Multi-Agent Matrix Games with Individual learners: How Exploration-Exploitation Strategies Impact the Emergence of Coordination

Julien Armand[1,2,3], **Tommy C. H. Lin**[1,2], **Maxime Heuillet**[1,2,3], **Audrey Durand**[1,2,3,4]

{julien.armand.1, tommy-chien-hsuan.lin.1, maxime.heuillet.1}@ulaval.ca, audrey.durand@ift.ulaval.ca

[1]**Département d'informatique et de génie logiciel, Université Laval, Canada**
[2]**Institut Intelligence et Données (IID), Université Laval, Canada**
[3]**Mila – Quebec AI Institute**
[4]**CIFAR AI Chair**

Abstract

Coordination between independent learning agents in a multi-agent environment is an important problem where AI systems may impact each others learning process. In this paper, we study how individual agents converge to optimal equilibrium in multi-agent where coordination is necessary to achieve optimality. Specifically, we cover the case of coordination to maximize every individual payoffs and coordination to maximize the collective payoff (cooperation). We study the emergence of such coordination behaviours in two-player matrix games with unknown payoff matrices and noisy bandit feedback. We consider four different environments along with widely used deterministic and stochastic bandit strategies. We study how different learning strategies and observation noise influence convergence to the optimal equilibrium. Our results indicate that coordination often emerge more easily from interactions between deterministic agents, especially when they follow the same learning behaviour. However, stochastic learning strategies appear to be more robust in the presence of many optimal joint actions. Overall, noisy observations often help stabilizing learning behaviours.

1 Introduction

Coordinating independent agents in multi-agent systems is a central problem in reinforcement learning (Wei & Luke, 2016). As autonomous agents are increasingly deployed to interactively learn from their environment, they may not be designed to reason over the presence of other learning agents in the environment. Therefore, to achieve optimal outcomes, the agents may have to learn to coordinate implicitly. Understanding implicit coordination has broader impacts into frontier applications such as autonomous fleets of cars, drones, or robots (Broecker et al., 2018; Toghi et al., 2021).

We consider the setting where independent learning agents (also referred to as *learners*) are unable to explicitly observe the actions and the outcomes of other agents (Claus & Boutilier, 1998). Such independent learners can treat each other as part of the environment. However, in this setting, the evolution of agents over time (through learning) translates into *non-stationary noise* on the outcomes observed by each agent (Laurent et al., 2011). Moreover, discovering the optimal joint action requires a coordinated exploration, while maintaining this optimal behaviour requires a coordinated exploitation. Independent learners are therefore at risk of facing the *alter-exploration* problem (Laurent et al., 2011), where they enter a vicious circle of uncoordinated exploration-exploitation preventing them for identifying the optimal joint action. On top of that, interactions between agent

actions can hinder convergence to a common, coordinated, equilibrium (Matignon et al., 2012). Understanding how the exploration-exploitation mechanisms underlying different learning strategies impact the emergence of coordinated behaviours therefore remains an open research question.

We study the emergence of coordination between independent learning agents under repeated matrix games. More specifically, we focus on two-player games, where the expected outcome of each player is described using a matrix on the joint actions (Robinson & Goforth, 2006). The typical setting assumes that the matrix is known to the players (Osborne & Rubinstein, 1994). It has been generalized to unknown matrices with bandit feedback in zero-sum games (O'Donoghue et al., 2021), where players only observe each others actions and a noisy payoff. However, due to the known zero-sum dynamics, players can deduce the reward obtained by the other player. As this is therefore not compatible with the definition of individual learners, we consider *truly unknown* matrices.

We also consider the traditional bandit feedback (Lattimore & Szepesvári, 2020) where players observe only their own rewards. Our work complements the very few prior results in this setting focused essentially on cooperation (Douglas et al., 2024), that is coordination on a collective objective, with noise-free observations.

We study interplays between exploration-exploitation strategies using state-of-the-art bandit methods in coordination and cooperation games, in the presence of specific challenges induced by the structure of payoff matrices and noisy observations.

2 Problem setting

We consider two-player repeated games characterized by *unknown* payoff matrices $R^{(1)}, R^{(2)} \in \mathbb{R}^{k\times m}$. On each round $t = 1, 2, \dots, T$ (with *unknown* horizon T), player 1 (*row* player) selects action $i_t \in \{1, \dots, k\}$ and player 2 (*column* player) selects action $j_t \in \{1, \dots, m\}$. Let $\mathcal{A} := \{1, \dots, k\} \times \{1, \dots, m\}$ denote the set of *joint actions*. The joint action $a_t = (i_t, j_t)$ is played and rewards are generated for both players[1]:

$$r_t^{(1)} = R_{a_t}^{(1)} + \eta_t^{(1)} \qquad r_t^{(2)} = R_{a_t}^{(2)} + \eta_t^{(2)},$$

where $\eta_t^{(1)}$ and $\eta_t^{(2)}$ are zero-mean noises, independent and identically distributed from a known distribution across time. Both players can only observe their own reward and not the actions of the other player. This is known as bandit feedback (Lattimore & Szepesvári, 2020).

The two-player zero-sum matrix games (O'Donoghue et al., 2021) correspond to a specific configuration of this setting where $R^{(1)} = -R^{(2)}$ and this information is known to the players. Therefore, each player can learn the motivation of the other player by learning their own payoff matrix. In our general setting, *the relationship between $R^{(1)}$ and $R^{(2)}$, if any, remains unknown to the players*.

Coordination games We say that coordination is required when the payoff matrices are such that the optimal joint action for both players is the same, that is $a_\star := \arg\max_{a\in\mathcal{A}} R^{(1)} = \arg\max_{a\in\mathcal{A}} R^{(2)}$. In this case, the optimal joint policy allows both players to maximize their individual profit. Note that all games where $R^{(1)} = R^{(2)}$ are coordination games by default. However, coordination can also involve payoff, that is $R^{(1)} \neq R^{(2)}$. In this case, players have different *motivations*, but their motivations are well-aligned such that the optimal joint action corresponds to each player maximizing their individual outcome *simultaneously*.

The performance of learning agents in a coordination game is evaluated using the *cumulative regret*:

$$\mathcal{R}^{(1)}(T) := \sum_{t=1}^{T} \left(R_{a_\star}^{(1)} - \mathbb{E}[R_{a_t}^{(1)}] \right) \qquad \mathcal{R}^{(2)}(T) := \sum_{t=1}^{T} \left(R_{a_\star}^{(2)} - \mathbb{E}[R_{a_t}^{(2)}] \right), \tag{1}$$

[1] Given a matrix M, $M_{(i,j)}$ denotes the element at row i and column j in M. For a joint action $a = (i, j)$, $M_a = M_{(i,j)}$.

that is the expected deviation between the cumulative rewards obtained with the optimal joint action and the cumulative rewards obtained by each player. Without loss of generality, we can focus on the cumulative regret of a single player since the performances of both players are tied together.

Cooperation games We denote as cooperation a sub-case of coordination, where the outcome is equally bad for all agents if all players pursue their individual profit; the optimal behaviour is to *coordinate on the best collective action*. Prisoner's Dilemma is a well-known example of cooperation game. Let R denote the collective payoff, defined such that the element at row i and column j corresponds to the minimum expected payoff over both players given joint action $a = (i, j)$: $R_a = \min_{p\in\{1,2\}} R_a^{(p)}$. We evaluate the performance of learning agents in a cooperation using the *collective cumulative regret*:

$$\mathcal{R}(T) := \sum_{t=1}^{T} \left(R_{a_\star} - \mathbb{E}[R_{a_t}]\right), \tag{2}$$

where the optimal collective joint action $a_\star := \arg\max_{a\in\mathcal{A}} R$ maximizes the minimum outcome.

3 Methodology

This section describes the considered environments (games) and individual learners (players) strategies, along with the design of the study.

3.1 Games

Four games are selected to capture different challenges faced by individual learners in a multi-agent system where coordination is required (Matignon et al., 2012): stochasticity, non-stationarity, alter-exploration, shadowed equilibrium, and Pareto-selection. All games are configured such as to have bounded expected outcomes contained in $[0, 1]$. We consider three pure coordination games (using $R^{(1)} = R^{(2)}$[2]) and one cooperation game to complement prior findings (Douglas et al., 2024).

`Simple` game We begin with a simple two-action coordination game characterized by a single optimal joint action $a_\star = (1, 1)$ to isolate common emergent coordination challenges:

$$R^{(1)} = R^{(2)} = \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix}.$$

This game can be considered as easy since the individual components of $a_\star$ can be identified even when the other player behaves randomly. However, in a learning setting, players policies may evolve over time, which can result in *non-stationary* stochastic rewards (from the perspective of a single player). Moreover, identifying the optimal joint action still requires efficient *joint exploration*, which is dependent on both players exploring sufficiently in a coordinated manner. Performance in this game is evaluated using the cumulative regret of player 1 (Equation 1).

`Pareto` game We investigate the impact of multiple optimal joint actions using a normalized variant of the Pareto-penalty coordination game from Claus & Boutilier (1998):

$$R^{(1)} = R^{(2)} = \begin{bmatrix} 1 & \gamma & 0 \\ \gamma & \beta & \gamma \\ 0 & \gamma & 1 \end{bmatrix}$$

with $0 < \gamma < \beta < 1$. In this game, players must learn to coordinate on one of the optimal solutions, that is $a_\star \in \{(1, 1), (3, 3)\}$, leading to the so-called *Pareto-selection* challenge (Matignon et al., 2012). One can strategically attribute payoffs (γ and β) to lead players into suboptimal solutions.

[2]Without loss of generality since the relationship between $R^{(1)}$ and $R^{(2)}$ is unknown to the players.

We consider an easy variant ($\gamma = 0$, $\beta = 0.2$) to isolate the Pareto-selection challenge, and a hard variant ($\gamma = 0.2$, $\beta = 0.8$) that induces a *shadowed equilibrium* (Matignon et al., 2012). Under this additional challenge, the optimal joint actions are non-distinguishable if the other player displays a uniformly random exploration behaviour, which is typically the case in the beginning of learning. Performance in this game is evaluated using the cumulative regret of player 1 (Equation 1).

`Prisonner's Dilemma` The `Prisoner's Dilemma` is a well-studied two-agent cooperation game generally described by payoff matrices:

$$R^{(1)} = \begin{bmatrix} \beta & 0 \\ 1 & \gamma \end{bmatrix} \qquad R^{(2)} = \begin{bmatrix} \beta & 1 \\ 0 & \gamma \end{bmatrix}$$

with $0 < \gamma < \beta < 1$ (Douglas et al., 2024). This game is characterized by a single optimal joint action $a_\star = (1, 1)$, that is the best collective action. Its complexity arises from the fact that, irrespective of the fixed action chosen by the opposing player, under such payoff matrices, agents are incentivized in selecting action 2, resulting into the Nash equilibrium $(2, 2)$ with outcome γ for both players. However, if players cooperate by choosing the optimal joint action $(1, 1)$, they obtain the highest collective outcome $\beta > \gamma$. We use $\beta = 0.6$ and $\gamma = 0.4$. Performance in this game is evaluated using the collective cumulative regret (Equation 2) with $R = \begin{bmatrix} \beta & 0 \\ 0 & \gamma \end{bmatrix}$.

3.2 Learning agent strategies

When considering independent learning agents, it is natural to model the stochasticity induced by the actions of other agents as reward noise. Akin to prior work focused on competition and cooperation (O'Donoghue et al., 2021; Douglas et al., 2024), we study the emergence of coordination (both individual and collective) using stochastic bandit agents aiming to maximize their individual profit. We consider both deterministic and stochastic widely used strategies to capture the influence of different exploration-exploitation mechanisms.

Let $N_i(t-1)$ and $S_i(t-1)$ respectively denote the number of times that action i was played up to time t (exclusively) and the sum of rewards obtained over these plays[3].

Deterministic player The well-known UCB strategies work by maintaining empirical estimates $\hat{\mu}_i(t-1) := S_i(t-1)/N_i(t-1)$, and select actions based on upper confidence bounds on these estimates (Auer, 2002). Given a fixed history of action plays and associate observations, the next action to play is computer deterministically (Lattimore & Szepesvári, 2020):

$$i_t = \underset{i \in \{1,\dots,k\}}{\arg\max}\, \hat{\mu}_i(t-1) + \sigma \sqrt{\frac{8 \ln(t)}{N_i(t-1)}}, \tag{3}$$

assuming that the stochasticity induced by the other player's policy combined with reward noise is σ-sub-Gaussian. UCB-based strategies typically explore actions at a logarithmic rate over the horizon. We also consider the KL-UCB variant (Garivier & Cappé, 2013), which uses confidence intervals directly derived from the bandit regret lower-bounds (Lai & Robbins, 1985):

$$i_t = \underset{i \in \{1,\dots,k\}}{\arg\max}\, \hat{\mu}_i(t-1) + \sigma \sqrt{\frac{2(\ln(t) + 3\ln\ln(t))}{N_i(t-1)}}. \tag{4}$$

Thanks to its tighter confidence intervals, KL-UCB explores slightly less frequently than UCB (although still at a log-rate). Both UCB and KL-UCB require that each action is played at least one in order for the upper confidence bounds to be computed. Therefore, during the first k rounds of the game, each action is played once in a random order.

[3] We take the perspective of player 1 (row player) without loss of generality.

Stochastic player We consider the widely recognized Thompson Sampling (TS) strategy (Thompson, 1933; Chapelle & Li, 2011; Russo & Van Roy, 2014). On round t, a TS agent selects action i_t based on samples $\theta_{i,t}$ from the posterior distributions associated with each action i. Formally, considering σ-sub-Gaussian noise and a Gaussian prior with prior mean μ_0 and variance σ_0^2:

$$\begin{aligned} &\theta_{i,t} \sim \mathcal{N}\left(m_{i,t}, s_{i,t}^2\right) \quad \text{for each action } i \in \{1, \ldots, k\} \\ &\quad \text{where} \quad m_{i,t} := \frac{\mu_0/\sigma_0^2 + S_i(t-1)/\sigma^2}{1/\sigma_0^2 + N_i(t-1)/\sigma^2} \quad \text{and} \quad s_{i,t}^2 := \left(\frac{1}{\sigma_0^2} + \frac{N_i(t-1)}{\sigma^2}\right)^{-1} \\ &i_t = \underset{i \in \{1,\ldots,k\}}{\arg\max}\, \theta_{i,t}. \end{aligned} \tag{5}$$

Unlike deterministic strategies (like UCB and KL-UCB), two TS agents who have observed exactly the same history of actions and rewards might recommend to different actions at time t.

3.3 Design of the study

We consider four games (`simple`, `Pareto` easy ($\gamma = 0$, $\beta = 0.2$), `Pareto` hard ($\gamma = 0.2$, $\beta = 0.8$), and `Prisoner's Dilemma`) with Gaussian noise on observations, that is $\eta_t^{(1)}, \eta_t^{(2)} \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$ with known variance σ^2_{noise}. We consider three noise levels per game: none ($\sigma^2_{\text{noise}} = 0$); low ($\sigma^2_{\text{noise}} = 0.01$); and high ($\sigma^2_{\text{noise}} = 1$). This results into $4 \times 3 = 12$ environment configurations. Note that the high-noise level was used previously to study competition on Rock-Paper-Scissors (O'Donoghue et al., 2021), whereas the noise-free setting was used in the Prisoner's Dilemma (Douglas et al., 2024). We introduce a low-noise level, aiming to capture the impact of small perturbations as noise can break symmetry in action selections by deterministic agents.

On each of these configurations, we evaluate five pairings of agent strategies. We study the case where both players follow the same learning strategy: UCB×UCB, KL-UCB×KL-UCB, and TS×TS. We study the interaction between deterministic and stochastic agent strategies: UCB×TS. Finally, we study the interaction between deterministic strategies that explore at slightly different rates, while following the same background logic: KL-UCB×UCB. Each pairing of agent strategies is run 500 times on each environment configuration, resulting into 5 agent pairings $\times$ 500 runs $\times$ 12 environment configurations, for a total of $30,000$ runs.

Each run is performed over a horizon of $T = 1000$ rounds. For each run, deterministic strategies initialization (one play for each action) is performed in a random order, that is not necessarily the same for both players. Therefore agents may not be exposed to the optimal joint actions in the first rounds. Stochastic agents are also configured to ensure that coordination is not induced through sampling alignment. To account for the noise induced by the other agent, all strategies are configured with a noise parameter (σ) that combines the observation noise σ_{noise} and the outcome range variance $[0, 1]$: $\sigma = \sqrt{\sigma^2_{\text{noise}} + 1/4}$.

For reproducibility, all code is available online.

4 Results

We present mean cumulative regret (Equation 1) and mean cumulative collective regret (Equation 2) for each agent pairing in each environment configuration. Appendix A contains additional results on the proportion of joint action selections by each agent pairing in each environment configuration.

4.1 Simple game: Stochastic strategies at higher-risk of alter-exploration

Figure 1 displays the mean cumulative regret of player 1 (Equation 1) on the `simple` game, for the three noise levels. We observe that the UCB×UCB and KL-UCB×KL-UCB pairings achieve always sub-linear regret, whereas TS-based pairings incur linear higher regret. Surprisingly, UCB×KL-UCB fails to converge to the optimal joint action in the noise-free setting. When

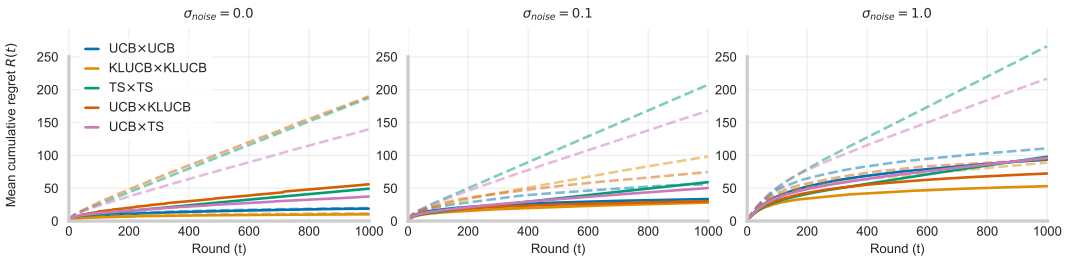

Figure 1: Average cumulative regret (plain lines) with standard deviation (dotted lines) of player 1 on the `simple` game (500 runs for each noise level).

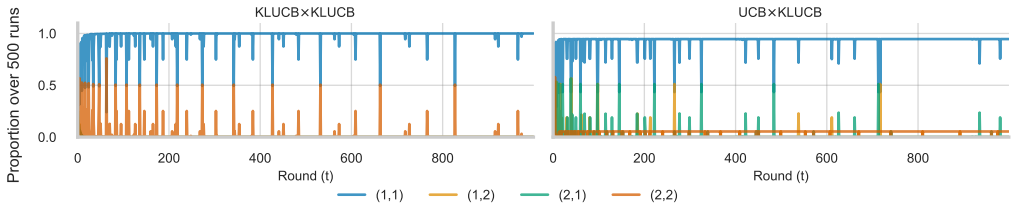

Figure 2: Proportion (over 500 runs) of joint action selections on every round of the `simple` game without noise ($\sigma_{\text{noise}} = 0$).

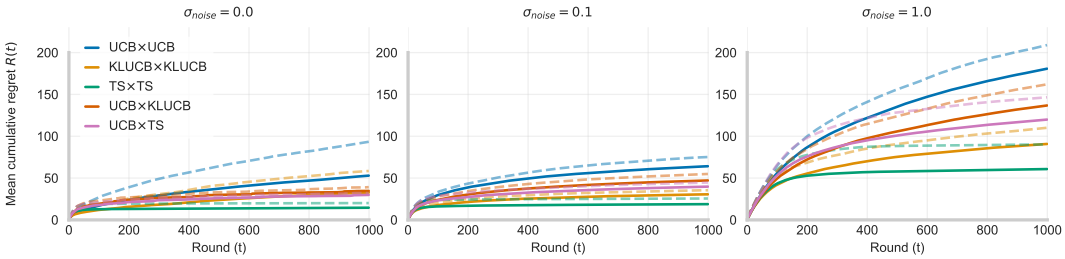

Figure 3: Average cumulative regret (plain lines) with standard deviation (dotted lines) of player 1 on the easy ($\gamma = 0$, $\beta = 0.2$) `Pareto` game (500 runs for each noise level).

looking at the proportion of joint action selections by KL-UCB×KL-UCB and UCB×KL-UCB pairings (Figure 2), we observe that KL-UCB×KL-UCB is characterized by synchronized exploration-exploitation phases, with the duration of exploitation phases increasing over time. However, results for the UCB×KL-UCB pairings show that synchronicity is not guaranteed when the two deterministic agents learn at slightly different speeds. This confirms that pathological alter-exploration cycles may prevent the emergence of coordination even in the simplest settings. Fortunately, agent-independent observation noise appears able to break such cycles (Figure 1, middle and right).

4.2 Pareto game: Pareto-sequences generally help, shadowed equilibria can be deadly

Figure 3 displays the mean cumulative regret of player 1 (Equation 1) on the easy ($\gamma = 0$, $\beta = 0.2$) `Pareto` game, for the three noise levels. We observe that all agent pairings achieve sub-linear regret across all noise levels, except UCB×UCB which achieves sub-linear regret only in presence of observation noise. These results indicate that having multiple joint optimal actions alone may not be a challenge in most real-world (noisy) settings.

Figure 4 displays the mean cumulative regret of player 1 (Equation 1) on the hard ($\gamma = 0.2$, $\beta = 0.8$) `Pareto` game, for the three noise levels. Recall that the hard variant includes the additional shadowed equilibrium challenge. We observe that all pairings except UCB×KL-UCB display linear

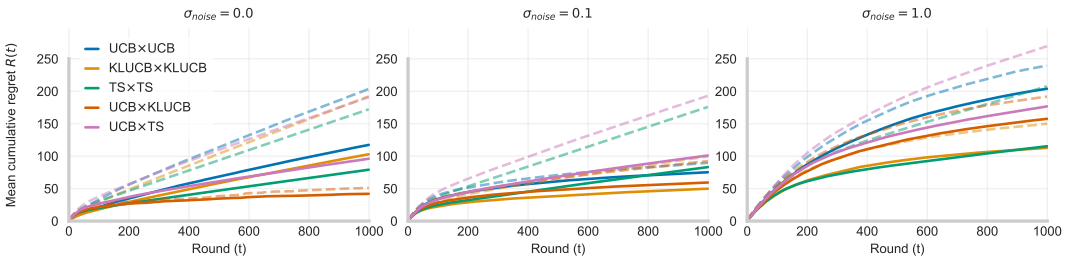

Figure 4: Average cumulative regret (plain lines) with standard deviation (dotted lines) of player 1 on the hard ($\gamma = 0.2$, $\beta = 0.8$) `Pareto` game (500 runs for each noise level).

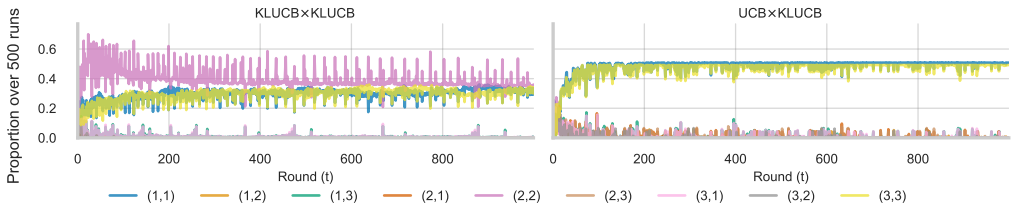

Figure 5: Proportion (over 500 runs) of joint action selections on every round of the hard ($\gamma = 0.2$, $\beta = 0.8$) `Pareto` game without noise ($\sigma_{\text{noise}} = 0$).

regret in the noise-free setting. When looking at the proportion of joint action selections by KL-UCB×KL-UCB and UCB×KL-UCB pairings (Figure 5), we observe that KL-UCB×KL-UCB fails to distinguish the optimal joint actions from the best suboptimal joint action $(2, 2)$, still showing coordination issues. On the other hand, UCB×KL-UCB converges to the optimal joint actions, suggesting that a lack of initial coordination could help to avoid suboptimal joint action $(2, 2)$ early in the game. These results highlight the difficulty of facing shadowed equilibria and suggests that intricate interplays between the learning agent strategies may be required to identify the optimal action under such conditions. Fortunately, we also observe that under sufficient observation noise, fully-deterministic agent pairings all achieve sub-linear regret. This however suggests that alter-exploration dynamics in TS-based pairings are more difficult to break.

4.3 Prisoner's Dilemma: Observation noise slows the emergence of cooperation

Figure 6 displays the mean cumulative collective regret (Equation 2) on the `Prisoner's Dilemma` game, for the three noise levels. We observe that UCB×UCB and KL-UCB×KL-UCB converge to the optimal collective action under the noise-free setting, confirming results from Douglas et al. (2024). However, it is unclear whether this result holds under noisy observation given the considered horizon. This suggest that optimal *collective* joint actions might be more difficult to identify than optimal (individual) joint actions. Interestingly, we observe plateaus in the cumulative regret of UCB×KL-UCB, indicating phases of convergence to the cooperation equilibrium that the agents appear unable to maintain.

Figure 7 displays the proportion of joint action selections by KL-UCB×KL-UCB in the noise-free and low-noise settings, highlighting again the subtlety of mechanisms at interplay in alter-exploration. In the noise-free setting (left), the agents manage to identify the cooperation equilibrium, exploring the Nash equilibrium in phases that appear to extend over time. In the low-noise setting (right), the agents quickly converge to the Nash equilibrium, but the frequency of the cooperation equilibrium increases over time. While noise does appear to slow down convergence, its does not seem to prevent it from emerging.

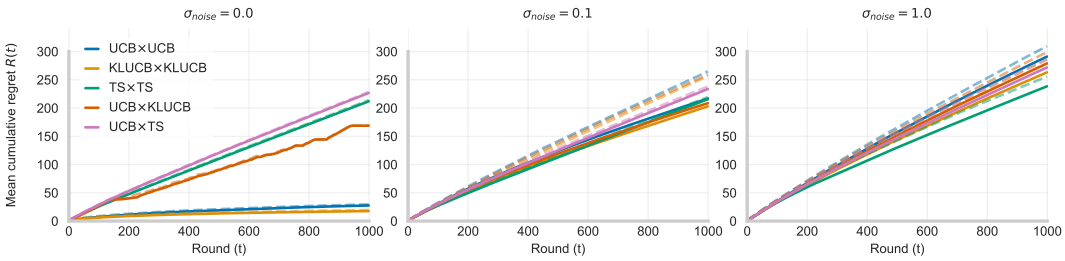

Figure 6: Average cumulative collective regret (plain lines) with standard deviation (dotted lines) on the `Prisoner's Dilemma` game (500 runs for each noise level).

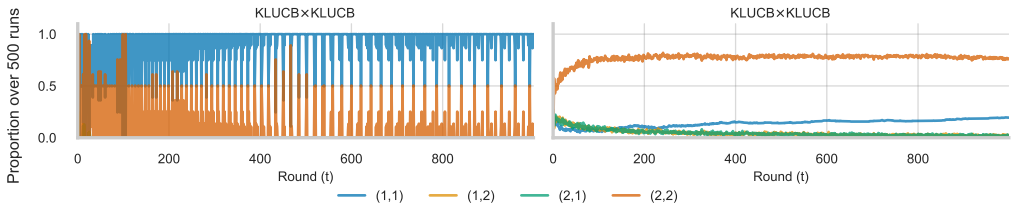

Figure 7: Proportion (over 500 runs) of joint action selections on every round of the `Prisoner's Dilemma` game in the noise-free ($\sigma_{\text{noise}} = 0$, left) and low-noise setting ($\sigma_{\text{noise}} = 0.1$, right).

5 Conclusion

In this work, we study the influence of the learning mechanisms on the emergence of coordination between independent agents. We focus on two-player matrix games where reward matrices are unknown to the players and players observe only their own noisy rewards (bandit feedback). This is among the very few works that consider noisy reward feedback (O'Donoghue et al., 2021) and, to our knowledge, the first work to also consider that the relationship between reward matrices is also unknown to the player.

Our results indicate that coordination tends to emerge more easily from interactions between deterministic agents, especially when they follow the same learning behaviour. This observation was made even in the `simple` game, where TS-based agents failed to converge to the optimal joint action. This suggest that their stable and predictable exploration-exploitation regimes might be less at risk of resulting in alter-exploration dynamics. This finding is interesting considering that Thompson Sampling (TS) is known to perform better under the traditional stochastic bandit setting (Chapelle & Li, 2011). Previous studies (Sadoune et al., 2024; Douglas et al., 2024) on variants of the `Prisoner's Dilemma` have also shown that UCB has a better potential to converge to the optimal joint action compared to stochastic strategies like ε-greedy or TS. However, our findings reveal a counter-example: the same mechanism being the success of determinism strategy pairing can also hamper coordination in the noise-free `Pareto` games. Fortunately, observation noise appears to be sufficient for allowing coordination to emerge in the presence of shadowed equilibria. Finally, our results on `Prisoner's Dilemma` complement prior results (Douglas et al., 2024) by showing that cooperation may still emerge from UCB-based agents under observation noise.

These results motivate further experiments, possibly on longer horizons, to confirm the observed behaviours under the noisier settings. A major challenge remains to parametrize game environments in such was as to isolate the studied challenges to produce meaningful conclusions. Finally, these results call for a theoretical analysis that could provide formal insights and explanations on the observed behaviours. As a final remark, it is important to note a negative impact of implicit coordination. In the field of pricing algorithms, coordination might lead to collusion and increase prices (Harrington, 2018; Calvano et al., 2020; Sadoune et al., 2024; Douglas et al., 2024).

Acknowledgments

This work was funded by IVADO (R3AI-R10). We also thank Mathieu Godbout, Randy Lefebvre, and Alexandre Larouche for their insightful feedback.

References

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 2002.

Bastian Broecker, Karl Tuyls, and James Butterworth. Distance-based multi-robot coordination on pocket drones. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 6389–6394, 2018.

Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, October 2020.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the 15th National Conference on Artificial Intelligence and 10th Innovative Applications of Artificial Intelligence Conference*, 1998.

Connor Douglas, Foster Provost, and Arun Sundararajan. Naive algorithmic collusion: When do bandit learners cooperate and when do they compete? In *Proceedings of the International Conference on Information Systems*, 2024.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference on Learning Theory*, 2013.

Joseph E Harrington. Developing competition law for collusion by autonomous price-setting agents. *Journal of Competition Law & Economics*, 14(3):331–363, September 2018.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Guillaume J Laurent, Laëtitia Matignon, and Nadine Le Fort-Piat. The world of independent learners is not markovian. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 2011.

Laetitia Matignon, Guillaume J. Laurent, and Nadine Le Fort-Piat. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge Engineering Review*, 2012.

Martin J. Osborne and Ariel Rubinstein. *A course in game theory*. MIT Press, 1994.

Brendan O'Donoghue, Tor Lattimore, and Ian Osband. Matrix games with bandit feedback. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*, 2021.

David Robinson and David Goforth. *The topology of the 2x2 games: a new periodic table*. Routledge, London New York, 2006.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Igor Sadoune, Marcelin Joanis, and Andrea Lodi. Algorithmic collusion and the minimum price Markov game, November 2024.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Behrad Toghi, Rodolfo Valiente, Dorsa Sadigh, Ramtin Pedarsani, and Yaser P. Fallah. Social coordination and altruism in autonomous driving. *CoRR*, abs/2107.00200, 2021.

Ermo Wei and Sean Luke. Lenient learning in independent-learner stochastic cooperative games. *Journal of Machine Learning Research*, 17(84):1–42, 2016.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Proportion of joint action selections

A.1 Simple game

Figures 8, 9, and 10 display the proportion (over 500 runs) of joint action selections on every round of the `simple` game for each agent strategy pairing, for each noise level respectively. Note that this game contains a single joint optimal action $a_\star = (1, 1)$.

Noise-free setting We observe (Figure 8) that both UCB×UCB and KL-UCB×KL-UCB pairings are characterized a lot of exploration during the first 100-200 rounds, followed by exploitation phases that become longer and longer over time. On the other hand, TS×TS quickly converge to a high-proportion of $a_\star$ selections that plateaus, maintaining a low but fixed selection proportion of suboptimal (but coordinated) action $(2, 2)$. This results into linear regret (Figure 1, left). Both UCB×KL-UCB and UCB×TS display a strong lack of coordination as they keep playing joint actions $(2, 1)$ and $(1, 2)$ that provide a null outcome (in expectation).

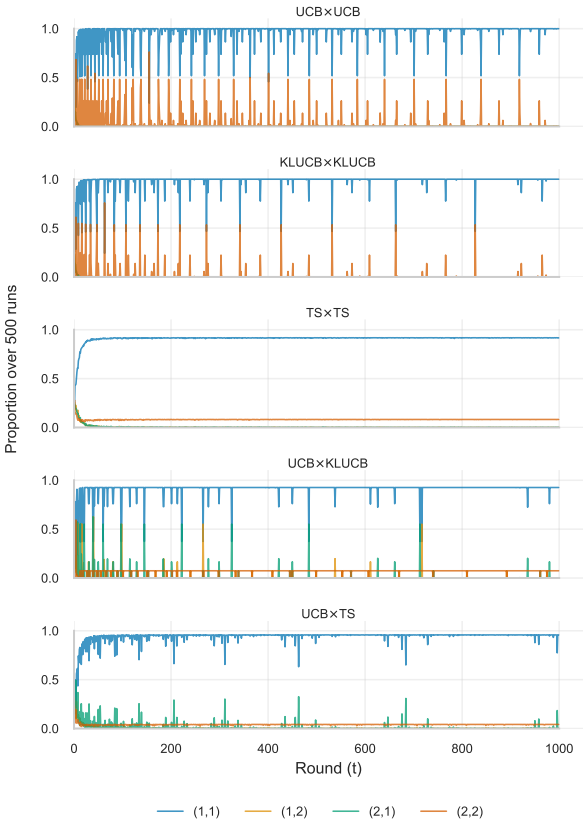

Figure 8: Proportion of joint action selections per agent strategy pairing on the `simple` game without noise. The joint optimal action is $a_\star = (1, 1)$.

Noisy settings In noisy settings (Figures 9 and 10), we observe that all agent strategy pairings quickly converge to a high-proportion of $a_\star$ selections that eventually plateaus, maintaining a low but fixed selection proportion of suboptimal (but coordinated) action $(2, 2)$. The higher the noise, the

slower the convergence. Surprisingly, UCB×UCB and KL-UCB×KL-UCB explore much less than in the noise-free setting. Moreover, miscoordinated exploration patterns that resulted into selections of joint actions $(2, 1)$ and $(1, 2)$ for UCB×KL-UCB and UCB×KL-UCB disappear in the presence of observation noise.

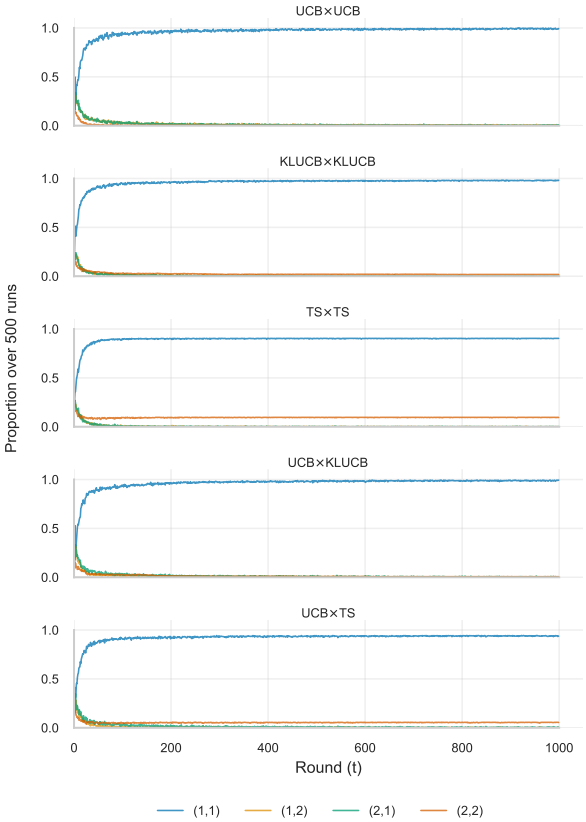

Figure 9: Proportion of joint action selections per agent strategy pairing on the `simple` game with low noise. The joint optimal action is $a_\star = (1, 1)$.

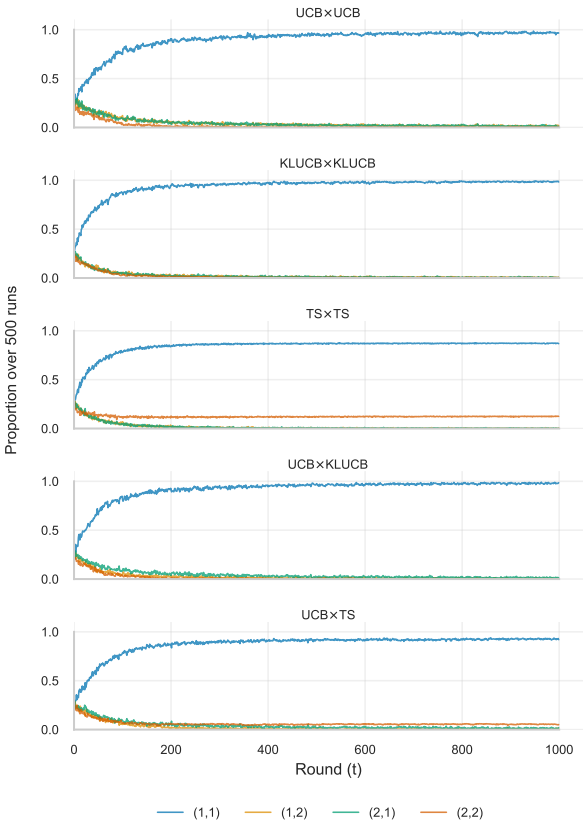

Figure 10: Proportion of joint action selections per agent strategy pairing on the `simple` game with high noise. The joint optimal action is $a_\star = (1, 1)$.

A.2 Pareto game (easy)

Figures 11, 12, and 13 display the proportion (over 500 runs) of joint action selections on every round of the easy ($\gamma = 0$, $\beta = 0.2$) `Pareto` game for each agent strategy pairing, for each noise level respectively. This game contains two joint optimal actions, that is $a_\star \in \{(1,1),(3,3)\}$.

Noise-free setting We observe (Figure 11) that UCB×UCB and KL-UCB×KL-UCB pairings seem to implicitly coordinate their exploration, with the exploration of non-null suboptimal joint action $(2,2)$ triggering switches between the two joint optimal actions. Since switches between optimal joint actions are not instantaneous, they result into some regret (Figure 3, left). On the other hand, TS×TS also alternates between the two optimal joint actions (with probabilities around 50% on each), but switches must we coordinated to obtain logarithmic regret (Figure 3, left).

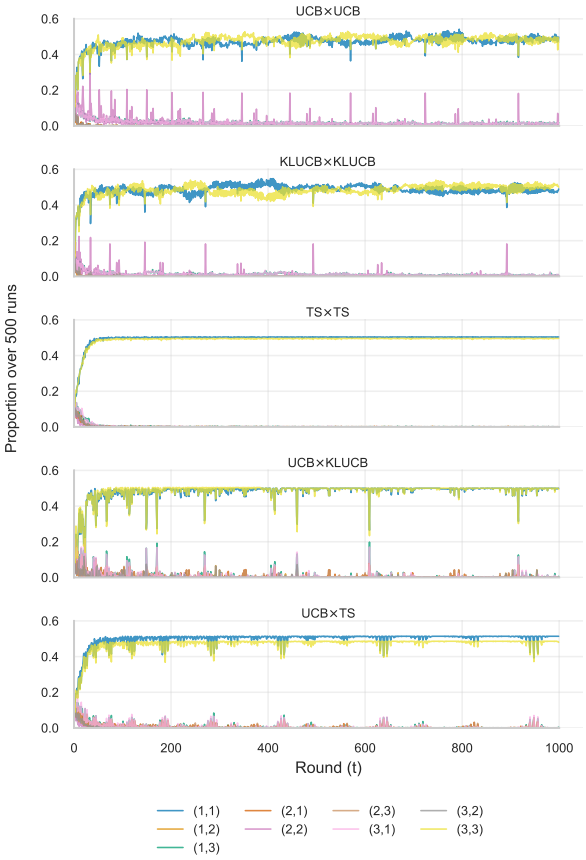

Figure 11: Proportion of joint action selections per agent strategy pairing on the easy ($\gamma = 0$, $\beta = 0.2$) `Pareto` game without noise. The joint optimal actions are $(1,1)$ and $(3,3)$.

Noisy settings In presence of observation noise (Figures 12 and 13), we observe that all pairings result in a behaviour that is similar to TS×TS in the noise-free setting (Figure 11). They alternate between the two optimal joint actions (with probabilities around 50% on each), but switches must be coordinated to obtain logarithmic regret (Figure 3, middle and right).

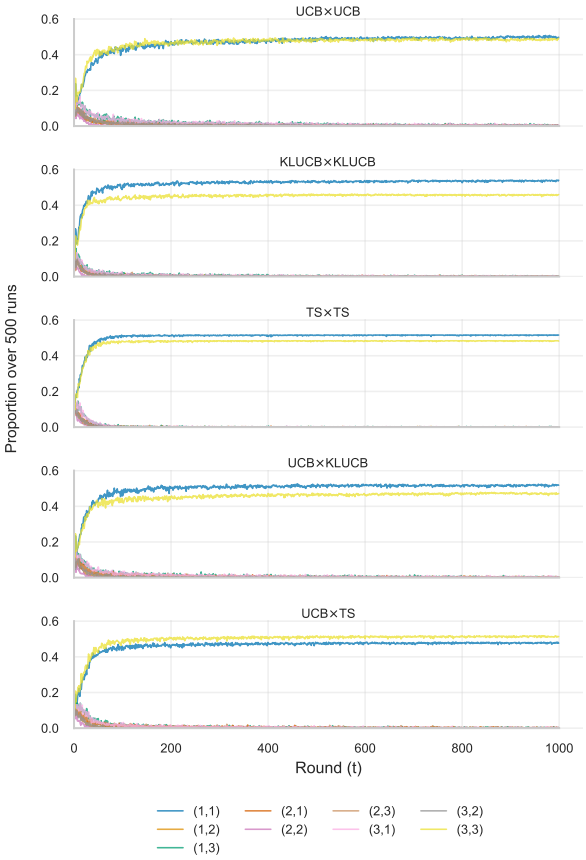

Figure 12: Proportion of joint action selections per agent strategy pairing on the easy ($\gamma = 0$, $\beta = 0.2$) `Pareto` game with low noise. The joint optimal actions are $(1, 1)$ and $(3, 3)$.

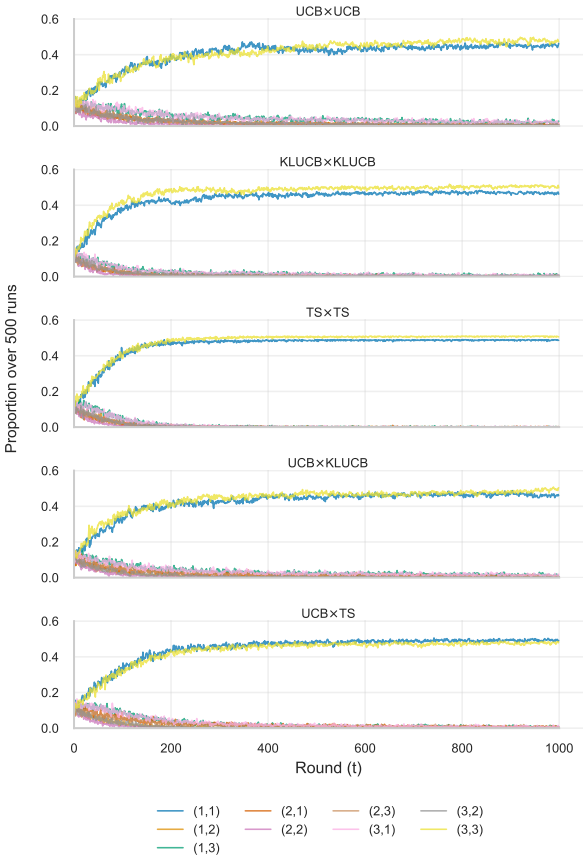

Figure 13: Proportion of joint action selections per agent strategy pairing on the easy ($\gamma = 0$, $\beta = 0.2$) `Pareto` game with high noise. The joint optimal actions are $(1, 1)$ and $(3, 3)$.

A.3 Pareto game (hard)

Figures 14, 15, and 16 display the proportion (over 500 runs) of joint action selections on every round of the hard ($\gamma = 0.2$, $\beta = 0.8$) `Pareto` game for each agent strategy pairing, for each noise level respectively. This game contains two joint optimal actions, that is $a_\star \in \{(1,1),(3,3)\}$, with several non-null suboptimal joint actions $(1,3)$, $(2,2)$, and $(3,1)$. The configuration is such that under a uniform random strategy of player 2, all actions appear to have the same expected outcome (0.4) for player 1.

Noise-free setting We observe (Figure 14) that all pairings but UCB×KL-UCB display linear regret and that UCB×UCB and KL-UCB×KL-UCB both initially converge to the high-paying suboptimal action $(2,2)$, followed by a slow convergence to the optimal joint actions (mixture). Although this suboptimal behaviour translates into linear regret (Figure 4, left), it still shows coordination between the players. On the other hand, TS-based pairings display a quick convergence towards at least one optimal joint action, with a constant proportion of plays remaining on the high-paying suboptimal joint action. Only UCB×KL-UCB manages to ignore the high-paying suboptimal joint action, suggesting that early miscoordination might have (luckily) prevented the agents from identifying action $(2,2)$.

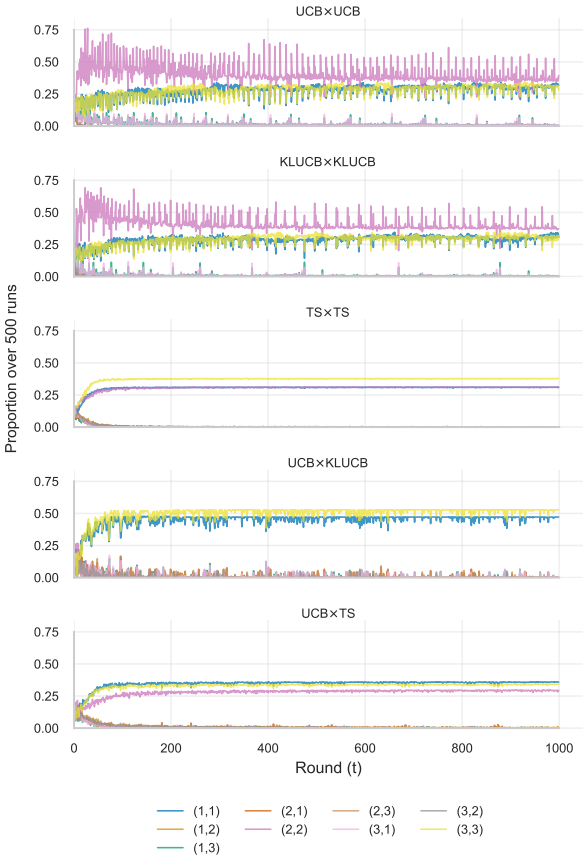

Figure 14: Proportion of joint action selections per agent strategy pairing on the hard ($\gamma = 0.2$, $\beta = 0.8$) `Pareto` game without noise. The joint optimal actions are $(1,1)$ and $(3,3)$.

Noisy settings In presence of observation noise (Figures 15 and 16), we observe that all pairings of deterministic agents converge to the optimal joint actions, showing proper coordination. Observation noise therefore appears to mitigate the arising difficulties when combining several suboptimal joint actions with several optimal joint actions in fully-deterministic agent pairings.

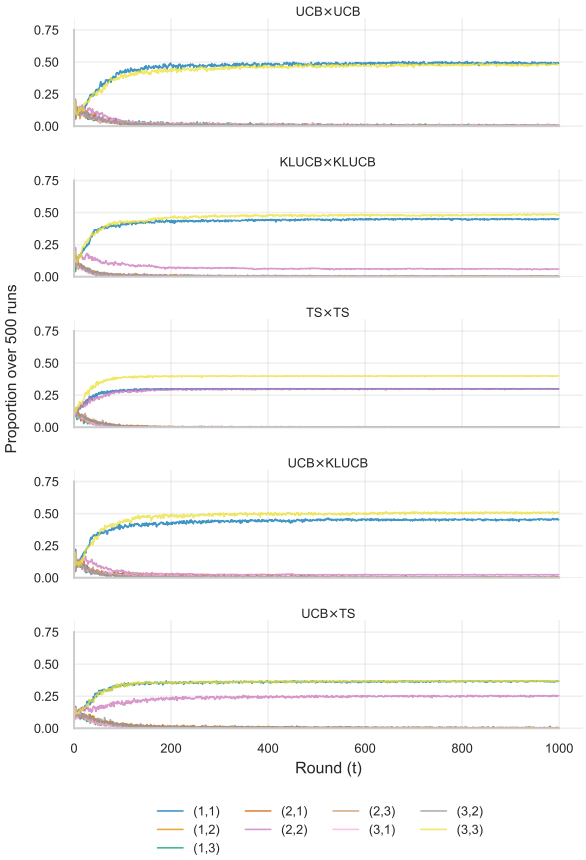

Figure 15: Proportion of joint action selections per agent strategy pairing on the hard ($\gamma = 0.2$, $\beta = 0.8$) `Pareto` game with low noise. The joint optimal actions are $(1, 1)$ and $(3, 3)$.

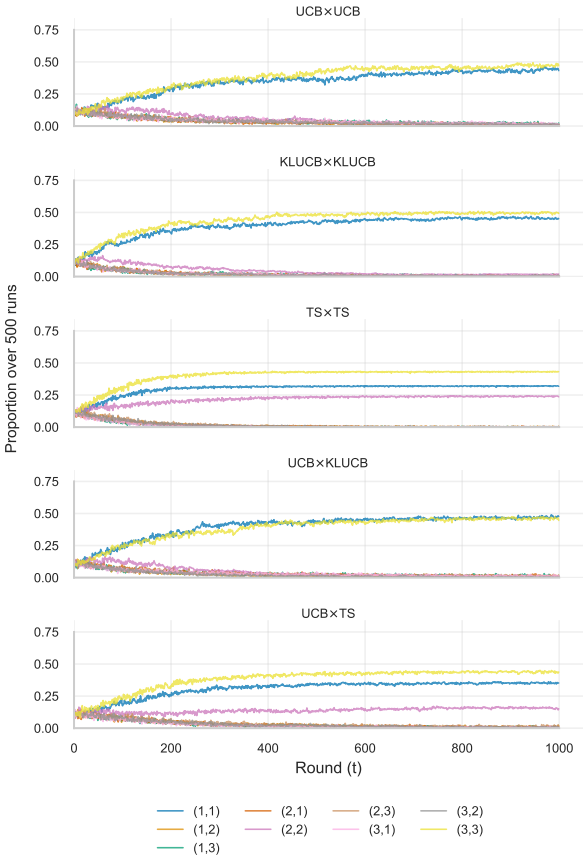

Figure 16: Proportion of joint action selections per agent strategy pairing on the hard ($\gamma = 0.2$, $\beta = 0.8$) `Pareto` game with high noise. The joint optimal actions are $(1, 1)$ and $(3, 3)$.

A.4 Prisoner's Dilemma

Figures 17, 18, and 19 display the proportion (over 500 runs) of joint action selections on every round of the `Prisoner's Dilemma` game for each agent strategy pairing, for each noise level respectively. This game contains one collective joint optimal action, that is $a_\star = (1, 1)$, and the Nash equilibrium at $(2, 2)$.

Noise-free setting We observe (Figure 17) that UCB×UCB and KL-UCB×KL-UCB exhibit coordinated, symmetric behavior: both players switch actions in the same rounds, alternately visiting (1, 1) and (2, 2). This coordination prevents them from playing the mismatched joint actions (1, 2) and (2, 1) that cause difficulties in this game. In contrast, TS×TS quickly converges to the Nash equilibrium, indicating that agents fail to synchronize as they pursue their individual best outcomes.Finally, the UCB×KL-UCB pairing exhibits an interesting pattern in the absence of noise: it shows phases of convergence to the cooperation equilibrium that transform into phases of convergence to the Nash equilibrium. These are probably triggered by individual, miscoordinated exploration, as shown by the high-proportion of selections for joint action $(2, 1)$.

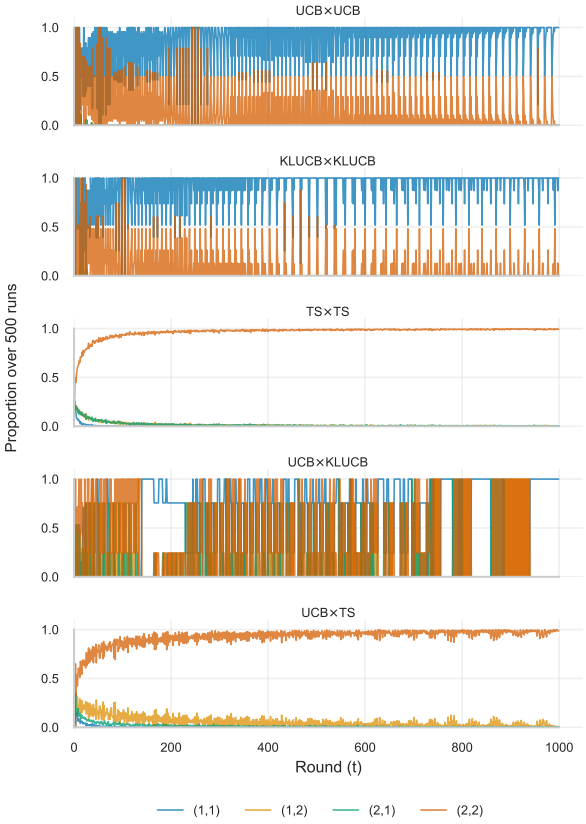

Figure 17: Proportion of joint action selections per agent strategy pairing on the `Prisoner's Dilemma` game without noise. The joint optimal action is $(1, 1)$.

Noisy settings We observe that under low observation noise (Figure 18), all pairings initially converge to the Nash equilibrium $(2, 2)$, the optimal joint action $a\star = (1, 1)$ being played the least. However, pairings of deterministic agents gradually increase their selection of $a_\star$ over time at the expense of the Nash equilibrium. Unfortunately, the horizon appears to be too short to confirm the phenomenon under high observation noise (Figure 19).

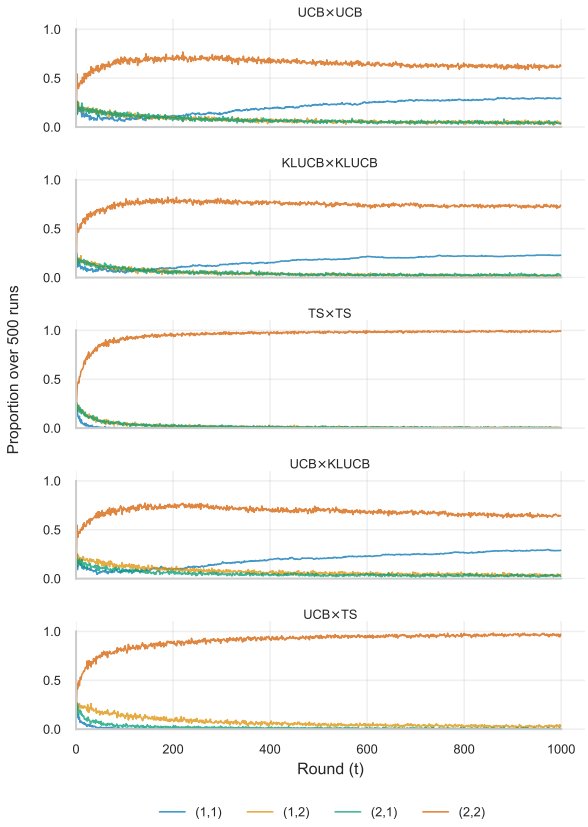

Figure 18: Proportion of joint action selections per agent strategy pairing on the `Prisoner's Dilemma` game with low noise. The joint optimal action is $(1, 1)$.

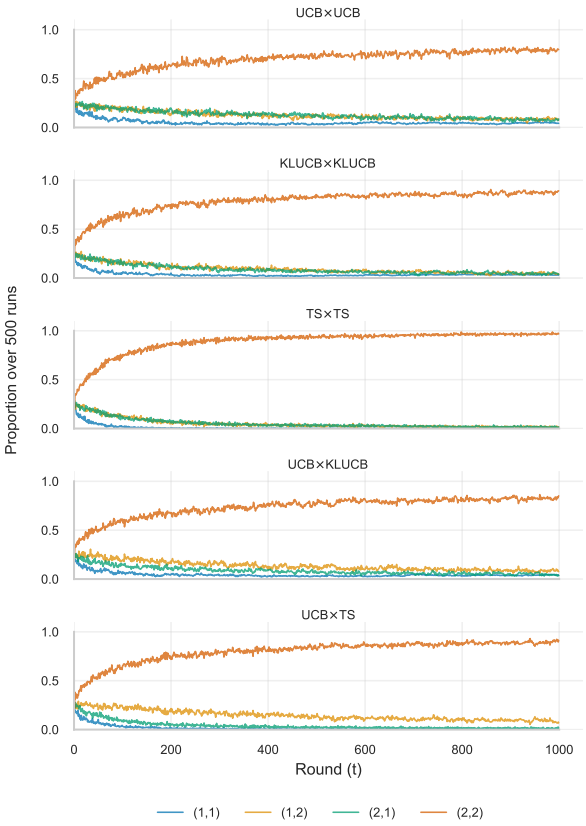

Figure 19: Proportion of joint action selections per agent strategy pairing on the `Prisoner's Dilemma` game with high noise. The joint optimal action is $(1, 1)$