \log n + \frac{\bar{v}}{\epsilon}$, this implies that

$$\tau \leq \frac{2(\epsilon \log n + \bar{v})(c_{\max} - c_{\min})^2}{\epsilon^2\delta^2}.$$

However, we assumed that the RQE resulted from the use of a $\tau > \frac{2(\epsilon \log n + \bar{v})(c_{\max} - c_{\min})^2}{\epsilon^2\delta^2}$ which is a contradiction. □

C Convergence Guarantee for the Meta Algorithm

In this section, we establish a convergence guarantee for the projected policy-gradient instantiation of our meta algorithm in the two-player finite matrix game setting (Theorem C.4). The argument proceeds in three stages: Section C.1 introduces the lifted four-player regularized game and the operator G whose variational inequality on $\mathcal{Z}_\delta$ characterizes its RQE; Section C.2 establishes an a priori interior bound on the RQE and two properties of G (Lipschitz continuity and strong monotonicity) that drive the analysis; Section C.3 combines these properties into a linear-convergence guarantee for the projected gradient iteration.

C.1 Lifted game and operator formulation

For each player $i \in \{1, 2\}$, let R_i denote the payoff matrix of the original two-player matrix game. We work with the lifted state

$$z = (x_1, x_2, p_1, p_2) \in \mathcal{Z}, \qquad \mathcal{Z} := \Delta(\mathcal{A}_1) \times \Delta(\mathcal{A}_2) \times \Delta(\mathcal{A}_2) \times \Delta(\mathcal{A}_1),$$

in which x_i is the mixed strategy of original player i and p_i is the mixed strategy of the adversary against player i. The original-player objective is

$$J_i(x_i, x_{-i}, p_i) = x_i^\top R_i p_i + \frac{1}{\tau_i}\,\mathrm{KL}(p_i, x_{-i}) - \epsilon_i H(x_i),$$

and the corresponding adversary objective is

$$\bar{J}_i(x_i, x_{-i}, p_i) = -x_i^\top R_i p_i - \frac{1}{\tau_i}\,\mathrm{KL}(p_i, x_{-i}).$$

For any $\delta > 0$, define the δ-interior set

$$\mathcal{Z}_\delta := \left\{ z = (x_1, x_2, p_1, p_2) \in \mathcal{Z} \;:\; \begin{array}{ll} x_i(a) \geq \delta & \forall a \in \mathcal{A}_i,\ i \in \{1, 2\}, \\ p_i(b) \geq \delta & \forall b \in \mathcal{A}_{-i},\ i \in \{1, 2\} \end{array} \right\}.$$

Thus $\mathcal{Z}$ is the ambient product of simplices, while $\mathcal{Z}_\delta$ is the closed convex feasible set used throughout the convergence argument. We define the lifted game operator $G : \mathcal{Z}_\delta \to \mathbb{R}^d$ by

$$G(z) := \Big(-\nabla_{x_1} J_1(z),\ -\nabla_{x_2} J_2(z),\ -\nabla_{p_1} \bar{J}_1(z),\ -\nabla_{p_2} \bar{J}_2(z) \Big), \tag{16}$$

where d is the total dimension of z.

Throughout, by an RQE of the lifted game we mean a Nash equilibrium $z^\star = (x_1^\star, x_2^\star, p_1^\star, p_2^\star)$: each $x_i^\star$ maximizes $J_i(\cdot, x_{-i}^\star, p_i^\star)$ over $\Delta(\mathcal{A}_i)$ and each $p_i^\star$ maximizes $\bar{J}_i(x_i^\star, x_{-i}^\star, \cdot)$ over $\Delta(\mathcal{A}_{-i})$. Since, for fixed x_{-i}, J_i is strictly concave in x_i and strictly convex in p_i, the minimax theorem shows that this holds if and only if $(x_1^\star, x_2^\star)$ is an RQE of the original game (Definition 3.1) and $p_i^\star$ attains the infimum in (1) at $(x_i^\star, x_{-i}^\star)$ [38, Proposition 1].

C.2 Quantitative properties of the operator G

The convergence proof rests on an a priori lower bound on the coordinates of any RQE (Lemma C.1) and two quantitative properties of G: Lipschitz continuity on $\mathcal{Z}_\delta$ (Lemma C.2) and strong monotonicity on $\mathcal{Z}_\delta$ (Lemma C.3).

Lemma C.1 (Interior bound on the RQE)**.** *Consider the lifted four-player game above for a two-player matrix game with bounded payoff matrices R_1 and R_2, and let*

$$z^\star = (x_1^\star, x_2^\star, p_1^\star, p_2^\star)$$

be any RQE. For each $i \in \{1, 2\}$, define the payoff-span constant

$$S_i := \max_{\substack{a\in\mathcal{A}_i \\ b\in\mathcal{A}_{-i}}} R_i(a, b) - \min_{\substack{a\in\mathcal{A}_i \\ b\in\mathcal{A}_{-i}}} R_i(a, b).$$

Then every coordinate of every block of $z^\star$ is bounded away from zero. More precisely,

$$x_i^\star(a) \geq \underline{\delta}_{x,i} := \frac{1}{|\mathcal{A}_i|} \exp\left(-\frac{S_i}{\epsilon_i}\right), \qquad \forall a \in \mathcal{A}_i,$$

and

$$p_i^\star(b) \geq \underline{\delta}_{p,i} := \frac{1}{|\mathcal{A}_{-i}|} \exp\left(-\frac{S_{-i}}{\epsilon_{-i}} - \tau_i S_i\right), \qquad \forall b \in \mathcal{A}_{-i}.$$

Consequently,

$$\underline{\delta} := \min\{\underline{\delta}_{x,1}, \underline{\delta}_{x,2}, \underline{\delta}_{p,1}, \underline{\delta}_{p,2}\} > 0 \tag{17}$$

satisfies $z^\star \in \mathcal{Z}_{\underline{\delta}}$.

Proof. For a vector v, write $\mathrm{sp}(v) := \max_j v_j - \min_j v_j$. Fix $i \in \{1, 2\}$ and define

$$u_i^\star := R_i p_i^\star \in \mathbb{R}^{|\mathcal{A}_i|}.$$

Since each coordinate $u_i^\star(a)$ is a convex combination of entries of R_i, we have $\mathrm{sp}(u_i^\star) \leq S_i$. Now $x_i^\star$ maximizes

$$x_i^\top R_i p_i^\star - \epsilon_i H(x_i)$$

over $\Delta(\mathcal{A}_i)$. The entropy term forces the optimizer into the simplex interior, so the KKT conditions give a multiplier $\alpha_i \in \mathbb{R}$ such that

$$(R_i p_i^\star)(a) - \epsilon_i\big(\log x_i^\star(a) + 1\big) - \alpha_i = 0, \qquad \forall a \in \mathcal{A}_i.$$

Rearranging and normalizing over a yields

$$x_i^\star(a) = \frac{\exp((R_i p_i^\star)(a)/\epsilon_i)}{\sum_{a'\in\mathcal{A}_i} \exp((R_i p_i^\star)(a')/\epsilon_i)}.$$

Therefore,

$$\begin{aligned} x_i^\star(a) &\geq \frac{\exp((R_i p_i^\star)(a)/\epsilon_i)}{|\mathcal{A}_i| \exp(\max_{a'}(R_i p_i^\star)(a')/\epsilon_i)} \\ &= \frac{1}{|\mathcal{A}_i|} \exp\left(-\frac{\max_{a'}(R_i p_i^\star)(a') - (R_i p_i^\star)(a)}{\epsilon_i}\right) \\ &\geq \frac{1}{|\mathcal{A}_i|} \exp\left(-\frac{\mathrm{sp}(R_i p_i^\star)}{\epsilon_i}\right) \geq \frac{1}{|\mathcal{A}_i|} \exp\left(-\frac{S_i}{\epsilon_i}\right) = \underline{\delta}_{x,i}. \end{aligned}$$

Next define $q_i^\star := R_i^\top x_i^\star \in \mathbb{R}^{|\mathcal{A}_{-i}|}$. Again, each coordinate $q_i^\star(b)$ is a convex combination of entries of R_i, so $\mathrm{sp}(q_i^\star) \leq S_i$. Because $x_{-i}^\star$ is strictly positive by the previous step, the adversary problem

$$\min_{p_i\in\Delta(\mathcal{A}_{-i})} \left\langle q_i^\star, p_i \right\rangle + \frac{1}{\tau_i} \mathrm{KL}(p_i, x_{-i}^\star)$$

has an interior minimizer $p_i^\star$. Its KKT conditions give a multiplier $\beta_i \in \mathbb{R}$ such that

$$q_i^\star(b) + \frac{1}{\tau_i}\left(\log \frac{p_i^\star(b)}{x_{-i}^\star(b)} + 1\right) - \beta_i = 0, \qquad \forall b \in \mathcal{A}_{-i}.$$

Hence $p_i^\star$ has the Gibbs form

$$p_i^\star(b) = \frac{x_{-i}^\star(b)\exp(-\tau_i q_i^\star(b))}{\sum_{b'\in\mathcal{A}_{-i}} x_{-i}^\star(b')\exp(-\tau_i q_i^\star(b'))}.$$

Using $\sum_{b'} x_{-i}^\star(b') = 1$, we obtain

$$\sum_{b'} x_{-i}^\star(b')\exp(-\tau_i q_i^\star(b')) \le \exp\Big(-\tau_i \min_{b'} q_i^\star(b')\Big).$$

Therefore, for every $b \in \mathcal{A}_{-i}$,

$$\begin{aligned} p_i^\star(b) &\ge x_{-i}^\star(b)\exp\Big(-\tau_i\big(q_i^\star(b) - \min_{b'} q_i^\star(b')\big)\Big) \\ &\ge x_{-i}^\star(b)\exp(-\tau_i \operatorname{sp}(q_i^\star)) \ge \underline{\delta}_{x,-i}\exp(-\tau_i S_i) = \underline{\delta}_{p,i}. \end{aligned}$$

Taking the minimum over the four blockwise lower bounds proves the claim. □

The next two lemmas provide explicit constants for Lipschitz continuity and strong monotonicity of the KL- and negative-entropy-induced operator G on $\mathcal{Z}_\delta$.

Lemma C.2 (Lipschitz continuity of G)**.** *Let $\|\cdot\|_{\mathrm{op}}$ denote the Euclidean operator norm of a matrix. Then G is L_δ-Lipschitz on $\mathcal{Z}_\delta$ with*

$$L_\delta := \left(\left(\tfrac{\epsilon_1}{\delta}+\|R_1\|_{\mathrm{op}}+\tfrac{1}{\tau_2\delta}\right)^2+\left(\tfrac{\epsilon_2}{\delta}+\|R_2\|_{\mathrm{op}}+\tfrac{1}{\tau_1\delta}\right)^2+\left(\|R_1\|_{\mathrm{op}}+\tfrac{1}{\tau_1\delta}\right)^2+\left(\|R_2\|_{\mathrm{op}}+\tfrac{1}{\tau_2\delta}\right)^2\right)^{1/2}.$$

Proof. Let $z = (x_1, x_2, p_1, p_2)$ and $z' = (x_1', x_2', p_1', p_2')$ lie in $\mathcal{Z}_\delta$. Since every coordinate of x_i, x_i', p_i, p_i' is at least δ, the scalar map $s \mapsto \log s$ has derivative bounded by δ^{-1} on the relevant interval. Hence, component-wise,

$$\|\log x_i - \log x_i'\|_2 \le \tfrac{1}{\delta}\|x_i - x_i'\|_2, \qquad \|\log p_i - \log p_i'\|_2 \le \tfrac{1}{\delta}\|p_i - p_i'\|_2.$$

Using

$$G(z) = \begin{pmatrix} -R_1 p_1 + \epsilon_1(\log x_1 + \mathbf{1}) \\ -R_2 p_2 + \epsilon_2(\log x_2 + \mathbf{1}) \\ R_1^\top x_1 + \tau_1^{-1}(\log p_1 - \log x_2 + \mathbf{1}) \\ R_2^\top x_2 + \tau_2^{-1}(\log p_2 - \log x_1 + \mathbf{1}) \end{pmatrix},$$

we obtain

$$\begin{aligned} \|G_1(z) - G_1(z')\|_2 &\le \|R_1\|_{\mathrm{op}}\|p_1 - p_1'\|_2 + \tfrac{\epsilon_1}{\delta}\|x_1 - x_1'\|_2, \\ \|G_2(z) - G_2(z')\|_2 &\le \|R_2\|_{\mathrm{op}}\|p_2 - p_2'\|_2 + \tfrac{\epsilon_2}{\delta}\|x_2 - x_2'\|_2, \\ \|G_3(z) - G_3(z')\|_2 &\le \|R_1\|_{\mathrm{op}}\|x_1 - x_1'\|_2 + \tfrac{1}{\tau_1\delta}\|p_1 - p_1'\|_2 + \tfrac{1}{\tau_1\delta}\|x_2 - x_2'\|_2, \\ \|G_4(z) - G_4(z')\|_2 &\le \|R_2\|_{\mathrm{op}}\|x_2 - x_2'\|_2 + \tfrac{1}{\tau_2\delta}\|p_2 - p_2'\|_2 + \tfrac{1}{\tau_2\delta}\|x_1 - x_1'\|_2. \end{aligned}$$

Summing these four bounds and applying the Cauchy–Schwarz inequality gives $\|G(z) - G(z')\|_2 \le L_\delta\|z - z'\|_2$, as claimed. □

Lemma C.3 (Strong monotonicity of G)**.** *Suppose $4\epsilon_1\tau_2\delta > 1$ and $4\epsilon_2\tau_1\delta > 1$. Define*

$$\mu_{i,\delta} := \frac{1}{2}\left(\epsilon_i + \frac{\delta}{\tau_{-i}} - \sqrt{\left(\epsilon_i - \frac{\delta}{\tau_{-i}}\right)^2 + \frac{1}{\tau_{-i}^2}}\right), \qquad i \in \{1,2\},$$

and let $\mu_\delta := \min\{\mu_{1,\delta}, \mu_{2,\delta}\} > 0$. Then G is μ_δ-strongly monotone on $\mathcal{Z}_\delta$. That is,

$$\langle G(z) - G(z'), z - z'\rangle \ge \mu_\delta\|z - z'\|_2^2.$$

Proof. For $z = (x_1, x_2, p_1, p_2) \in \mathcal{Z}_\delta$, let

$$D_{x_i} := \operatorname{diag}(x_i^{-1}), \qquad D_{p_i} := \operatorname{diag}(p_i^{-1}).$$

From the explicit formula for G, its Jacobian is

$$\nabla G(z) = \begin{bmatrix} \epsilon_1 D_{x_1} & 0 & -R_1 & 0 \\ 0 & \epsilon_2 D_{x_2} & 0 & -R_2 \\ R_1^\top & -\tau_1^{-1} D_{x_2} & \tau_1^{-1} D_{p_1} & 0 \\ -\tau_2^{-1} D_{x_1} & R_2^\top & 0 & \tau_2^{-1} D_{p_2} \end{bmatrix}.$$

Therefore the symmetric part is

$$\frac{\nabla G(z) + \nabla G(z)^\top}{2} = \begin{bmatrix} \epsilon_1 D_{x_1} & 0 & 0 & -(2\tau_2)^{-1} D_{x_1} \\ 0 & \epsilon_2 D_{x_2} & -(2\tau_1)^{-1} D_{x_2} & 0 \\ 0 & -(2\tau_1)^{-1} D_{x_2} & \tau_1^{-1} D_{p_1} & 0 \\ -(2\tau_2)^{-1} D_{x_1} & 0 & 0 & \tau_2^{-1} D_{p_2} \end{bmatrix}.$$

The payoff blocks cancel from the symmetric part. After permuting the coordinates to (x_1, p_2, x_2, p_1), this matrix becomes block diagonal with blocks

$$\mathcal{M}_1(z) := \begin{bmatrix} \epsilon_1 D_{x_1} & -(2\tau_2)^{-1} D_{x_1} \\ -(2\tau_2)^{-1} D_{x_1} & \tau_2^{-1} D_{p_2} \end{bmatrix}, \qquad \mathcal{M}_2(z) := \begin{bmatrix} \epsilon_2 D_{x_2} & -(2\tau_1)^{-1} D_{x_2} \\ -(2\tau_1)^{-1} D_{x_2} & \tau_1^{-1} D_{p_1} \end{bmatrix}.$$

Since these blocks are diagonal across coordinates, it suffices to bound the smallest eigenvalue of the corresponding 2×2 coordinate matrices. For $a \in \mathcal{A}_1$,

$$M_{1,a}(z) = \begin{bmatrix} \epsilon_1 / x_1(a) & -1/(2\tau_2 x_1(a)) \\ -1/(2\tau_2 x_1(a)) & 1/(\tau_2 p_2(a)) \end{bmatrix} = \frac{1}{x_1(a)} \begin{bmatrix} \epsilon_1 & -1/(2\tau_2) \\ -1/(2\tau_2) & x_1(a)/(\tau_2 p_2(a)) \end{bmatrix}.$$

Now $x_1(a) \in [\delta, 1]$, $p_2(a) \in [\delta, 1]$, so $x_1(a)/p_2(a) \geq \delta$ and $x_1(a)^{-1} \geq 1$. Hence

$$\lambda_{\min}\big(M_{1,a}(z)\big) \geq \lambda_{\min} \begin{bmatrix} \epsilon_1 & -1/(2\tau_2) \\ -1/(2\tau_2) & \delta/\tau_2 \end{bmatrix} = \mu_{1,\delta}.$$

The condition $4\epsilon_1\tau_2\delta > 1$ is exactly the positivity condition for this 2×2 determinant, so $\mu_{1,\delta} > 0$. The same argument gives $\lambda_{\min}\big(M_{2,b}(z)\big) \geq \mu_{2,\delta} > 0$ for every $b \in \mathcal{A}_2$. Therefore,

$$\frac{\nabla G(z) + \nabla G(z)^\top}{2} \succeq \mu_\delta I, \qquad \forall z \in \mathcal{Z}_\delta.$$

Now let $z, z' \in \mathcal{Z}_\delta$, set $v := z - z'$, and define $\phi(t) := z' + tv$. Since $\mathcal{Z}_\delta$ is convex, $\phi(t) \in \mathcal{Z}_\delta$ for all $t \in [0, 1]$. Then

$$\begin{aligned} \langle G(z) - G(z'),\, z - z' \rangle &= \int_0^1 v^\top \nabla G(\phi(t)) v \, dt = \int_0^1 v^\top \frac{\nabla G(\phi(t)) + \nabla G(\phi(t))^\top}{2} v \, dt \\ &\geq \mu_\delta \int_0^1 \|v\|_2^2 \, dt = \mu_\delta \|z - z'\|_2^2. \end{aligned}$$

Thus G is μ_δ-strongly monotone on $\mathcal{Z}_\delta$. □

C.3 Linear convergence of projected policy gradient

Combining the three lemmas above, we obtain our main convergence result: on the feasible set $\mathcal{Z}_{\underline{\delta}}$ given by Lemma C.1, the projected policy-gradient iteration converges linearly to the unique RQE.

Theorem C.4 (Projected policy gradient converges to the RQE in two-player matrix games)**.** *Consider the lifted four-player regularized game associated with a two-player matrix game with bounded payoff matrices R_1, R_2, and let $\underline{\delta} > 0$ be the interior bound from Lemma C.1. Assume the regularization parameters satisfy*

$$4\epsilon_1\tau_2\underline{\delta} > 1, \qquad 4\epsilon_2\tau_1\underline{\delta} > 1, \tag{18}$$

and let $L_{\underline{\delta}}$ and $\mu_{\underline{\delta}}$ denote, respectively, the Lipschitz constant from Lemma C.2 and the strong-monotonicity modulus from Lemma C.3, both evaluated at $\delta = \underline{\delta}$. Then:

1. *the lifted game admits a unique RQE $z^\star \in \mathcal{Z}_{\underline{\delta}}$;*

2. *for every initialization $z^0 \in \mathcal{Z}_{\underline{\delta}}$ and every step size satisfying*

$$0 < \eta < \frac{2\mu_{\underline{\delta}}}{L_{\underline{\delta}}^2},$$

the projected policy-gradient iterates

$$z^{k+1} = \Pi_{\mathcal{Z}_{\underline{\delta}}}\big(z^k - \eta G(z^k)\big) \tag{19}$$

converge linearly to $z^\star$: setting $\rho_{\underline{\delta}} := 1 - 2\eta\mu_{\underline{\delta}} + \eta^2 L_{\underline{\delta}}^2 \in [0,1)$,

$$\|z^k - z^\star\|_2^2 \le \rho_{\underline{\delta}}^k \, \|z^0 - z^\star\|_2^2, \quad \forall k \ge 0. \tag{20}$$

Proof. We write $\delta := \underline{\delta}$ throughout the proof and proceed in four steps.

Step 1: Existence, variational-inequality characterization, and uniqueness. By the Gibbs variational principle, the infimum in (1) equals $-\tau_i^{-1} \log \sum_{b \in \mathcal{A}_{-i}} x_{-i}(b) \exp\big(-\tau_i (R_i^\top x_i)(b)\big)$, so $U_i^{\tau_i,\epsilon_i}$ is continuous on $\Delta(\mathcal{A}_1)\times\Delta(\mathcal{A}_2)$ and concave in x_i. Hence the original game has an RQE [48] and, by the equivalence in Section C.1, so does the lifted game. Let $z^\star = (x_1^\star, x_2^\star, p_1^\star, p_2^\star)$ be any RQE; by Lemma C.1, $z^\star \in \mathcal{Z}_\delta$ and each block of $z^\star$ lies in the relative interior of its simplex. Since $z^\star$ is an RQE of the lifted four-player game, each block is a best response to the others. For each $i \in \{1,2\}$, the function $x_i \mapsto J_i(x_i, x_{-i}^\star, p_i^\star)$ is concave and differentiable at $x_i^\star$, while $p_i \mapsto \bar{J}_i(x_i^\star, x_{-i}^\star, p_i)$ is concave and differentiable at $p_i^\star$. Therefore the first-order optimality condition for concave maximization on a convex set yields

$$\begin{aligned}
\big\langle -\nabla_{x_1} J_1(z^\star),\, x_1 - x_1^\star \big\rangle &\ge 0, \qquad \forall x_1 \in \Delta(\mathcal{A}_1),\\
\big\langle -\nabla_{x_2} J_2(z^\star),\, x_2 - x_2^\star \big\rangle &\ge 0, \qquad \forall x_2 \in \Delta(\mathcal{A}_2),\\
\big\langle -\nabla_{p_1} \bar{J}_1(z^\star),\, p_1 - p_1^\star \big\rangle &\ge 0, \qquad \forall p_1 \in \Delta(\mathcal{A}_2),\\
\big\langle -\nabla_{p_2} \bar{J}_2(z^\star),\, p_2 - p_2^\star \big\rangle &\ge 0, \qquad \forall p_2 \in \Delta(\mathcal{A}_1).
\end{aligned}$$

Adding these four inequalities and using the definition of G in (16), we obtain

$$\langle G(z^\star),\, z - z^\star \rangle \ge 0, \qquad \forall z \in \mathcal{Z}_\delta. \tag{21}$$

Hence the RQE is a solution of $\mathrm{VI}(G, \mathcal{Z}_\delta)$.

To show uniqueness, suppose $z^\star, \tilde{z}^\star \in \mathcal{Z}_\delta$ both solve $\mathrm{VI}(G, \mathcal{Z}_\delta)$. Plugging each into (21) with the other as test point and adding gives

$$\langle G(z^\star) - G(\tilde{z}^\star),\, z^\star - \tilde{z}^\star \rangle \le 0.$$

On the other hand, by Lemma C.3,

$$\langle G(z^\star) - G(\tilde{z}^\star),\, z^\star - \tilde{z}^\star \rangle \ge \mu_\delta \|z^\star - \tilde{z}^\star\|_2^2.$$

Therefore $\|z^\star - \tilde{z}^\star\|_2^2 = 0$, so $z^\star = \tilde{z}^\star$. This proves part (1).

Step 2: Euclidean projection fixed-point identity. We claim that

$$z^\star = \Pi_{\mathcal{Z}_\delta}\big(z^\star - \eta G(z^\star)\big). \tag{22}$$

Indeed, the Euclidean projection $\hat{z} := \Pi_{\mathcal{Z}_\delta}(z^\star - \eta G(z^\star))$ is characterized by

$$\big\langle z^\star - \eta G(z^\star) - \hat{z},\, z - \hat{z} \big\rangle \le 0, \qquad \forall z \in \mathcal{Z}_\delta.$$

Setting $\hat{z} = z^\star$, the left-hand side becomes

$$\big\langle z^\star - \eta G(z^\star) - z^\star,\, z - z^\star \big\rangle = -\eta \langle G(z^\star),\, z - z^\star \rangle \le 0$$

by (21). Hence $z^\star$ satisfies the projection optimality condition, proving (22).

Step 3: One-step contraction estimate. Using (22) and the nonexpansiveness of Euclidean projection onto the closed convex set $\mathcal{Z}_\delta$,

$$\begin{aligned}\|z^{k+1} - z^\star\|_2 &= \|\Pi_{\mathcal{Z}_\delta}(z^k - \eta G(z^k)) - \Pi_{\mathcal{Z}_\delta}(z^\star - \eta G(z^\star))\|_2 \\ &\le \|(z^k - \eta G(z^k)) - (z^\star - \eta G(z^\star))\|_2.\end{aligned}$$

Squaring and expanding gives

$$\|z^{k+1} - z^\star\|_2^2 \le \|z^k - z^\star\|_2^2 - 2\eta\langle G(z^k) - G(z^\star),\, z^k - z^\star\rangle + \eta^2\|G(z^k) - G(z^\star)\|_2^2.$$

Since $z^k, z^\star \in \mathcal{Z}_\delta$, applying Lemmas C.2 and C.3 yields

$$\langle G(z^k) - G(z^\star),\, z^k - z^\star\rangle \ge \mu_\delta\|z^k - z^\star\|_2^2, \qquad \|G(z^k) - G(z^\star)\|_2 \le L_\delta\|z^k - z^\star\|_2.$$

Substituting into the previous inequality yields the one-step contraction

$$\|z^{k+1} - z^\star\|_2^2 \le \big(1 - 2\eta\mu_\delta + \eta^2 L_\delta^2\big)\|z^k - z^\star\|_2^2. \tag{23}$$

Step 4: Linear convergence. Let $\rho_\delta := 1 - 2\eta\mu_\delta + \eta^2 L_\delta^2$. The step-size condition $0 < \eta < 2\mu_\delta/L_\delta^2$ gives $\rho_\delta < 1$. Moreover, $\mu_\delta \le L_\delta$ (by Cauchy–Schwarz combined with the L_δ-Lipschitz bound), so the quadratic ρ_δ has minimum value $1 - \mu_\delta^2/L_\delta^2 \ge 0$. Therefore $\rho_\delta \in [0, 1)$, and iterating (23) yields

$$\|z^k - z^\star\|_2^2 \le \rho_\delta^k\, \|z^0 - z^\star\|_2^2, \qquad \forall k \ge 0. \qquad \square$$

D Mathematical Background of SRPO

In this section, we provide a formal framework for risk aversion and entropy regularization in MARL, and present the motivation for our algorithm SRPO. We first introduce the mathematical framework of discounted general-sum Markov games, and then discuss how to extend risk aversion and entropy regularization from normal-form games to Markov games. Next, we present the pseudocode of SRPO and a short discussion. To better motivate SRPO, we provide policy gradient theorems for each original player and adversary as well as their performance difference lemmas (PDLs), and introduce how to (mathematically) obtain the SRPO loss from the PDLs.

D.1 Discounted General-sum Markov Games

We consider a discounted N-player general-sum Markov game is specified by a tuple:

$$\mathcal{MG} = \{\mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, \{r_i\}_{i=1}^N, \gamma, P, \rho_0\},$$

where $\mathcal{S}$ is the state space of the underlying MDP, $\mathcal{A}_i$ is the action space of player $i \in [N]$, and we use the notation $\mathcal{A} = \prod_{i=1}^N \mathcal{A}_i$ to denote the product action space of both players. We assume that $|\mathcal{S}|$ and $|\mathcal{A}_i|$ are all finite. Here, $r_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function of player i, which we assume to be deterministic. We use $\mathbf{r}$ to denote the joint reward function $\mathbf{r} := (r_i)_{i=1}^N$. Moreover, $\gamma \in [0, 1)$ is the discount factor and $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, where $\Delta(\mathcal{S})$ is the probability simplex of $\mathcal{S}$ and $P(s'|s, \mathbf{a})$ is the probability of the next state being s' given the current state s and the current actions $\mathbf{a} = (a_i)_{i=1}^N$ of the players. We use $\rho_0 \in \Delta(\mathcal{S})$ to denote the initial state distribution.

We focus on *Markov policies*, the class of policies where the action selection probability only depends on the current state instead of the entire gameplay trajectory, i.e., $\pi = (\pi_i)_{i=1}^N$ where $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i), i \in [N]$. Given a product Markov policy π, without considering risk aversion and entropy regularization, player i has an expected discounted cumulative reward given by $\mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t r_i(s_t, a_t)]$.

To incorporate risk aversion and entropy regularization in discounted infinite-horizon Markov games, we slightly overload the notations in normal-form games and consider the following risk-adjusted objective of player i that maximizes $f_i(\pi_i, \pi_{-i}) = \min_{p_i : \mathcal{S} \to \Delta(\mathcal{A}_{-i})} J_i(\pi_i, \pi_{-i}, p_i)$ where J_i is defined as

$$J_i(\pi_i, \pi_{-i}, p_i) = \mathbb{E}_{\pi, p, s_0 \sim \rho_0}\left[\sum_{t=0}^\infty \gamma^t \left(r_i(s_t, \mathbf{a}_t) + \frac{1}{\tau_i} D_i\left(p_i, \pi_{-i}; s_t\right) - \epsilon_i \nu_i(\pi_i; s_t)\right)\right] \tag{24}$$

where the joint actions $\mathbf{a}_t$ are sampled through $a_{i,t} \sim \pi_i(\cdot|s_t)$ and $\mathbf{a}_{-i,t} \sim p_i(\cdot|s_t)$ and the next state s_{t+1} is sampled from $s_{t+1} \sim P(\cdot|s_t, \mathbf{a}_t)$, and the notations of $D_i(p_i, \pi_{-i}; s)$ and $\nu_i(\pi_i; s)$ are abbreviations of $D_i(p_i(\cdot|s), \pi_{-i}(\cdot|s))$ and $\nu_i(\pi_i(\cdot|s))$ respectively. Here, we allow for more generality of the regularizers, and, if we want to be consistent with the main body, we can choose D_i to be KL, and ν_i to be negative entropy H.

Given a set of original player policies π_i and adversarial policies p_i, we define the value function for each state $s \in \mathcal{S}$ as:

$$V_i^{\pi,p}(s) = \mathbb{E}_{\pi,p,s_0=s}\left[\sum_{t=0}^{\infty}\gamma^t\left(r_i(s_t,\mathbf{a}_t) + \frac{1}{\tau_i}D_i\left(p_i,\pi_{-i};s_t\right) - \epsilon_i\nu_i(\pi_i;s_t)\right)\right] \tag{25}$$

so that $J_i(\pi, p) = \mathbb{E}_{s\sim\rho_0}[V_i^{\pi,p}(s)]$, and the Q function as:

$$Q_i^{\pi,p}(s,\mathbf{a}) = r_i(s,\mathbf{a}) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,\mathbf{a})}V_i^{\pi,p}(s'). \tag{26}$$

It is easy to verify that

$$V_i^{\pi,p}(s) = \pi_i(\cdot|s)^T Q_i^{\pi,p}(s,\cdot)p_i(\cdot|s) + \frac{1}{\tau_i}D_i(p_i,\pi_{-i};s) - \epsilon_i\nu_i(\pi_i;s). \tag{27}$$

Finally, we define given policies π_i, p_i, we define the discounted state visitation probability as

$$d_{s_0}^{\pi_i,p_i}(s) \coloneqq (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr_{\pi_i,p_i}(s_t = s|s_0) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t e_{s_0}^T(P^{\pi_i,p_i})^t e_s, \tag{28}$$

where e_s denotes the one-hot vector corresponding to state s. To learn the RQE of the Markov game, we have to optimize the policies π_i to maximize (and p_i to minimize) the risk-adjusted objective (24). Here, we provide some useful tools for optimizing these policies.

D.2 Details of SRPO

When using PPO [51], the objective of agent i is given by

$$\mathcal{L}_i^{\text{IPPO}}(\theta_i,\theta_{-i}) = \mathcal{L}_i^{\text{CLIP}}(\theta_i,\theta_{-i}) - \epsilon_i H(\theta_i)(o_i^t), \tag{29}$$

where the clipped surrogate objective is

$$\mathcal{L}_i^{\text{CLIP}}(\theta_i,\theta_{-i}) = \mathbb{E}_t\Big[\min\Big(r_i^t(\theta_i)\hat{A}_i^t(\theta_i,\theta_{-i}), \text{clip}(r_i^t(\theta_i),1-\delta,1+\delta)\hat{A}_i^t(\theta_i,\theta_{-i})\Big)\Big],$$

with $r_i^t(\theta_i) = \frac{\pi_{\theta_i}(a_i^t|o_i^t)}{\pi_{\theta_{i,\text{old}}}(a_i^t|o_i^t)}$ being the importance sampling ratio, $\hat{A}_i^t(\theta_i,\theta_{-i})$ the estimated advantage at time t (and depends on θ_{-i} as it is computed using samples of the policies of all agents), $\epsilon_i H(\theta_i)(o_i^t)$ a (negative) entropy regularizer, and o_i^t the observation of agent i at time t.

We can now follow the meta-algorithm above to derive strategically risk-averse proximal policy optimization. Specifically, the objective of agent i is

$$\mathcal{L}_i^{\text{SRPO}}(\theta_i,\phi_i) = \mathcal{L}_i^{\text{CLIP}}(\theta_i,\phi_i) - \epsilon_i H(\theta_i)(o_i^t). \tag{30}$$

Since the PPO objective already includes entropy regularization, the additional entropy term in (6) is superfluous. The objective of player i's adversary is then

$$\bar{\mathcal{L}}_i^{\text{SRPO}}(\phi_i,(\theta_i,\theta_{-i})) = -\mathcal{L}_i^{\text{CLIP}}(\theta_i,\phi_i) - \frac{1}{\tau_i}\text{KL}(\phi_i,\theta_{-i}). \tag{31}$$

Overall, SRPO consists of running gradient descent on these objectives, and the iteration structure and complexity of SRPO is very similar to that of IPPO.

We now present the pseudocode for SRPO in Algorithm 1. Notice that although the agents and adversaries are maximizing two losses $\mathcal{L}_i^{\text{SRPO}}$ and $\bar{\mathcal{L}}_i^{\text{SRPO}}$ respectively, we unify these two losses by considering the joint loss:

$$\mathcal{L}_{i,\text{joint}}^{\text{SRPO}}(\theta_i,\theta_{-i},\phi_i) = \mathcal{L}_i^{\text{CLIP}}(\theta_i,\phi_i) + \frac{1}{\tau_i}\text{KL}(\phi_i,\theta_{-i}) - \epsilon_i H(\theta_i)(o_i^t). \tag{32}$$

The update of agent i can be seen as maximizing $\mathcal{L}_{i,\text{joint}}^{\text{SRPO}}$ as a function of θ_i, and the update of adversary i can be seen as minimizing $\mathcal{L}_{i,\text{joint}}^{\text{SRPO}}$ as a function of ϕ_i.

Algorithm 1 Strategically Risk-averse Policy Optimization (SRPO) for $2N$ agents

Initialize: Policies $\{\pi_{\theta_i}\}_{i=1}^N$, adversaries $\{\pi_{\phi_i}\}_{i=1}^N$, and critics $\{V_{\psi_i}\}_{i=1}^N$.
Parameters: Risk aversion $\{\tau_i\}_{i=1}^N$, entropy regularization $\{\epsilon_i\}_{i=1}^N$, learning rates $\eta_\theta, \eta_\phi, \eta_\psi$, clipping δ.
for iteration $k = 1, 2, \dots$ **do**
  for each agent $i \in \{1, \dots, N\}$ **do**
    // Data Collection
    Run joint policy $\boldsymbol{\pi}_{\theta_i,\phi_i} = (\pi_{\theta_i}, \pi_{\phi_i})$ of agent i and adversary i in the environment for T steps.
    Store trajectories $\mathcal{D} = \{(o_i^t, a_i^t, r_i^t)\}_{i=1,t=1}^{N,T}$.
    Compute advantage estimates $\hat{A}_i^t$ using local critic V_{ψ_i}.
    // Maximin-Optimization
    for step in steps **do**
      # Update Agent by maximizing $\mathcal{L}_i^{\mathrm{SRPO}}$
      $\theta_i \leftarrow \theta_i + \eta_\theta \nabla_{\theta_i} \mathcal{L}_i^{\mathrm{SRPO}}$
      # Update Adversary by maximizing $\bar{\mathcal{L}}_i^{\mathrm{SRPO}}$
      $\phi_i \leftarrow \phi_i + \eta_\phi \nabla_{\phi_i} \bar{\mathcal{L}}_i^{\mathrm{SRPO}}$
      # Update Local Critic
      $\psi_i \leftarrow \psi_i - \eta_\psi \nabla_{\psi_i} \mathcal{L}_i^{VF}(\psi_i)$
    end for
  end for
  Update old policy parameters: $\theta_{i,\text{old}} \leftarrow \theta_i$ for all i.
end for

Discussion The iteration structure of SRPO is very similar to that of IPPO. In the data collection phase, unlike IPPO, agent i computes the advantage estimate based on rollouts played against adversary i instead of other real agents. In the optimization phase, both agent i and their adversary update their policies to maximize the SRPO losses (30) and (31). Here, we have introduced an additional critic with parameter ψ_i with the MSE loss function $\mathcal{L}_i^{VF}(\theta_i, \theta_{-i})$, as a standard practice for variance reduction in IPPO, which is only used when computing the advantage estimates.

D.3 Policy Gradient Theorems

Policy gradient theorems are the most fundamental theoretical tool for nearly all policy-based algorithms. As the foundation of our framework, we provide the expression of policy gradients for the original agents and the adversaries under the risk-adjusted objectives, respectively. Although our results are the first to present these, similar results for single-agent version can be found in [26].

For the original agents, the policy gradient theorem has the following form:

Lemma D.1 (Policy Gradient for π_i)**.** *The gradient $\nabla_{\pi_i} V_i^{\pi,p}(s)$ can be written as*

$$\nabla_{\pi_i(\cdot|x)} V_i^{\pi,p}(s) = \frac{1}{1-\gamma} d_s^{\pi_i,p_i}(x) \left[\mathbb{E}_{\mathbf{a}_{-i} \sim p_i(\cdot|x)} [Q_i^{\pi,p}(x, \cdot, \mathbf{a}_{-i})] - \epsilon_i \nabla_{\pi_i} \nu_i(\pi_i; x) \right].$$

If π_i is parameterized by θ_i, we have

$$\nabla_{\theta_i} V_i^{\pi,p}(s) = \frac{1}{1-\gamma} \mathbb{E}_{x \sim d_s^{\pi_i,p_i}, \mathbf{a} \sim (\pi_i \otimes p_i)(\cdot|x)} \left[\nabla_{\theta_i} \log \pi_i(a_i|x) \left(Q_i^{\pi,p}(x, \mathbf{a}) - \epsilon_i \frac{\partial \nu_i(\pi_i; x)}{\partial \pi_i(a_i|x)} \right) \right],$$

where $\mathbf{a} = (a_i, \mathbf{a}_{-i})$.

Proof. We have

$$\begin{aligned}\nabla_{\pi_i(\cdot|x)} V_i^{\pi,p}(s) =& \langle \frac{\partial}{\partial \pi_i(\cdot|x)} Q_i^{\pi,p}(s,\cdot), (\pi_i \otimes p_i)(\cdot|s) \rangle \\ &+ \left(\mathbb{E}_{a_{-i}\sim p_i(\cdot|s)}[Q_i^{\pi,p}(s,\cdot,a_{-i})] - \epsilon_i \nabla_{\pi_i} \nu_i(\pi_i;s)\right) \mathbb{I}[s=x] \\ =& \gamma \mathbb{E}_{s'\sim P^{\pi_i,p_i}(s)}[\nabla_{\pi_i} V_i^{\pi,p}(s')] \\ &+ \left(\mathbb{E}_{a_{-i}\sim p_i(\cdot|s)}[Q_i^{\pi,p}(s,\cdot,a_{-i})] - \epsilon_i \nabla_{\pi_i} \nu_i(\pi_i;s)\right) \mathbb{I}[s=x] \\ =& \dots \\ =& \frac{1}{1-\gamma} d_s^{\pi_i,p_i}(x) \left[\mathbb{E}_{a_{-i}\sim p_i(\cdot|x)}[Q_i^{\pi,p}(x,\cdot,a_{-i})] - \epsilon_i \nabla_{\pi_i} \nu_i(\pi_i;x)\right].\end{aligned}$$

Similarly, for the parameterized case, we have

$$\begin{aligned}&\frac{\partial}{\partial \pi_i(a_i|x)} V_i^{\pi,p}(s) \\ &\quad= \left\langle \frac{\partial}{\partial \pi_i(a_i|x)} Q_i^{\pi,p}(s,\cdot), (\pi_i \otimes p_i)(\cdot|s) \right\rangle \\ &\qquad+ \left(\mathbb{E}_{a_{-i}\sim p_i(\cdot|s)}[Q_i^{\pi,p}(s,a_i,a_{-i})] - \epsilon_i \frac{\partial}{\partial \pi_i(a_i|x)} \nu_i(\pi_i;s)\right) \mathbb{I}[s=x] \\ &\quad= \gamma \mathbb{E}_{s'\sim P^{\pi_i,p_i}(s)}[\frac{\partial}{\partial \pi_i(a_i|x)} V_i^{\pi,p}(s')] \\ &\qquad+ \left(\mathbb{E}_{a_{-i}\sim p_i(\cdot|s)}[Q_i^{\pi,p}(s,a_i,a_{-i})] - \epsilon_i \frac{\partial}{\partial \pi_i(a_i|x)} \nu_i(\pi_i;s)\right) \mathbb{I}[s=x] \\ &\quad= \dots \\ &\quad= \frac{1}{1-\gamma} d_s^{\pi_i,p_i}(x) \left[\mathbb{E}_{a_{-i}\sim p_i(\cdot|x)}[Q_i^{\pi,p}(x,a_i,a_{-i})] - \epsilon_i \frac{\partial}{\partial \pi_i(a_i|x)} \nu_i(\pi_i;x)\right],\end{aligned}$$

therefore,

$$\begin{aligned}\nabla_{\theta_i} V_i^{\pi,p}(s) =& \sum_{x,a_i} \frac{\partial}{\partial \pi_i(a_i|x)} V_i^{\pi,p}(s) \nabla_{\theta_i} \pi_i(a_i|x) \\ =& \frac{1}{1-\gamma} \sum_{x,a_i} d_s^{\pi_i,p_i}(x) \left[\mathbb{E}_{a_{-i}\sim p_i(\cdot|x)}[Q_i^{\pi,p}(x,a_i,a_{-i})] - \epsilon_i \frac{\partial}{\partial \pi_i(a_i|x)} \nu_i(\pi_i;x)\right] \nabla_{\theta_i} \pi_i(a_i|x) \\ =& \frac{1}{1-\gamma} \mathbb{E}_{x\sim d_s^{\pi_i,p_i}} \mathbb{E}_{a\sim(\pi_i\otimes p_i)(\cdot|x)} \left[\nabla_{\theta_i} \log \pi_i(a_i|x) \left(Q_i^{\pi,p}(x,a) - \epsilon_i \frac{\partial}{\partial \pi_i(a_i|x)} \nu_i(\pi_i;x)\right)\right].\end{aligned}$$

This concludes the proof. □

Similarly, we have the policy gradient for p_i as follows:

Lemma D.2 (Policy gradient for p_i)**.** *The gradient $\nabla_{p_i} V_i^{\pi,p}(s)$ can be written as*

$$\nabla_{p_i(\cdot|x)} V_i^{\pi,p}(s) = \frac{1}{1-\gamma} d_s^{\pi_i,p_i}(x) \left[\mathbb{E}_{a_i\sim\pi_i(\cdot|x)}[Q_i^{\pi,p}(x,a_i,\cdot)] + \frac{1}{\tau_i} \nabla_{p_i} D_i(p_i,\pi_{-i};x)\right]$$

and if p_i is parameterized by $\bar{\theta}_i$, we have

$$\nabla_{\bar{\theta}_i} V_i^{\pi,p}(s) = \frac{1}{1-\gamma} \mathbb{E}_{x\sim d_s^{\pi_i,p_i}, \mathbf{a}\sim(\pi_i\otimes p_i)(\cdot|x)} \left[\nabla_{\bar{\theta}_i} \log p_i(\mathbf{a}_{-i}|x) \left(Q_i^{\pi,p}(x,\mathbf{a}) + \frac{1}{\tau_i} \frac{\partial D_i(p_i,\pi_{-i};x)}{\partial p_i(\mathbf{a}_{-i}|x)}\right)\right],$$

where $\mathbf{a} = (a_i, \mathbf{a}_{-i})$.

Proof. We have

$$\begin{aligned}
\nabla_{p_i(\cdot|x)} V_i^{\pi,p}(s) =& \nabla_{p_i(\cdot|x)}(\langle Q_i^{\pi,p}(s,\cdot),(\pi_i \otimes p_i)(\cdot|s)\rangle) + \frac{1}{\tau_i}\nabla_{p_i} D_i(p_i,\pi_{-i};s)\mathbb{I}[s=x] \\
=& \langle \frac{\partial}{\partial p_i(\cdot|x)} Q_i^{\pi,p}(s,\cdot),(\pi_i \otimes p_i)(\cdot|s)\rangle \\
&+ \left(\mathbb{E}_{a_i\sim\pi_i(\cdot|s)}[Q_i^{\pi,p}(s,a_i,\cdot)] + \frac{1}{\tau_i}\nabla_{p_i} D_i(p_i,\pi_{-i};s)\right)\mathbb{I}[s=x] \\
=& \gamma \mathbb{E}_{s'\sim P^{\pi_i,p_i}(s)}[\nabla_{p_i(\cdot|x)} V_i^{\pi,p}(s')] \\
&+ \left(\mathbb{E}_{a_i\sim\pi_i(\cdot|s)}[Q_i^{\pi,p}(s,a_i,\cdot)] + \frac{1}{\tau_i}\nabla_{p_i} D_i(p_i,\pi_{-i};s)\right)\mathbb{I}[s=x] \\
=& \dots \\
=& \frac{1}{1-\gamma} d_s^{\pi_i,p_i}(x)\left[\mathbb{E}_{a_i\sim\pi_i(\cdot|x)}[Q_i^{\pi,p}(x,a_i,\cdot)] + \frac{1}{\tau_i}\nabla_{p_i} D_i(p_i,\pi_{-i};x)\right],
\end{aligned}$$

and similarly for the parameterized case,

$$\begin{aligned}
&\nabla_{\bar{\theta}_i} V_i^{\pi,p}(s) \\
=& \frac{1}{1-\gamma}\mathbb{E}_{x\sim d_s^{\pi_i,p_i}}\mathbb{E}_{a\sim(\pi_i\otimes p_i)(\cdot|x)}\left[\nabla_{\bar{\theta}_i}\log p_i(a_{-i}|x)\left(Q_i^{\pi,p}(x,a) + \frac{1}{\tau_i}\frac{\partial}{\partial p_i(a_{-i}|x)} D_i(p_i,\pi_{-i};x)\right)\right].
\end{aligned}$$

This concludes the proof. □

D.4 Performance Difference Lemmas

Another crucial tool for quantifying the difference in value functions (hence expected returns) between different policies through the advantage function is the *performance difference lemma* (PDL). For our 4-player risk-adjusted game, we state the PDLs for original agents and adversaries below:

Lemma D.3 (Performance Difference Lemma for π_i)**.** *For two policies π_i and π_i' of player i, we have*

$$V_i^{\pi_i,z}(s) - V_i^{\pi_i',z}(s) = \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_i,p_i},\mathbf{a}\sim(\pi_i\otimes p_i)(\cdot|s')}\left[A_i^{\pi_i',z}(s',\mathbf{a}) - \epsilon_i\nu_i(\pi_i;s')\right] \tag{33}$$

where the advantage function A_i is defined as:

$$A_i^{\pi_i,z}(s,\mathbf{a}) = Q_i^{\pi_i,z}(s,\mathbf{a}) - V_i^{\pi_i,z}(s) + \frac{1}{\tau_i}D_i(z;s).$$

Here, z is the shorthand notation of $z = (\pi_{-i}, p_i)$.

Proof. We can write the difference in value function as

$$\begin{aligned}
&V_i^{\pi_i',z}(s) - V_i^{\pi_i,z}(s) \\
=& \mathbb{E}_{\pi_i',z,s_0=s}\left[\sum_{t=0}^{\infty}\gamma^t\left(r_i(s_t,\mathbf{a}_t) + \frac{1}{\tau_i}D_i\left(z;s_t\right) - \epsilon_i\nu_i(\pi_i';s_t)\right)\right] - V_i^{\pi_i,z}(s) \\
=& V_i^{\pi_i',z}(s) - \mathbb{E}_{(a_{i,0},\mathbf{a}_{-i,0})\sim(\pi_i\otimes p_i)(\cdot|s)}\left[r_i(s,\mathbf{a}_0) + \frac{1}{\tau_i}D_i(z;s) - \epsilon_i\nu_i(\pi_i;s) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,\mathbf{a}_0)}V_i^{\pi_i',z}(s')\right] \\
&+ \mathbb{E}_{(a_{i,0},\mathbf{a}_{-i,0})\sim(\pi_i\otimes p_i)(\cdot|s)}\left[r_i(s,\mathbf{a}_0) + \frac{1}{\tau_i}D_i(z;s) - \epsilon_i\nu_i(\pi_i;s) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,\mathbf{a}_0)}V_i^{\pi_i',z}(s')\right] - V_i^{\pi_i,z}(s) \\
=& \mathbb{E}_{(a_{i,0},\mathbf{a}_{-i,0})\sim(\pi_i\otimes p_i)(\cdot|s)}\left[V_i^{\pi_i',z}(s) - Q_i^{\pi_i',z}(s,\mathbf{a}_0) - \frac{1}{\tau_i}D_i(z;s) + \epsilon_i\nu_i(\pi_i;s)\right] \\
&+ \gamma\mathbb{E}_{(a_{i,0},\mathbf{a}_{-i,0})\sim(\pi_i\otimes p_i)(\cdot|s)}\mathbb{E}_{s'\sim P(s,\mathbf{a}_0)}\left[V_i^{\pi_i',z}(s') - V_i^{\pi_i,z}(s')\right] \\
=& \cdots \\
=& \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_i,p_i}}\mathbb{E}_{\mathbf{a}\sim(\pi_i\otimes p_i)(\cdot|s')}\left[V_i^{\pi_i',z}(s') - Q_i^{\pi_i',z}(s',\mathbf{a}) - \frac{1}{\tau_i}D_i(z;s') + \epsilon_i\nu_i(\pi_i;s')\right]
\end{aligned}$$

This concludes the proof. □

Lemma D.4 (Performance Difference Lemma for p_i)**.** *For two policies p_i and p_i' of adversary i, we have:*

$$V_i^{p_i,\pi}(s) - V_i^{p_i',\pi}(s) = \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_i,p_i},\mathbf{a}\sim(\pi_i\otimes p_i)(\cdot|s')}\left[A_i^{p_i',\pi}(s',\mathbf{a}) + \frac{1}{\tau_i}D_i(p_i,\pi_{-i};s')\right]$$

where

$$A_i^{p_i,\pi}(s,\mathbf{a}) = Q_i^{p_i,\pi}(s,\mathbf{a}) - V_i^{p_i,\pi}(s) - \epsilon_i\nu_i(\pi_i;s).$$

Proof. We can write the value function difference as

$$\begin{aligned}
&V_i^{p_i,\pi}(s) - V_i^{p_i',\pi}(s)\\
=&V_i^{p_i,\pi}(s) - \mathbb{E}_{(a_{i,0},\mathbf{a}_{-i,0})\sim(\pi_i\otimes p_i)(\cdot|s)}\left[r_i(s,\mathbf{a}_0) + \frac{1}{\tau_i}D_i(p_i,\pi_{-i};s) - \epsilon_i\nu_i(\pi_i;s) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,\mathbf{a}_0)}V_i^{p_i',\pi}(s')\right]\\
&+\mathbb{E}_{(a_{i,0},\mathbf{a}_{-i,0})\sim(\pi_i\otimes p_i)(\cdot|s)}\left[r_i(s,\mathbf{a}_0) + \frac{1}{\tau_i}D_i(p_i,\pi_{-i};s) - \epsilon_i\nu_i(\pi_i;s) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,\mathbf{a}_0)}V_i^{p_i',\pi}(s')\right] - V_i^{p_i',\pi}(s)\\
=&\gamma\mathbb{E}_{(a_{i,0},\mathbf{a}_{-i,0})\sim(\pi_i\otimes p_i)(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,\mathbf{a}_0)}[V_i^{p_i,\pi}(s') - V_i^{p_i',\pi}(s')]\\
&+\mathbb{E}_{(a_{i,0},\mathbf{a}_{-i,0})\sim(\pi_i\otimes p_i)(\cdot|s)}\left[Q_i^{p_i',\pi}(s,\mathbf{a}_0) + \frac{1}{\tau_i}D_i(p_i,\pi_{-i};s) - \epsilon_i\nu_i(\pi_i;s) - V_i^{p_i',\pi}(s)\right]\\
=&\dots\\
=&\frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{\pi_i,p_i}}\mathbb{E}_{\mathbf{a}\sim(\pi_i\otimes p_i)(\cdot|s')}\left[Q_i^{p_i',\pi}(s',\mathbf{a}) + \frac{1}{\tau_i}D_i(p_i,\pi_{-i};s') - \epsilon_i\nu_i(\pi_i;s') - V_i^{p_i',\pi}(s')\right].
\end{aligned}$$

This concludes the proof. □

D.5 Motivating the SRPO Objective

Having stated the policy gradient theorems and PDLs, we now motivate our SRPO objective following the same rationale as that in TRPO [50]. Let $z = (\pi^0, p^0)$ be the joint policy at the current timestep, by Lemma D.3, consider policy update for π_i, we can write the value function for any policy π_i as

$$J_i^{\pi_i,\pi_{-i}^0,p_i^0} = J_i^{\pi^0,p^0} + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_i,p_i^0},\mathbf{a}\sim(\pi_i\otimes p_i^0)(\cdot|s)}\left[A_i^{\pi_i^0,\pi_{-i}^0,p_i^0}(s,\mathbf{a}) - \epsilon_i\nu_i(\pi_i;s)\right],$$

where d^{π_i,p_i^0} is the state distribution for the initial state distribution ρ_0. When π_i is constrained to be close to π_i^0, we use the following surrogate objective first proposed in TRPO [50] to approximate the expectation term by replacing the sample distribution from d^{π_i,p_i^0} to $d^{\pi_i^0,p_i^0}$:

$$\mathbb{E}_{s\sim d^{\pi_i^0,p_i^0},\mathbf{a}\sim(\pi_i^0\otimes p_i^0)(\cdot|s)}\left[\frac{\pi_i(a_i|s)}{\pi_i^0(a_i|s)}A_i^{\pi^0,p_i^0}(s,\mathbf{a}) - \epsilon_i\nu_i(\pi_i;s)\right], \tag{34}$$

where the approximation $d^{\pi_i^0,p_i^0} \approx d^{\pi_i,p_i^0}$ holds when π^0 stays close to π, and importance sampling through the ratio $\frac{\pi_i(a_i|s)}{\pi_i^0(a_i|s)}$ on a_i. Similarly by Lemma D.4, we have the following surrogate objective for p_i as follows:

$$\mathbb{E}_{s\sim d^{\pi_i^0,p_i^0},\mathbf{a}\sim(\pi_i^0\otimes p_i^0)(\cdot|s)}\left[\frac{p_i(\mathbf{a}_{-i}|s)}{p_i^0(\mathbf{a}_{-i}|s)}A_i^{p_i^0,\pi^0}(s,\mathbf{a}) + \frac{1}{\tau_i}D_i(p_i,\pi_{-i};s)\right]. \tag{35}$$

The original TRPO objective suggests optimizing (34) and (35) subject to the constraints

$$\mathbb{E}_{d^{\pi_i^0,p_i^0}}[\mathrm{KL}(\pi_i(\cdot|s)\|\pi_i^0(\cdot|s))] \le \delta,$$

and

$$\mathbb{E}_{d^{\pi_i^0,p_i^0}}[\mathrm{KL}(p_i(\cdot|s)\|p_i^0(\cdot|s))] \le \delta,$$

respectively. A later adaptation PPO [51] replaces these hard constraints with the clipped surrogate objective. Following this adaptation and replacing the policy to condition on observations instead of the global states, we obtain our SRPO losses (30) and (31) respectively.



Figure 6: Overcooked Gridworld.



Figure 7: Tag.

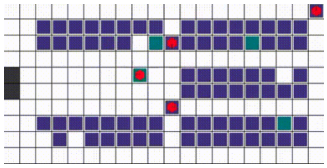


Figure 8: RWARE.

E Experimental Details

E.1 Overall Setup

Across all experiments, we compare SRPO against IPPO under a unified training and evaluation protocol. Each method is trained from scratch using multiple random seeds, and performance is reported as episodic return averaged over rollouts.

Training. Agents are trained for a fixed interaction budget per environment, with periodic evaluation during training. Unless otherwise specified, SRPO and IPPO agents share the same architecture, optimizer settings, amount of interaction with the environment, and entropy regularization, differing only in the inclusion of the strategic risk-averse objective. Given the same total amount of interactions, SRPO takes roughly twice the training time as IPPO.

Cross-play evaluation. To assess unseen partner generalization, we adopt a **cross-play** protocol. For each environment, we collect a population of independently trained agents and evaluate all pairwise combinations without further learning. Cross-play performance is reported as a matrix, where each entry corresponds to the average episodic return of a fixed agent pair evaluated over multiple episodes. This protocol measures zero-shot coordination with previously unseen partners.

E.2 Overcooked Gridworld

Environments. We implemented a revised version of Overcooked AI, shown in Figure 6. The task is defined on a 5×5 grid with two agents acting simultaneously. Agents start from one of two symmetric initial configurations in the top left and must cooperatively pick up onions from two fixed sources and deliver them to a central pot. Each agent selects from five discrete actions (up, down, left, right, stay); non-stay actions incur a private movement cost of 0.2. Agent updates are applied sequentially in random order each step, with collisions blocked and penalized by 2.0. An agent can carry at most one onion; stepping onto an available onion grants a shared reward of +1 and removes the onion, while delivering an onion by attempting to move into the (solid) pot yields a shared reward of +10. Each onion independently respawns with probability 0.2 per step. Observations are fully shared and consist of both agents' positions, carry status, and onion availability. In this setup, rewards combine private movement penalties with shared team rewards, inducing cooperative behavior under individual costs.

Training and evaluation. Each agent is trained for approximately 2×10^6 environment interactions. Both the actor and the critic are two-layer MLPs with 64 hidden units and $\tanh$ activations. Training statistics are recorded every 10240 interactions, where each evaluation point reports the average return over 5 rollouts of length 128, and each update performs 4 PPO epochs over the full batch with the clipped surrogate objective ($\epsilon_{\text{clip}} = 0.2$), GAE ($\gamma = 0.99$ and $\lambda = 0.95$), and Adam at learning rate 6×10^{-4}. We perform 30 independent runs for each method. IPPO agents use an entropy coefficient of $\epsilon = 0.1$, while SRPO agents use $\tau = 10$ with the same entropy coefficient.

For the FCP baseline [57], we first train $N = 10$ self-play seeds with a single shared actor-critic controlling both seats, snapshotting $K = 3$ checkpoints per seed (early, mid, and final), yielding a frozen partner pool of $N \times K = 30$ partners of varying skill levels. We then train an FCP best-response policy with PPO against this pool, where at each episode every parallel environment independently samples a partner index and an FCP seat assignment; only the FCP policy receives gradient updates. To keep the FCP network's input distribution consistent across seat assignments, we apply a slot-swap to the observation when FCP occupies seat 1 so that its own seat features always appear first, and we mirror this swap at cross-play evaluation time.

For the main experiments, we construct a 60×60 cross-play matrix consisting of 30 IPPO agents and 30 SRPO agents. Each entry is averaged over 10 evaluation episodes of length 128. We report cross-play results for both the training environment and the held-out test environment. The results are shown in Figure 2, and comparison with the FCP baseline is reported in Figure 9.

On the FCP baseline. Although FCP is a natural reference point for partner generalization, we find it both computationally expensive and empirically weaker than SRPO on this task, as shown in Figure 9. In terms of scalability, FCP requires a two-stage pipeline: N self-play runs must complete before any best-response training can begin, and the FCP run itself consumes the same 2×10^6 interaction budget as a single IPPO/SRPO seed. In terms of cross-play return, FCP underperforms SRPO primarily because the partner pool is generated by self-play, which tends to converge to similar policies in this gridworld. As a result, the FCP best response effectively learns to free-ride rather than developing robust coordination behavior.

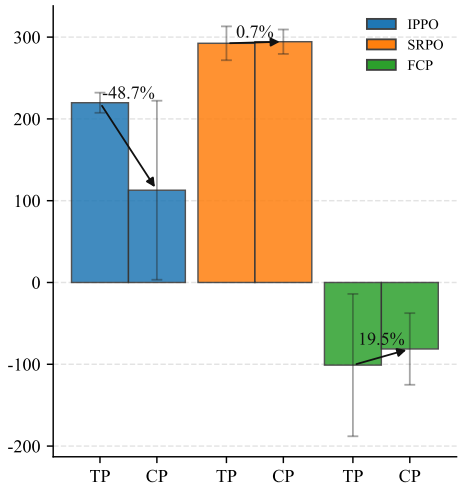


Figure 9: Difference between training performance (TP) and cross-play performance (CP), reported as mean and standard deviation in Overcooked Gridworld. IPPO performance drops sharply in cross-play, with lower average return and larger variance, while SRPO remains stable. FCP also achieves relatively low returns when paired with itself or with FCP agents from other seeds, suggesting that the self-play partner pool induces free-riding rather than robust coordination.

E.3 Tag

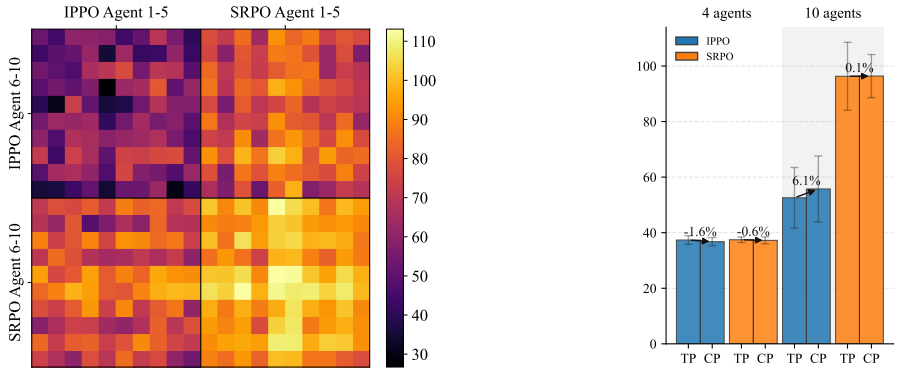
Environment. We evaluate on the PettingZoo Multi-Agent Particle Environment (MPE) simple_tag_v3 [59], configured with 1 runner (prey), 2 chasers, and 2 static obstacles, using discrete actions and a horizon of 100 cycles per episode. The environment is shown in Figure 7. Rewards are taken directly from the environment and scaled by 0.1, and an episode terminates when any agent is terminated or truncated.

Training and evaluation. Each agent is trained for approximately 3×10^7 environment interactions. Both the actor and the critic are two-layer MLPs with 128 hidden units and ReLU activations. Training performance is recorded every 40960 interactions, where each evaluation point reports the average episodic return over 64 rollouts of length 100, and each update performs 4 PPO epochs over the full batch with the clipped surrogate objective ($\epsilon_{\text{clip}} = 0.2$), GAE ($\gamma = 0.99$ and $\lambda = 0.95$), and Adam at learning rate 10^{-3}. We perform 30 independent runs for each method. IPPO agents use an entropy coefficient of $\epsilon = 0.01$, while SRPO agents use the risk aversion parameter $\tau = 10$ with the same entropy coefficient. To evaluate partner generalization, we construct a 60×60 cross-play matrix consisting of 30 IPPO agents and 30 SRPO agents. Each matrix entry corresponds to the average return obtained by a fixed agent pair, evaluated over 100 episodes of length 100 without further learning. In addition to cross-play on the training environment, we evaluate all agent pairs in a held-out test environment with a different runner configuration and report these results separately. The results are shown in Figure 3.

E.4 Multi-Robot Warehouse

Environment. We evaluate on the Multi-Robot Warehouse environment [43], configured as `rware-tiny-4ag-easy-v2` with 4 agents, discrete actions, and 71-dimensional partial observations. The environment is shown in Figure 8. Rewards are taken directly from the environment, augmented with a small pickup-shaping bonus of 0.01 awarded the first step an agent grasps a requested shelf, and an episode terminates only at the environment-imposed horizon.

Training and evaluation. Each agent is trained for 2×10^7 environment interactions. Both the actor and the critic are two-layer MLPs with 64 hidden units and $\tanh$ activations. Rollouts use 16 parallel environments of length 500, and each update performs 4 PPO epochs over minibatches of size 2048 with the clipped surrogate objective ($\epsilon_{\text{clip}} = 0.2$), GAE ($\gamma = 0.99$ and $\lambda = 0.95$), and Adam at learning rate 5×10^{-4}. Both IPPO and SRPO use parameter sharing across agents and an entropy coefficient of $\epsilon = 0.05$. We perform 10 independent runs for each method. To evaluate partner generalization, we construct a 20×20 cross-play matrix between the 10 IPPO and 10 SRPO agents, where each entry reports the average return obtained by a fixed agent pair evaluated over rollouts without further learning. The results are shown in Figure 10.



(a) RWare 10 Agents.

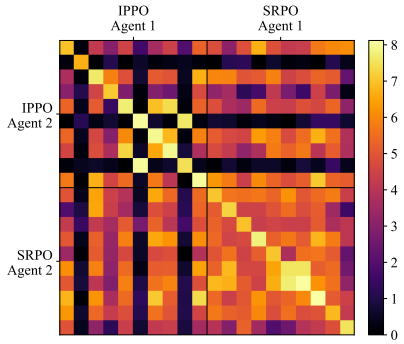
(b) TP-CP gaps.

Figure 10: RWARE cross-play performances of SRPO ($\epsilon = 0.05, \tau = 20$) and IPPO ($\epsilon = 0.05$) in the ten-agent setting. SRPO does not sacrifice performance in the training (4 agents TP), but achieve much better CP and TP performances in the ten-agent cases.

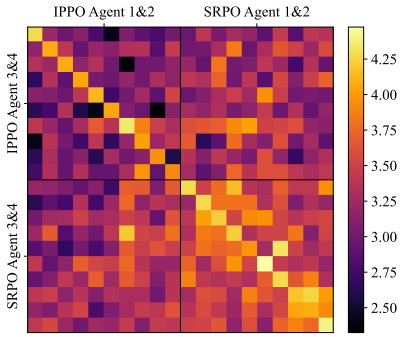
E.5 Hanabi

Environments. Hanabi is a cooperative card game where players act as distracted pyrotechnicians who must work together to launch a spectacular fireworks display by playing cards in the correct sequence, with the twist that you can see everyone's cards except your own. As such, Hanabi is a canonical benchmark for collaboration. At each turn, the active player chooses from moves such as playing a card to the shared firework piles, discarding a card to regain an information token, or hinting another player about the color or rank of cards in their hand; hints consume a finite pool of information tokens, and illegal plays consume life tokens. Following Lauffer et al. [30], we consider a version with 3 colors and 3 ranks. We conduct our experiments in both the 2 player and the 4 player settings.

Training and evaluation. Each run uses a single GPU with 1 training thread and 1000 rollout threads, collecting trajectories of episode length 100. We train for 3×10^7 environment steps with PPO updates using 15 epochs and 1 minibatch per update. We set the learning rates to 7×10^{-4} (actor) and 1×10^{-3} (critic), and set the initialization gain to 0.01. We set the entropy regularization to 0.001 and the risk aversion parameter $\tau = 0.01$. In the 2-player setting, policies use a 2-layer MLP with hidden size 128. In the 4-player setting, policies use a 2-layer MLP with hidden size 256, and each agent is randomly set to be adversary agent during each rollout for SRPO. For each setting, we construct a 20×20 cross-play matrix consisting of 10 IPPO agents and 10 SRPO agents, trained independently with different seeds. Each entry is averaged over 50 evaluation episodes of length 100. The results are shown in Figures 11 and 12.

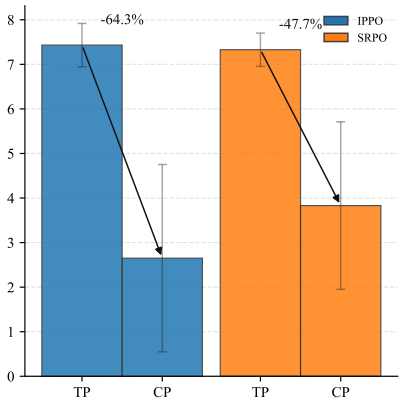


(a) Hanabi 2-player.

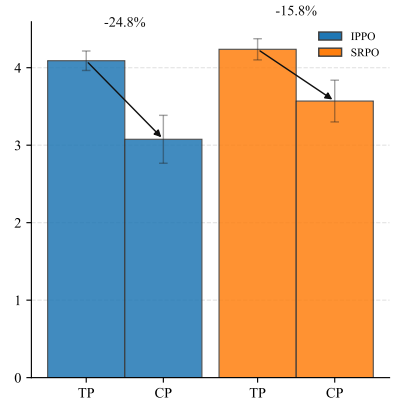


(b) Hanabi 4-player.

Figure 11: Cross-play performances of SRPO ($\epsilon = 0.001, \tau = 0.01$) and IPPO ($\epsilon = 0.001$) in both 2-player and 4-player Hanabi games. Each square represents the average reward of two agents across 50 runs of length 100.



(a) Hanabi 2-player.



(b) Hanabi 4-player.

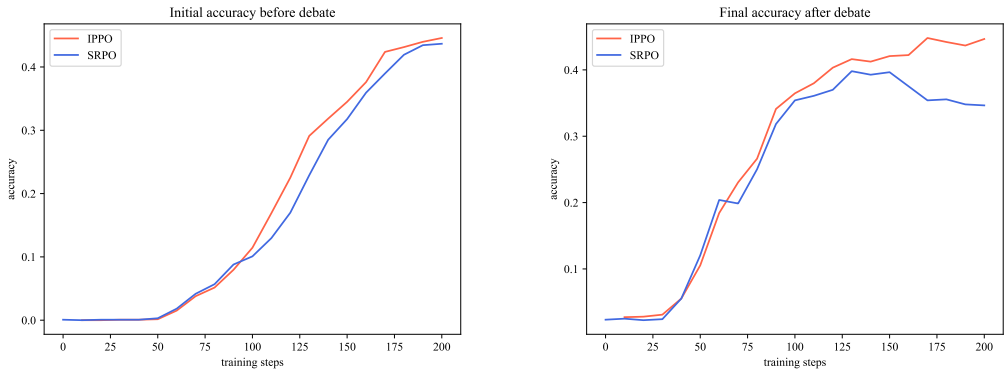
Figure 12: Performance change (mean and standard deviation) between Training Performance (TP) and Cross-play Performance (CP): the performance of IPPO drastically decreases, with lower average and larger standard deviation in cross-play, while the performance of SRPO drops less severely.

E.6 Multi-agent debate on GSM8k

Environment. In this setup, two agents engage in a three-round iterative debate protocol. In the first round, each agent independently observes the question and produces its own reasoning and answer. In subsequent rounds, each agent observes the original question as well as both agents' outputs from the previous round, and then refines its response. Successfully solving a problem therefore requires more than producing a correct answer in isolation: each agent must reinforce correct reasoning while remaining robust to potentially misleading or incorrect proposals from its teammate. This naturally induces a cooperative yet adversarial interaction, where robustness to the partner's policy plays a critical role in achieving reliable joint performance.

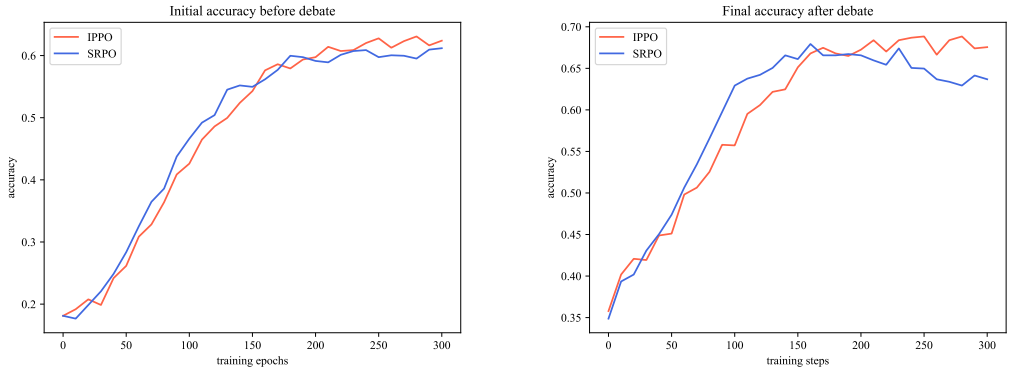
Training and evaluation. Both IPPO (the existing state-of-the-art) and SRPO agents are trained using multiple base language models, including Qwen2.5-0.5B-Instruct (Q0.5B) and Qwen2.5-3B-Instruct (Q3B) [68], as well as Qwen3-0.6B (Q0.6B) and Qwen3-4B-Instruct-2507 (Q4B) [69], using the verl training framework [53]. Here, we set the temperature $T = 1$ for training, the entropy coefficient to be $\epsilon = 0$ for both IPPO and SRPO, and the risk degree $\tau = 10$ for SRPO. Qwen2.5-0.5B-Instruct is trained with 200 steps on two NVIDIA H100s, which takes around 8 hours. Qwen3-0.6B is trained with 300 steps on 2 h100s, which takes around 12 hours. Qwen2.5-3B-Instruct is trained with 200 steps on four NVIDIA H100s, which takes around 12 hours. Qwen3-4B-Instruct-2507 is trained with 350 steps on four NVIDIA H100s, which takes around 24 hours. Every 10 steps, we evaluate the performances on a held-out validation set using the same debate setup. The training curves are shown in Figures 13 to 16. Due to the cost of LLM RL training, we report a single training run per

setting. For evaluation, we set the temperature to 0.1 and evaluate on the validation set for 5 runs, and the results are reported in Table 1. Compared with IPPO, SRPO primarily affects coordination rather than individual reasoning. During training, SRPO achieves slightly lower final debate accuracy than IPPO because it interacts with an adversarial partner, whereas IPPO does not. During evaluation, however, SRPO consistently achieves more robust cross-play performance across model sizes, as shown in Table 1.



(a) Initial accuracy before debate. (b) Final accuracy after debate.

Figure 13: The training curve for SRPO and IPPO with Qwen2.5-0.5B-Instruct. We point out that SRPO is trained against an adversary—meaning lower training reward (post debate) is expected as the adversary learns to mislead the agent to minimize reward.



(a) Initial accuracy before debate. (b) Final accuracy after debate.

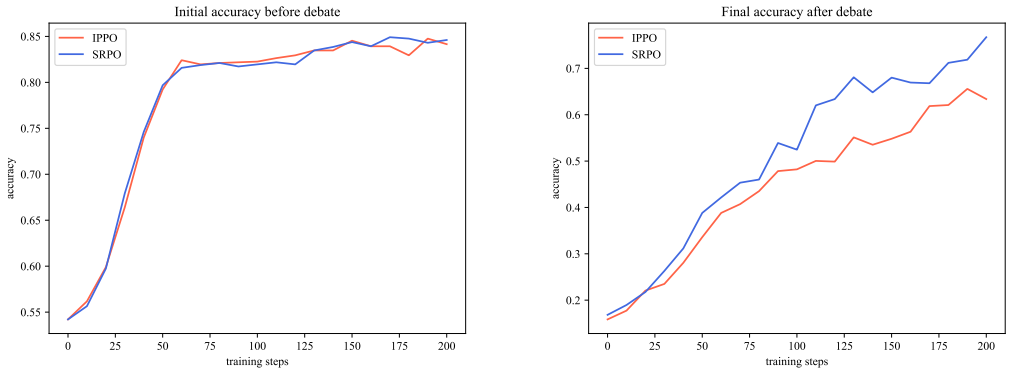
Figure 14: The training curve for SRPO and IPPO with Qwen3-0.6B. We point out that SRPO is trained against an adversary—meaning lower training reward (post debate) is expected as the adversary learns to mislead the agent to minimize reward.

E.7 Ablation study

We conduct ablation studies on both the risk aversion parameter τ and the entropy coefficient ϵ in the Overcooked and Tag environments, with additional ablation studies on τ in RWARE. The results are shown in Figures 17 and 18. While prior work [17] demonstrates that tuning the entropy coefficient ϵ can improve cross-play performance of IPPO in the Hanabi environment, we find that this strategy does not generalize to Overcooked or Tag. In these environments, entropy tuning alone fails to yield robust partner generalization, suggesting that entropy regularization is not a principled mechanism for addressing coordination under partner shifts.

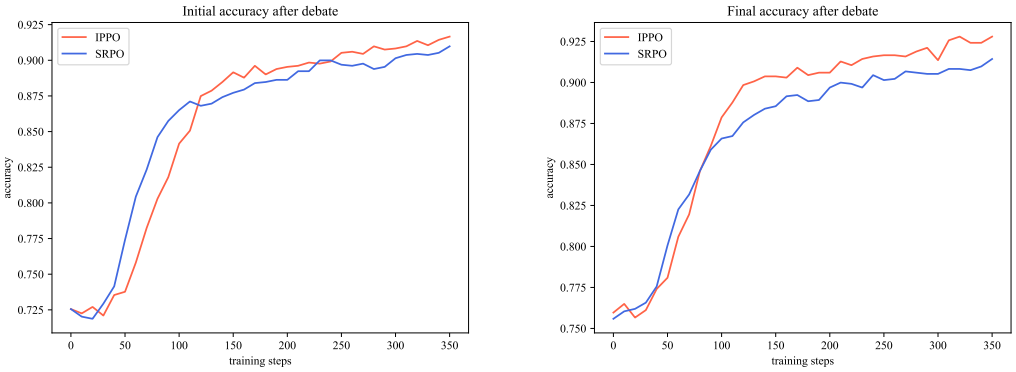
F Broader Impact

This work advances collaborative MARL by training agents that generalize to unseen partners and contribute proportionally to shared objectives, with potential benefits for human–AI teaming and



(a) Initial accuracy before debate. (b) Final accuracy after debate.

Figure 15: The training curve for SRPO and IPPO with Qwen2.5-3B-Instruct. We point out that SRPO is trained against an adversary—meaning lower training reward (post debate) is expected as the adversary learns to mislead the agent to minimize reward.



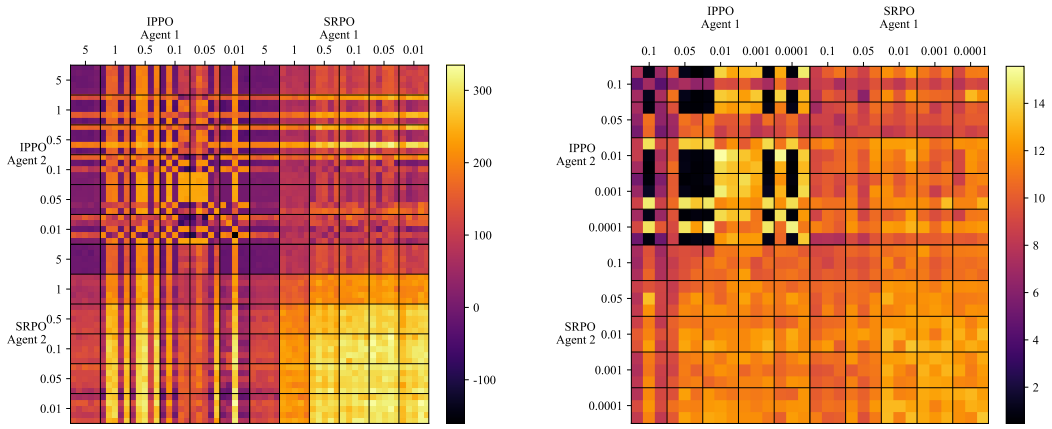
(a) Initial accuracy before debate. (b) Final accuracy after debate.

Figure 16: The training curve for SRPO and IPPO with Qwen3-4B-Instruct-2507. We point out that SRPO is trained against an adversary—meaning lower training reward (post debate) is expected as the adversary learns to mislead the agent to minimize reward.

multi-agent foundation model systems. We do not release any new high-risk artifacts and do not foresee direct negative societal impacts beyond those generic to advances in MARL.

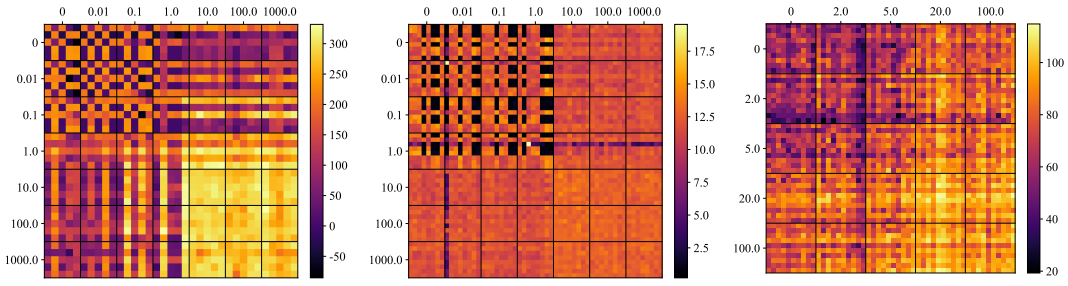
G Limitations

SRPO maintains an additional adversary policy ϕ_i per learning agent, which roughly doubles the per-iteration cost relative to IPPO in the fully decentralized form. However, this overhead can be eliminated when agents are exchangeable by sharing a single agent and a single adversary policy across players (as in our RWARE and large-population runs). The GSM8K multi-agent debate experiment (Section E.6) uses a single training seed per (model, algorithm) configuration due to the cost of fine-tuning Qwen-models, so its reported confidence intervals reflect only evaluation-rollout variability; we present it as a preliminary result complementing the multi-seed evidence (10–30 seeds) from Overcooked, Tag, RWARE, and Hanabi. Finally, our equilibrium-improvement and free-riding theorems apply respectively and provably to continuous quadratic aggregative games and to finite-action collaborative games with private costs; outside this scope (e.g., Tag) SRPO still improves cross-play robustness, but we do not claim improvements in training reward.



(a) Overcooked. (b) Tag.

Figure 17: Ablation study on the entropy parameter ϵ.



(a) Overcooked. (b) Tag. (c) RWARE (10 agents).

Figure 18: Ablation study on the degree of risk aversion parameter τ; $\tau = 0$ refers to the risk-neutral case, i.e., IPPO. SRPO is robust to choices of τ and achieves consistent gains over IPPO across orders of magnitude of τ.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The four contributions stated in the introduction (incentivizing collaboration, alleviating free-riding, the SRPO algorithm, and empirical validation) correspond directly to Theorem 4.1, Theorem 4.4, Section 5, and Section 6, respectively.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated Limitations section is provided in Section G.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Each theorem statement explicitly lists its assumptions, and complete proofs are provided in the appendices (Sections B.1, B.2, C and D). Proof sketches and discussion of the assumptions appear alongside the formal statements in the main text.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The SRPO algorithm is fully specified in Section 5, and Section E provides the environment configurations, network architectures, optimizer settings, hyperparameters (including τ, ϵ, learning rates, batch sizes, training horizons), number of seeds, and evaluation protocols for all four environments. We also release the code.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the public code repository (linked in Section 6) containing implementations of SRPO and IPPO for all four environments along with configuration files. All datasets and environments used (Overcooked Gridworld, PettingZoo MPE, Hanabi, GSM8K) are publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section E reports all training details for each environment, including optimizer (Adam), learning rates, PPO clipping and GAE parameters, entropy and risk-aversion coefficients, network sizes, number of seeds, interaction budgets, and the cross-play evaluation protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For the Overcooked, Tag, RWARE, and Hanabi experiments, all cross-play and training-vs-cross-play summary plots report mean and standard deviation across multiple independent training seeds (10–30 per setting) and multiple evaluation episodes per pair;

the variability captured (random seed plus stochastic environment rollouts) is described in Section E. The LLM debate experiment (Section E.6) is run with a single training seed per (model, algorithm) configuration due to compute constraints, and the reported confidence intervals therefore reflect only evaluation-rollout variability.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section E reports the interaction budget per run for each environment and the GPU type and wall-clock time for the LLM debate experiments. SRPO's additional compute cost relative to IPPO (an extra adversary update per iteration, mitigated by parameter sharing in exchangeable-agent settings) is discussed in Section G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is theoretical and methodological, evaluated on standard public benchmarks and datasets, with no human-subject data or sensitive content. Anonymity is preserved in the submitted code repository during review.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss societal impact in the impact statement: improved partner generalization in collaborative MARL (including human–AI teams and multi-agent LLM systems) is likely to be net-positive. The paper introduces no new datasets or generative models that pose immediate misuse risks.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any pre-trained generative model, scraped dataset, or other artifact with a high risk of misuse. We release only training and evaluation code for cooperative MARL benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets used (Overcooked Gridworld, PettingZoo MPE, Hanabi Learning Environment, GSM8K, Qwen2.5/Qwen3, Llama-3.2, the verl framework) are cited via their original publications. These assets are released under permissive open-source licenses (MIT/Apache 2.0/Llama Community License) and are used in accordance with their terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release the code repository implementing SRPO with documentation, configuration files, and instructions to reproduce the experiments in each of the four environments.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects, so IRB review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs (Qwen2.5, Qwen3, and Llama-3.2) are used as policies in the multi-agent debate experiment on GSM8K (Section E.6); their role, the training framework (verl), the prompting/debate protocol, and the hyperparameters used are described there. Otherwise, LLMs are not part of the core methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.