Exploitation by Exploration: 2-player Repeated 2×2 Games with Unknown Rewards

Adam M. Sykulski, Niall M. Adams and Nicholas R. Jennings

Abstract

Many Aladdin problems involve autonomous agents interacting in environments where they must learn and act at the same time. In this report, we consider a specific class of problems where agents have no prior knowledge of the rewards received for the actions they select, which may be typical when agents are acting in a dynamic and uncertain domain. This uncertainty means that agents have to learn as they play, which creates an exploration-exploitation tradeoff to each agent when selecting an action. We use results from both game theory and decision theory to make insights into how agents should act in an unknown environment, and effectively balance this exploration-exploitation tradeoff, which is dependent on the behaviour of the other agents in the environment.

In more detail, we investigate 2-player repeated 2×2 games where the payoff (or reward) structure is unknown *a priori* and the rewards received are observed with noise. We prove that, when an agent selects between the 2 actions using non-explorative strategies, convergence to a Nash equilibrium is not guaranteed in the absence of any additional exploration. Furthermore, we show that an agent that explores using ϵ-greedy exploration, can exploit a non-explorative agent to gain a larger reward in finite time, but only for certain game structures. To this end, approximations of the reward to each agent are constructed for all finite-length 2×2 games, for both explorative and non-explorative strategies, such that the optimal amount of exploration can be approximated. We make use of conditional independence patterns in the decision process, which allow our approximations to scale linearly in the length of the game.

1 Introduction

In many real-world problems, an agent has to make repeated decisions in an environment where the outcomes of these decisions are uncertain or unknown. Moreover, these outcomes are often dependent on decisions made by other agents. Such environments are modelled and investigated using Multi-Agent Systems (MAS). Consider a disaster management scenario (Ramchurn et al., 2008). Emergency service vehicles (the agents) may have to make simultaneous decisions in a potentially unknown environment, for example sequentially deciding which buildings to evacuate. Such systems are often decentralised, meaning coordinated actions are difficult to achieve. Furthermore, there is an interdependence between the optimal action of each agent, so an agent may perform poorly if it ignores the presence of other agents. For example, a collection of ambulances might all convene in the same location, rather than cover the disaster-affected area in a more effective manner.

The need to consider other decision makers, motivates a game theoretic approach to studying such problems. Most game theory literature considers scenarios where each agent has at least some prior information about the reward structure. In contrast, we consider repeated games where the reward structure is completely unknown *a priori*. Rewards are observed with noise thus the agents have to *learn as they play*. In the disaster management scenario, for example, agents are unlikely to know the success of each action (ie. choosing which building to evacuate) before arriving on the scene and furthermore, the success (or reward) of the chosen action is unlikely to be clear immediately, it is instead observed with noise. This framework has been well studied (Claus and Boutilier, 1998; Chapman et al., 2009; Babes et al., 2009; Marden et al., 2009) but findings have been restricted to proving convergence to Nash equilibria in 2-player games

for various homogeneous strategies in self-play. There have been few inroads however in finding strategies that maximise reward in finite time against both homogeneous and heterogeneous opponents, which is the focus of our work, in particular as this is more relevant and applicable to real-world scenarios, especially disasters – where the disaster period is of finite length, and agents are likely to make decisions in different ways.

As noted in Claus and Boutilier (1998), the problem of learning rewards in games can be viewed as a distributed bandit problem. The multi-armed bandit problem (Sutton and Barto, 1998) is a single-agent learning problem where the agent is faced with the challenge of finding the optimal action from a set of actions with unknown expected rewards. Such problems inherently suffer from the exploration-exploitation tradeoff, where the agent must choose between what it believes is the best action (exploitation) and trying alternative actions for potential future benefit (exploration). The extension of the bandit framework to multiple agents, such that the agents face inter-dependent decision making problems, therefore captures many of the multi-agent learning problems inherent in MAS – the need to learn how to act in an unknown environment in the presence of other decision makers.

In this work, we focus on interactions between two agents, where each agent must sequentially choose between two actions. This characterisation is still useful in a wider setting (with more agents and actions) as, in particular, many real-world situations can be modelled in this way (Govindan and Wilson, 2010). Specifically, the actions of all opposing agents are represented as the action of one agent and all sub-optimal actions are treated as the alternative action to the optimal. This simplification of the problem, provides a drastic reduction in the complexity of finding solutions in such systems, as compared with models that have large numbers of agents or actions. Of course, this reduction comes at the expense of capturing some of the differences between agents and their various action choices, particularly in a system with many heterogeneous agents that have a wide array of action choices. Larger systems will be considered as part of future work (see Section 10), and are a natural extension of the techniques used in this report. Nevertheless, the findings in this report are novel in the 2-agent, 2-action, setting (known as 2×2 games) and are therefore fundamental to understanding the need for exploratory behaviour in an unknown reward setting with multiple agents.

In more detail, for 2×2 games, we provide a proof that if both agents ignore the need to explore actions and instead act myopically (using either fictitious play or greedy action selection), then the strategies can converge to non-Nash equilibria, which are often suboptimal. Moreover, we demonstrate that non-explorative agents can easily be exploited by other agents that do explore. We use strategies from the multi-armed bandit framework, which address the exploration-exploitation tradeoff, and combine them with game theoretic strategies that consider the presence of an opponent. We provide simulation results for our novel strategies from various motivating examples, to show that an explorative agent can *exploit by exploring* – ie. gain a higher reward at the expense of the opponent's, with a suitably tuned exploration parameter. This result motivates a theoretical analysis of the problem, where we show that the expected rewards for different combinations of strategies, can be efficiently approximated and then used to find near optimal ϵ for finite-length games. These novel findings can be used to construct strategies that can learn the exploration parameter on-line, where the optimal value is dependent on both the reward structure and the type of opponent faced. This work therefore provides the first analysis detailing the need for agents to consider exploration in multi-agent sequential decision making problems, with unknown rewards – and furthermore is the first to formulate methods to approximate the optimal level of exploration to each agent.

This paper is structured as follows. Section 2 outlines the framework and case study games used in this report. Section 3 defines two separate types of multi-agent learners: individual and joint-action learners. We also propose several different strategies for selecting actions. Section 4 investigates some of our case games with strategies using no exploration and Section 5 outlines a proof showing the possible convergence of such strategies to non-Nash equilibria. Sections 6 and 7 investigate our case games but this time with the agents using explorative strategies. In Sections 8 and 9 we formulate a full calculation for the expected rewards for each combination of strategies for all 2×2 games and construct an approximation that scales linearly (in the length of the game). Section 10 is reserved for conclusions and future work.

2 Framework and Case Study Games

Agent A and agent B repeatedly play a stage game where they both choose between action 1 and action 2 at each time-step, $t = 1, 2, 3, \ldots, T$. Agent $k = A, B$ receives a reward $r_k(t)$, where:

$$r_A(t) = a(i, j) + \eta_t, \tag{1}$$
$$r_B(t) = b(i, j) + \nu_t, \tag{2}$$

where $i, j \in \{1, 2\}$ are the actions picked by agents A and B respectively, at time t. η_t and ν_t are noise processes which we presume to be i.i.d. Gaussian and centred at zero with variance σ_η^2 and σ_ν^2 respectively. The expected reward matrix of the stage game therefore is:

		Agent B	
		Action 1	Action 2
Agent A	Action 1	$a(1,1), b(1,1)$	$a(1,2), b(1,2)$
	Action 2	$a(2,1), b(2,1)$	$a(2,2), b(2,2)$

The agents' objective is to maximise their cumulative reward $R_k(T) = \sum_{t=1}^{T} r_k(t)$ (for $k = A, B$). The agents observe their own reward and also observe the action selected by the opponent.

Throughout this report we consider two well-studied repeated games to illustrate the performance of our various strategies. Together, these games characterise conflicts that frequently arise between reward-maximising agents. In later sections, we will see that both games demonstrate differing sub-optimal and non-Nash behaviour with un-explorative strategies and furthermore, the optimal level of exploration is markedly different due to the underlying reward structure. This motivates the inclusion of both case games. Note that each case game represents the expected reward for each joint action – the actual reward received is observed with noise.

Case Game 1: Matching Pennies

Consider the following zero-sum stage game, known as matching pennies:

		Agent B	
		Action 1	Action 2
Agent A	Action 1	$1, -1$	$-1, 1$
	Action 2	$-1, 1$	$1, -1$

Agent A wishes to match (both agents select the same action) and agent B prefers to not match. There is hence no pure strategy Nash equilibrium, instead the Nash equilibrium is a pair of mixed strategies where both agents select each action with probability 0.5.

Case Game 2: Prisoner's Dilemma

We also consider the following non-zero sum stage game, known as the prisoner's dilemma:

		Agent B	
		Action 1	Action 2
Agent A	Action 1	$1, 1$	$-2, 2$
	Action 2	$2, -2$	$-1, -1$

The Nash equilibrium is for both agents to "defect" and select action 2, despite the fact that this joint action pair is not Pareto efficient (both agents could gain a higher reward by "cooperating" through selecting action 1). In a repeated game setting, however, there exist several strategies that can outperform playing the Nash strategy for every stage game (Rogers et al., 2007, and references therein).

3 On-line Learning Strategies

The agents do not know the reward structure of the stage game *a priori*. The agents therefore have two distinct learning operations: **estimating** the rewards and **adapting** to the opponent's strategy. These have to be handled concurrently with reward seeking behaviour, otherwise agents can easily select sub-optimal actions. There are two distinguishable forms of multi-agent learning that the agents can use to sequentially estimate and adapt (Claus and Boutilier, 1998). **Independent learners** (ILs) would apply learning in the classical sense, ignoring the existence of the other agent. Conversely, **Joint action learners** (JALs) would make decisions based on their own past actions in conjunction with those of the opponent. JALs are more sophisticated in that they use observations of both the reward received and the action selected by the opponent to learn. Specifically, JALs estimate rewards in the joint action space and can adapt to the opponent by making inferences on its past history of actions. ILs however, only use observations of the reward and ignore information about its opponent – therefore ILs estimate rewards in an individual action space and adapt using this information only.

In various numerical simulations performed in Claus and Boutilier (1998) for cooperative games, ILs were found to perform not much differently from JALs with only slightly slower convergence to a Nash equilibrium. Nevertheless, this convergence was guaranteed by using Boltzmann (also known as SoftMax) exploration (Luce, 1959) and it is not clear how performance might differ between ILs and JALs with no added exploration, particularly in finite time. It is for these reasons that we consider both ILs and JALs in our analysis.

3.1 Independent Learners (ILs)

After selecting an action, the agents only observe their own reward and which action the opponent selected. ILs however, choose to ignore the opponent and make inferences based on the reward received only. The decision problem is then analogous to a bandit problem (Sutton and Barto, 1998), where the agent must sequentially choose which of two arms to pull. The estimated expected reward of each arm ($\widehat{a}(1)$ and $\widehat{a}(2)$ for agent A) can be updated at time t using recursive averaging:

$$\widehat{a}(i) \leftarrow \widehat{a}(i) + \frac{1}{n_i^A(t)} \left(r_A(t) - \widehat{a}(i) \right), \tag{3}$$

for $i = 1, 2$ when action i is selected, and similarly for agent B. $n_i^A(t)$ is the number of times agent A has selected action i prior to time t. Note that this form of recursive averaging is identical to Q-learning (Watkins and Dayan, 1992) in a stateless setting, where the learning parameter ($\lambda = 1/n_i^A(t)$) decays proportionately to the number of samples.

The agents must then use these estimated rewards to select an action to play at the next iteration. In the absence of any exploration, the obvious way to do this is to adopt a greedy strategy and select the action with the higher valued estimate – we refer to this strategy as **bandit greedy**.

Exploration of actions is required to guarantee convergence to Nash equilibria (Claus and Boutilier, 1998), but can also improve performance in finite time for bandit problems (Vermorel and Mohri, 2005). For these reasons, we consider strategies with exploration in our framework. There are many methods of exploration that can be used from the bandit literature such as ϵ-greedy (Watkins, 1989), Boltzmann, interval estimation (Kaelbling, 1993) and upper confidence bounds (UCB) (Auer et al., 2002). We consider only the ϵ-greedy strategy – primarily because it is documented to perform best in a variety of bandit problems (Vermorel and Mohri, 2005; Auer et al., 2002; Pavlidis et al., 2008). In addition, the ϵ-greedy concept is very simple and easy to implement (with only one tuning parameter). Moreover, the main weakness of ϵ-greedy is not relevant to a 2-action problem like ours. Specifically, in a problem with 3 or more actions, exploration is performed uniformly across the action space (unlike the other above mentioned strategies). This means explorative actions that are almost definitely sub-optimal can be selected as frequently as explorative actions that have a much higher potential to be optimal – which appears counterintuitive. This issue is of no significance in a 2-action game where there is only one explorative action to select from. We therefore define the **bandit ϵ-greedy** strategy as

follows:

$$\text{with probability} \begin{cases} \epsilon & \text{select action with highest estimated reward} \\ 1-\epsilon & \text{select action with lowest estimated reward} \end{cases} \tag{4}$$

3.2 Joint Action Learners (JALs)

Conversely, JALs learn on the joint action space by observing the actions selected by the opponent. In the 2-player, 2-action game studied here, each agent has 4 running estimates of rewards: $\widehat{a}(i,j)$ and $\widehat{b}(i,j)$ for $i,j = 1,2$. These can again be updated by agent A and B respectively at time t, using recursive averaging:

$$\widehat{a}(i,j) \leftarrow \widehat{a}(i,j) + \frac{1}{n_{i,j}(t)} \left(r_A(t) - \widehat{a}(i,j) \right), \tag{5}$$

updated when agent A selects action i and B selects j (and similarly for agent B). $n_{i,j}(t)$ is the number of times joint action $\{i,j\}$ has been selected up to time t. Notice again that this is analogous to stateless Q-learning as performed in Claus and Boutilier (1998).

The agents must use these joint action reward estimates, together with the past history of actions, to select the next action. There are several game-theoretic methods with which this can be done, including fictitious play (Brown, 1951), adaptive play (Young, 1993), regret-matching strategies (Marden et al., 2007) and one-shot Nash (Fudenberg and Maskin, 1986). In this research, we consider the agents using fictitious play. This strategy does not require knowledge of the opponent's rewards (required to calculate one-shot Nash and regret-matching strategies) and is also suitable for a game with a static reward process (an adaptive play strategy would be more naturally suited to a dynamic reward process). Furthermore, fictitious play (with known rewards) has been shown to converge to a Nash equilibrium for a variety of games including zero-sum games (Brown, 1951), potential games (Monderer and Shapley, 1996), games that are solvable by iterated elimination of dominant strategies (Nachbar, 1990) and more recently all $2\times N$ games where an appropriate tie-breaking rule is used to separate actions of equal preference (Berger, 2005). Games that converge to a Nash equilibrium under fictitious play are said to have the Fictitious Play Property (FPP).

We can define the **fictitious play** strategy as follows. The agents select the best response to the empirical frequency of actions selected by the opponent thus far. Suppose that at time t, agent B has selected action 1 on $n_1^B(t)$ past plays (and action 2 for t-$n_1^B(t)$). Action 1 is therefore a best response to agent A if and only if:

$$n_1^B(t)a(1,1) + (t - n_1^B(t))a(1,2) > n_1^B(t)a(2,1) + (t - n_1^B(t))a(2,2), \tag{6}$$

and action 2 is a best response if the inequality is reversed. The agents, however, do not know the true values of the rewards and instead select the **predicted best response**, using the estimated rewards from (5) rather than the unknown true values.

Without any exploration, there is no guarantee that fictitious play will converge to a Nash equilibrium, due to the fact that rewards are observed with noise. Claus and Boutilier (1998) use Boltzmann exploration with fictitious play for cooperative games and Chapman et al. (2009) use ϵ-greedy exploration with fictitious play for non-cooperative and potential games – both show that their strategies converge to a Nash equilibrium for specific games, if the exploration parameter, ϵ, decays to zero. For the same reasons given earlier, we pursue ϵ-greedy exploration – the method often works well in finite time and when used with fictitious play converges to a Nash equilibrium. We therefore define the **ϵ-FP** strategy as follows:

$$\text{with probability} \begin{cases} \epsilon & \text{select action that is the predicted best response} \\ 1-\epsilon & \text{select action that is the predicted worst response} \end{cases} \tag{7}$$

3.3 Summary

We have constructed strategies for both ILs and JALs and also explorative and non-explorative agents. Henceforth we refer to these strategies as they are denoted in Table 1.

		On-line Learning	
		Joint Action Leaner (JAL)	Individual Leaner (IL)
Exploration	No	Type I: fictitious play	Type II: bandit greedy
	Yes	Type III: ϵ-FP	Type IV: ϵ-greedy

Table 1: Different types of strategies for on-line learning agents

For both ILs and JALs we initialise these strategies with both agents selecting each action once (in a random order) as would often be done in a bandit problem (Auer et al., 2002). We therefore do not assume each joint action is sampled once, as this requires coordinated initialisation – although all our results and proofs in this report could be extended to deal with this and other initialisation procedures.

4 Non-explorative Strategies

In this section we consider agents playing with non-explorative strategies (ie. Type I or Type II strategies from Table 1). These strategies would perform well in a setting with known rewards, but we are interested in the impact of no explicit exploration when rewards are unknown. We present simulation results for both case games, where each stage game is repeated 50 times – this length of game is sufficiently long such that agents can learn in a noisy environment but short enough such that fast learners are rewarded.

4.1 Case Game 1: Matching pennies

• Type I: fictitious play vs. Type I: fictitious play

If both agents use fictitious play with full knowledge of the rewards then both agents would converge to the mixed strategy Nash equilibrium. With no prior knowledge and no exploration, however, no such convergence is guaranteed. Figure 1 shows the average proportion of times that Action 1 is selected by each agent, for the matching pennies game of length 50 (over 100,000 repeats). Each subplot shows results for games with different noise variances (where top left is the lowest variance and bottom right is the highest). For low noise variances, the agents seem to converge towards the Nash equilibrium without any additional exploration. As the magnitude of the noise increases, however, the agents increasingly play pure strategies. This is due to a lack of exploration. For example, agent B may have observed an unusually low reward for Action 2 in an early round of the game and calculates Action 1 to be the dominant strategy. Action 2 is never revisited to correct this error and the agent is subsequently exploited by agent A (who selects Action 1 to match). This pattern can explain all 4 possible combinations of pure strategies being played for certain games. The initial observations are the most crucial, as this is when reward estimates are furthest from their true values, and can therefore have the biggest impact on the long-term convergence of the strategies. Different initialisation procedures will also effect the long-term convergence of a strategy. Specifically, a longer and more explorative initialisation will result in more games converging to Nash equilibria, for the same strategy, than with a shorter initialisation sequence.

• Type II: bandit greedy vs. Type II: bandit greedy

If both agents are ILs using the bandit greedy strategy, then even with full knowledge of rewards, convergence to the Nash equilibrium is not guaranteed. Figure 2 (left) displays average rewards from the same setup as Figure 1, except both agents are ILs. The noise variance has been deliberately set low and despite this both players are playing pure strategies when they are ILs – approximately half the games favour agent A and the remainder agent B. The lack of exploration immediately forces one agent to commit to one action first and then the other exploits this choice.

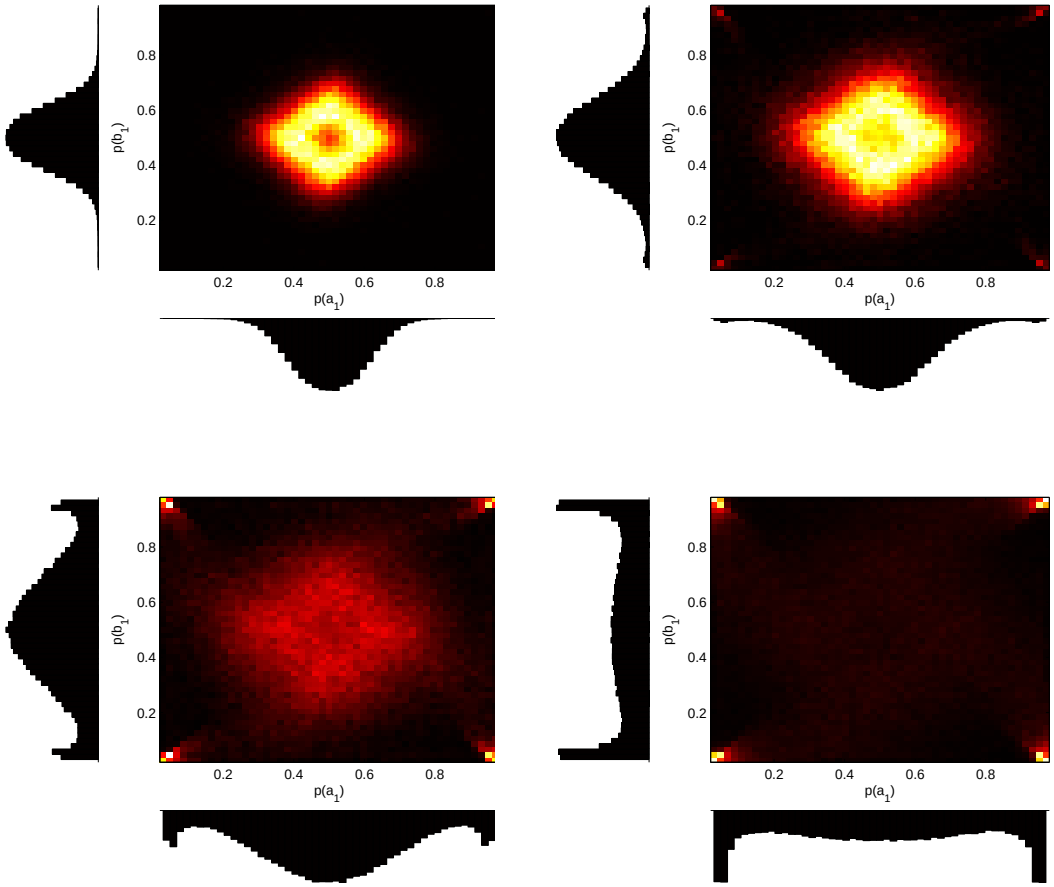

Figure 1: Density plots showing the proportion of times Action 1 is selected by agent A ($p(a_1)$) and agent B ($p(b_1)$) in the matching pennies game of length 50 over 100,000 repeats for noise variances of 0.25 (top left), 0.5 (top right), 1 (bottom left) and 2 (bottom right). Both agents are JALs using Fictitious Play.

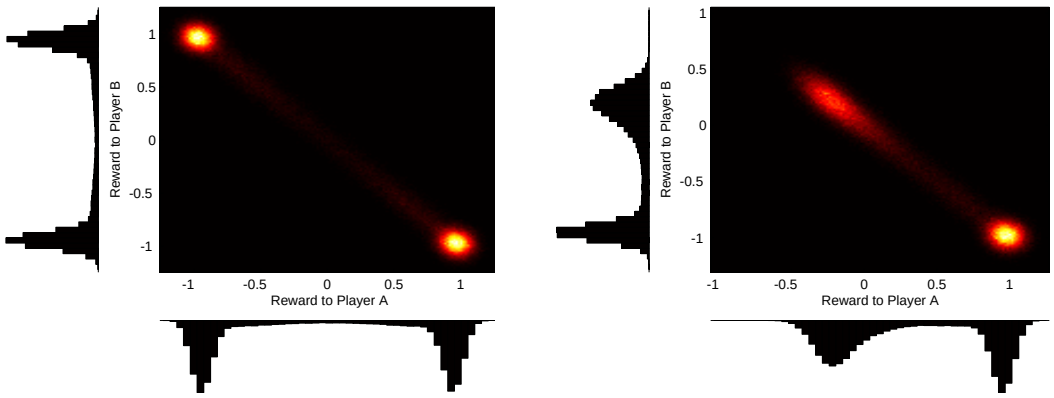

Figure 2: Density plots showing the average rewards for agent A and agent B in the matching pennies game of length 50 over 100,000 repeats for a noise variances of 0.25, where both agents are ILs (left) and agent A is a JAL and agent B an IL (right).

• **Type I: fictitious play vs. Type II: bandit greedy**

In Figure 2 (right), agent A is a JAL (using fictitious play) and agent B is an IL. Some games resulted in plays close to the Nash equilibrium but the majority converged to pure strategies where agent A (the JAL) exploits agent B. Learning on the joint action space allows agent A to distinguish between matched and un-matched actions and exploit agent B who plays a bandit problem and often settles on one action. This advocates the use of joint action learning, over independent learning, when the action taken of the opponent is observed.

4.2 Case Game 2: Prisoner's Dilemma

• **Type I: fictitious play vs. Type I: fictitious play**

For the prisoner's dilemma game, fictitious play would immediately converge to both agents defecting with known rewards, as this is a dominant action and hence is a best response regardless of the frequency of the opponent's actions. Figure 3 displays average rewards, in the unknown rewards setting, for 2 fictitious players playing the prisoner's dilemma game of length 50 (for 100,000 repeats), with low noise variance (left) and high variance (right). Notice that although around half of the games have converged to the agents defecting and receiving the Nash equilibrium reward of -1, other games have both agents "cooperating" and selecting the dominated pure strategy. This happens by chance from the noisy reward estimates. When the noise variance is high some games also converge to cooperate/defect. This non-Nash convergence is again attributed to the lack of exploration of the joint action space, which has resulted in incorrectly calculated best responses, especially when the noise variance is high. This selection of dominated strategies has actually resulted in a higher expected reward to each agent than if the rewards were known, due to the Pareto optimality of cooperating. In particular, the clustering of rewards around $(1, 1)$ has improved the expected reward from -1 (with known rewards) to -0.01 when the noise variance is low and -0.19 when the noise variance is high.

• **Type II: bandit greedy vs. Type II: bandit greedy**

Figure 4 (left) displays the same results for two ILs using bandit greedy (left). For this game setting, the ILs have performed similarly to JALs with almost identical convergence characteristics – although the larger clusters of average reward values suggest that the convergence has been slightly slower (this was also found to be the case in Claus and Boutilier (1998)).

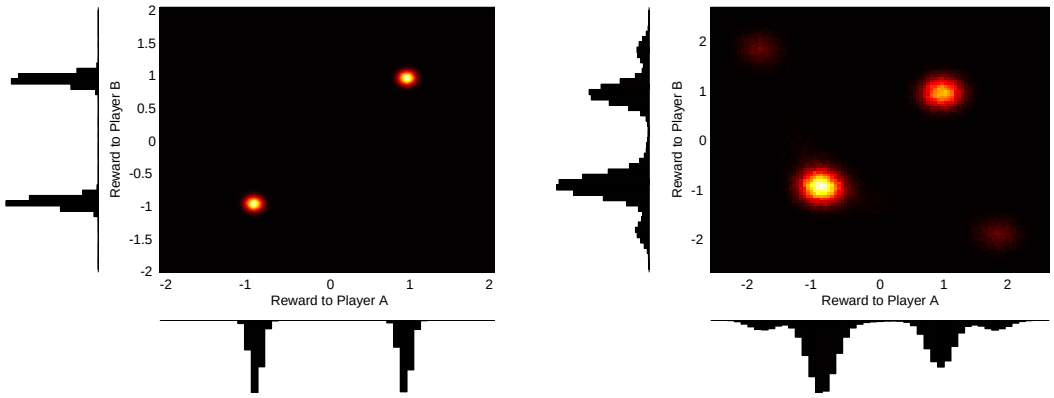

Figure 3: Density plots showing the average rewards for agent A and agent B in the prisoner's dilemma game of length 50 over 100,000 repeats for a noise variances of 0.25 (left) and 2 (right). Both agents are JALs using Fictitious Play.

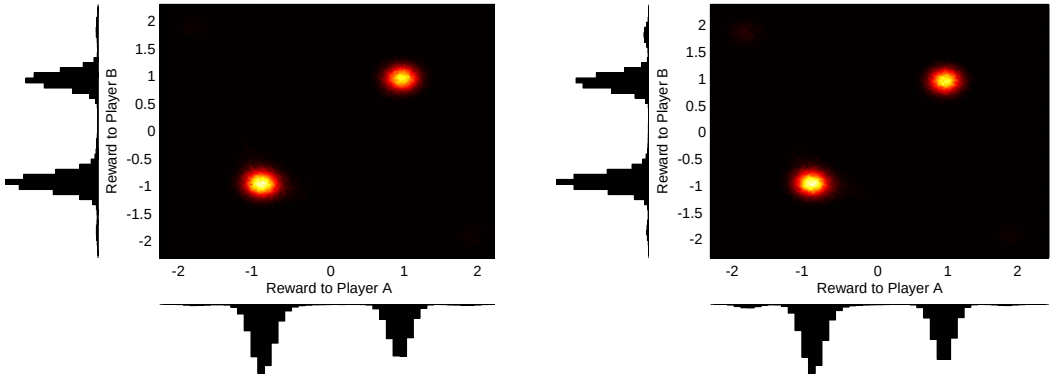

Figure 4: Density plots showing the average rewards for agent A (R_A) and agent B (R_B) in the prisoner's dilemma game of length 50 over 100,000 repeats for a noise variances of 0.25, where both agents are ILs (left) and agent A is a JAL and agent B an IL (right).

• **Type I: fictitious play vs. Type II: bandit greedy**

In Figure 4 (right) we show results for a JAL against an IL (right). On this occasion, the IL has performed as well as the JAL with very similar action selection behaviour. This is because, for this reward structure, the problem is more like a bandit problem – with one action dominating the other. Explicit knowledge of the joint action space is therefore of no particular benefit if the agent is indifferent between the actions of the opponent and trying to calculate a best response.

5 Convergence of non-explorative strategies

2×2 games, with the genericity assumption[1], can have either 0, 1 or 2 pure strategy Nash equilibria, 0 or 1 mixed strategy Nash equilibria and at least 1 Nash equilibrium overall (Dixit et al., 2004). In the previous section we demonstrated that, with unknown rewards, two fictitious players will not necessarily converge to playing a Nash equilibrium strategy in finite time. A

[1]The genericity assumption (Pruzhansky, 2003) states that an agent has a preferred action for every fixed action of the opponent, specifically $a(1,1) \neq a(2,1)$, $a(1,2) \neq a(2,2)$, $b(1,1) \neq b(1,2)$ and $b(2,1) \neq b(2,2)$.

number of simulations resulted in both agents sticking to a non-Nash pure strategy. In fact, it can be proven that in a 2×2 game (with the genericity assumption) the joint strategies will converge to one of 4 or 5 points: the 4 pure strategy profiles (at least two of which are non-Nash) and possibly a mixed strategy Nash equilibrium, if one exists.

We now provide a sketch of the proof, which consists of four parts. First, we show that the game cannot converge to any joint mixed strategy other than a Nash equilibrium. This is clear as both agents would then be sampling from all 4 points of the joint action space infinitely often, which would make all reward estimates of (5) converge to their true values. Fictitious play with known rewards converges to a Nash equilibrium – therefore converging to a non-Nash joint mixed strategy is not possible.

Secondly, we show that the game cannot converge to one pure strategy and one mixed strategy. In this instance, the mixed strategy player would have 2 reward estimates that are infinitely sampled and under the genericity assumption these values would be different – hence the agent converges to a pure strategy and cannot remain on a mixed strategy.

Thirdly, we show that convergence to any of the 4 pure strategy combinations is attainable. Consider, without loss of generality, both agents converging to action 1. There is a probability greater than 0 that $\widehat{a}(2,1) < a(1,1)$ and $\widehat{b}(1,2) < b(1,1)$, as the noise is unbounded. We need to prove that there is a probability greater than 0 such that $\widehat{a}(1,1) > \widehat{a}(2,1)$ perpetually.

In this formulation, the estimate $\widehat{a}(1,1)$ is an average of i.i.d Gaussian samples, $X_i \sim \mathcal{N}(a_{1,1}, \sigma_\eta^2)$. Suppose that $\delta = a(1,1) - \widehat{a}(2,1)$ where $\delta > 0$. We are therefore trying to prove that:

$$\prod_{j=1}^{\infty} P\left(\frac{\sum_{i=1}^{j} X_i}{j} > a(1,1) - \delta\right) > 0 \quad \text{for} \quad \delta > 0, \tag{8}$$

which after some rearranging is equivalent to proving that if:

$$Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\delta, 1), \quad W_j = \sum_{i=1}^{j} Y_i, \tag{9}$$

then,

$$\left(\prod_{j=1}^{\infty} P(W_j > 0)\right) > 0 \quad \text{for} \quad \delta > 0 \tag{10}$$

The process W_j is a random walk, or a discretised Brownian motion with positive drift δ. We have from Chang (1999) that the probability of continuous Brownian motion W_t (with positive drift δ) never falling below zero for $1 \leq t \leq \infty$ is $e^{-2\delta^2} > 0$. The probability for the discretised version is therefore bounded below by this. We have hence proved that there is a probability greater than 0 such that $\widehat{a}(1,1) > \widehat{a}(2,1)$ perpetually. The proof for $\widehat{b}(1,1) > \widehat{b}(1,2)$ follows by symmetry. Therefore, there is a probability greater than 0 that both agents converge to action 1, and hence to any of the 4 pure strategy combinations (by symmetry).

Finally, we show that convergence to a mixed strategy Nash equilibrium is possible (if one exists). For a mixed strategy Nash equilibrium to exist we require, without loss of generality, $a(1,1) > a(2,1)$ and $a(2,2) > a(1,2)$ (and similarly for B). It follows from the previous proof that there is a positive probability that $\widehat{a}(1,1) > \widehat{a}(2,1)$ and $\widehat{a}(2,2) > \widehat{a}(1,2)$ perpetually (and similarly for B). It follows that in such cases, convergence to a pure strategy is only possible if it is a Nash equilibrium. If there are no pure strategy Nash equilibria then all actions will be infinitely explored and convergence to the mixed strategy is guaranteed. Convergence to a mixed strategy Nash equilibrium is hence possible, if one exists. This completes the sketch proof to our claim that the game can converge to 4 or 5 joint strategy profiles.

Without the genericity assumption, there can be infinitely many Nash equilibria strategies. Following the same reasoning as above, it can be seen that fictitious play with unknown rewards can converge to any of these Nash equilibria and also non-Nash strategies where at least one agent is playing a pure strategy. Therefore, the strategies cannot both converge to non-Nash mixed strategies, as otherwise the joint action space is infinitely explored, which would guarantee convergence to a Nash strategy.

6 Explorative vs. Non-explorative Strategies

In Sections 4 and 5 we analysed ILs and JALs with no exploration and proved that convergence to non-Nash pure strategies is possible. Now we consider the impact of introducing explorative actions. As defined earlier, we consider an ϵ-FP strategy for JALs (Type III) and an ϵ-greedy strategy for ILs (Type IV). First we consider only agent A selecting explorative actions for our two case games, to see whether a non-explorative player can be exploited.

6.1 Case Game 1: Matching pennies

• Type III: ϵ-FP vs. Type I: fictitious play

Figure 5 displays simulated results for JALs (i.e. ϵ-FP against fictitious play) for the matching pennies game of length 50 (with a relatively high noise variance of 1). In the left figure, agent A selects the predicted worst response 10% of the time and in the right figure 30%. In both cases agent A has received a higher reward than agent B (there are more matched actions than unmatched actions), despite this being a zero-sum symmetric game. Agent A has exploited agent B for two key reasons:

- The agents have no longer both converged to pure strategies where actions are not matched and agent A receives a low reward. This is the first benefit of exploration to agent A – the agent is no longer exploited by the opponent as the exploration causes agent A to learn that this is not a best response (compare with Figure 1 (bottom left) where agent A is occasionally exploited).
- Agent B, in the absence of any exploration, is still sometimes playing a pure strategy and agent A has learnt to exploit this (by matching) for close to $(1-\epsilon)$% of plays and gain a high reward. Moreover, notice that agent B selects pure strategies more often when $\epsilon = 0.3$. This feature can be attributed to the fact that agent A is playing a more mixed strategy (due to the added exploration) which gives agent B rewards close to 1 (rather than -1) more often. Consequently, the exploration of agent A prevents agent B from switching action as its predicted best response action is less likely to change.

The second benefit of exploration is of particular interest – exploiting agent B too often, such that it consistently receives a low utility, is more likely to make the agent switch action. Therefore agent A benefits from a high exploration parameter even if it has learnt the expected reward values. This feature can be viewed as agent A explicitly managing its mixed strategy to maintain long term rewards. Higher values of ϵ come at the cost of less frequent exploitation, but has the benefit that the opposing agent is easier to exploit. This exploitation yields a high utility to the explorative agent – far greater than the Nash equilibrium utility of 0.

To explain these results in more detail, the benefit of explicitly managing a mixed strategy can be derived from theoretical reasoning used in Section 5 (where we proved non-Nash convergence of non-explorative strategies). In particular, notice that there is a positive probability that a non-explorative agent never switches action and continuously selects a pure action. In such instances this positive probability for agent (B), is greater if the expected utility of the pure action is kept, with a higher likelihood, above a required level (which is $a(1,1)$ in Equation (8)) – which can be done by agent A exploring and selecting the suboptimal action. The tradeoff, however, is exploring too much such that player B is rewarded (above the Nash expected reward) despite continuously playing this pure action.

Figure 6 (left) displays the expected reward to agent A for $\epsilon \in [0,1]$, for a selection of noise variances, for the game of length 50. Agent A has benefited from exploring by having a positive reward in a symmetric zero-sum game, when $0 < \epsilon < 0.5$. It is easy to see that $\epsilon = 0.5$ will yield an expected reward of 0 for all noise variances as agent A selects each action exactly 50% of the time. For $\epsilon > 0.5$, agent A selects the predicted worst response more often than not and thus allows agent B to take a positive reward. The optimal value of ϵ is larger for small noise variances, which initially appears counter-intuitive. After all, lower noise variance corresponds

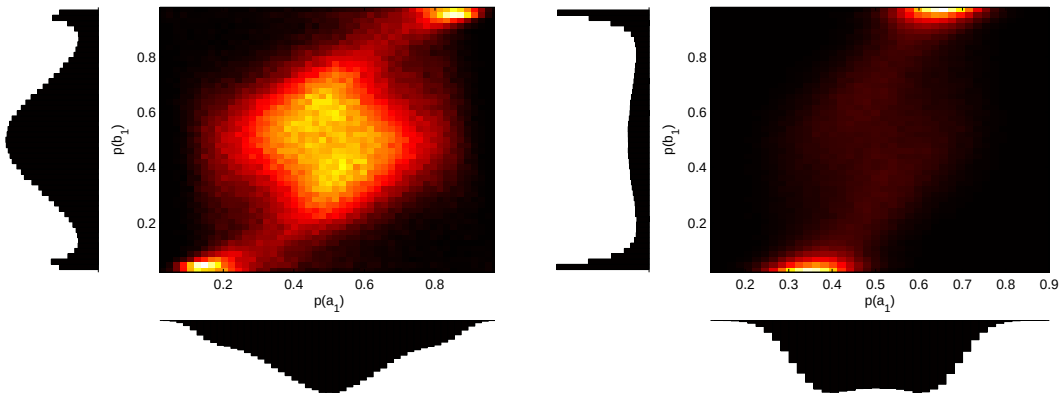

Figure 5: Density plots showing the proportion of times Action 1 is selected by agent A ($p(a_1)$) and agent B ($p(b_1)$) in the matching pennies game of length 50 over 100,000 repeats for a noise variances of 1, where agent B is a fictitious player and agent A is playing ϵ-FP with ϵ=0.1 (left) and ϵ=0.3 (right).

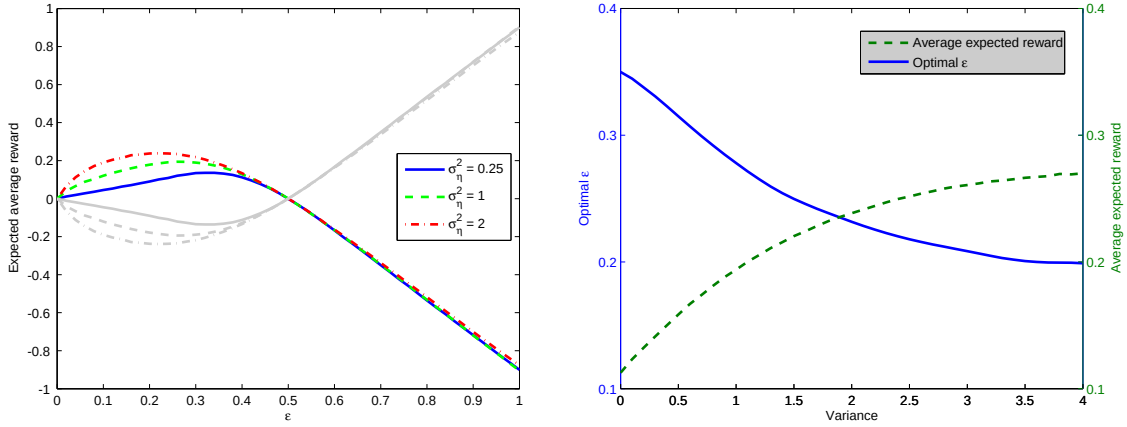

Figure 6: ϵ/-FP against fictitious play in the matching pennies game of length 50. (left) The Average expected reward to agent A and agent B (shaded) for the range of ϵ values for different noise variances and (right) Optimal ϵ and corresponding average expected reward over a range of noise variances.

to an easier learning problem. It must be remembered, however, that the only way agent A can gain a positive reward is to keep agent B on a pure strategy. For low noise variances this is hard to do (see Figure 1 (top left)), as the opponent quickly learns the correct best response of a mixed strategy. Agent A therefore has to keep its strategy very mixed in order to keep agent B's predicted response on the pure action. Conversely, large variances make agent B's predicted best responses more erroneous. Agent A can afford to exploit this more often and hence gain a higher expected reward, by playing with a smaller ϵ value.

Figure 6 (right) displays the optimal value of ϵ for a range of noise variances, along with the corresponding average expected rewards. The pattern emerges that higher noise variances, correspond to lower optimal ϵ, which in turn correspond to higher potential rewards. Note that the optimal value of ϵ is of course unknown to the agent *a priori*, however the theoretical derivation of approximated expected rewards in Section 9 can be used to construct methods for the agent to tune the value of ϵ as it plays.

- **Type IV: ϵ-greedy vs. Type II: bandit greedy**

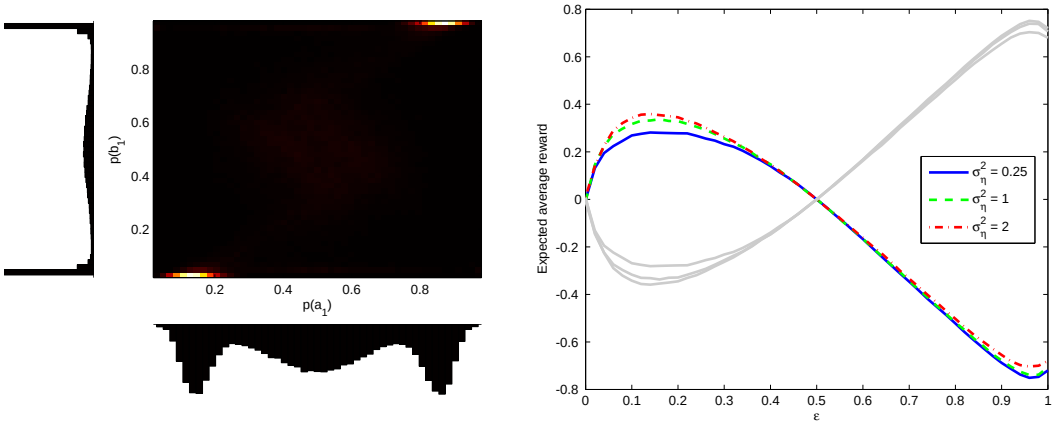

Figure 7: (left) Density plot showing the proportion of times Action 1 is selected by agent A ($p(a_1)$) and agent B ($p(b_1)$) in the matching pennies game of length 50 over 100,000 repeats for a noise variances of 0.25, where agent B is playing bandit greedy and agent A is playing ϵ-greedy with ϵ=0.1. (right) Average expected reward to agent A and agent B (shaded) for the range of ϵ values for different noise variances.

Figure 7 (left) considers the same matching pennies game except with ILs, where agent A uses ϵ-greedy (with $\epsilon = 0.1$) and agent B uses bandit greedy with no exploration. Agent A has again learnt to not be exploited with unmatched pure strategies and has learnt to often exploit agent B. The introduction of this exploration, however, has now introduced some convergence to mixed strategy equilibria (compare with Figure 2 (left)). Figure 7 (right) displays the expected reward to agent A for $\epsilon \in [0, 1]$, for a selection of noise variances. The properties of the results are similar to Figure 6 (left) for JALs in that the explorative agent can exploit with $0 < \epsilon < 0.5$. For ILs however, the optimal ϵ is smaller and less dependent on the noise variance. This lower value can be attributed to the fact that the greedy strategy learns more slowly than fictitious play and hence agent B can be exploited at a higher frequency without forcing it to change action.

- **Type III: ϵ-FP vs. Type II: bandit greedy**
- **Type IV: ϵ-greedy vs. Type I: fictitious play**

Finally, we show results for the same setup except in Figure 8 (left) agent A uses ϵ-FP and agent B uses bandit greedy and (right) agent A uses ϵ-greedy and agent B uses fictitious play. ϵ-FP already exploits bandit greedy when $\epsilon = 0$ (see Figure 2 (right)) and only benefits from a non-zero ϵ when the noise variance is high – so exploration is not always required to maximise reward. ϵ-greedy, however, is able to recover the deficit when $\epsilon = 0$ and exploit the fictitious player for certain values of $\epsilon < 0.5$. It can be concluded from these results that exploiting a non-explorative fictitious player requires careful management of the agent's mixed strategy (and hence a large ϵ) to prevent the opponent from switching strategies. In contrast, a bandit greedy strategy can be exploited with less exploration, as the opponent here is an IL and is therefore slower to learn that it should switch action. Nevertheless, with a high noise variance, the agent always benefits from some exploration in order to accurately learn the expected reward matrix.

6.2 Case Game 2: Prisoner's Dilemma

- **Type III: ϵ-FP vs. Type I: fictitious play**
- **Type IV: ϵ-greedy vs. Type II: bandit greedy**

We now briefly return to the Prisoner's Dilemma example. Figure 9 displays results for ϵ-FP against fictitious play and ϵ-greedy against bandit greedy. In both scenarios, agent A has a better reward than agent B, for low values of $\epsilon > 0$. For JALs the reward is maximised with $\epsilon = 0$

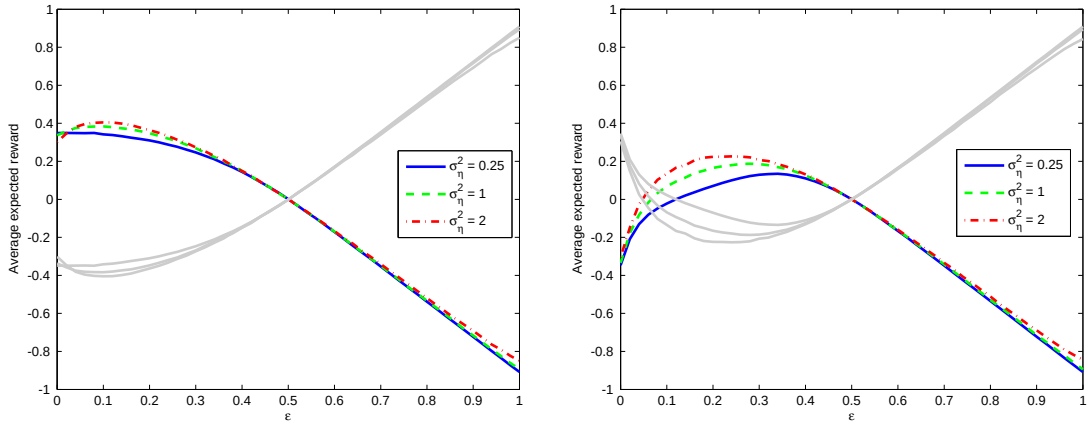

Figure 8: Average expected reward to agent A and agent B (shaded), in the matching pennies game, for the range of ϵ values for different noise variances where (left) agent A uses ϵ-FP and agent B uses bandit greedy and (right) agent A uses ϵ-greedy and agent B uses FP.

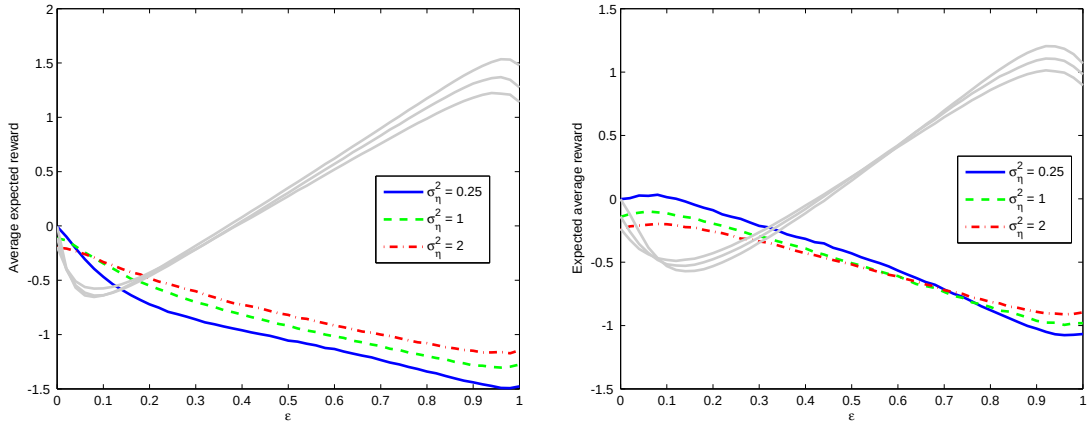

Figure 9: Average expected reward to agent A and agent B (shaded), in the prisoner's dilemma game, for the range of ϵ values for different noise variances where (left) agent A uses ϵ-FP and agent B uses fictitious play and (right) agent A uses ϵ-greedy and agent B uses bandit greedy.

and for ILs with $\epsilon \approx 0.1$. These lower optimal ϵ values (compared with matching pennies) can be attributed to the fact that agent B cannot be kept on the dominated strategy (cooperate) by playing mixed strategy. In addition, the prisoners dilemma is an unusual type of game where the Nash equilibrium is Pareto dominated – so learning and playing the true best response quickly results in lower rewards.

6.3 Summary

We have seen that exploring the action space, by playing the action that is estimated to perform worst, can in fact be beneficial to the agent's reward in finite time. An explorative agent can outperform a non-explorative agent as it learns the rewards from the action space more quickly and can then exploit the opponent, particularly if the opponent is stuck playing non-Nash strategies. The optimal value of ϵ, the exploration parameter, is quite varied depending on the structure of the game and the noise variance. In particular, if the game has a mixed strategy Nash equilibrium, then the agent can *exploit by exploring*, i.e. explicitly manage its mixed strategy, with a high ϵ value, to maintain long term rewards. Nevertheless, even with only pure strategy Nash equilibria, a small amount of exploration can still exploit a non-explorative learner.

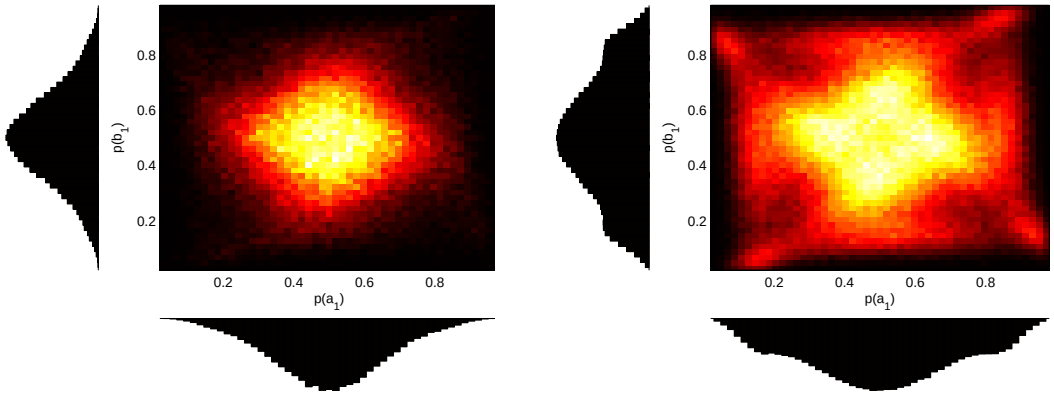

Figure 10: Density plots showing the average rewards for agent A and agent B in the matching pennies game of length 50 over 100,000 repeats for a noise variances of 1, where both agents are JALs using ϵ-FP (left) and both agents are ILs using ϵ-greedy (right). ϵ=0.1 for all strategies.

7 Explorative Strategies

In this section we consider both agents using explorative strategies. As documented in Claus and Boutilier (1998) and Chapman et al. (2009), a suitable exploration strategy will ensure convergence to a Nash equilibrium. The exploration parameter, ϵ, has to decay at a suitable rate, such that each action is infinitely explored but also action selections are greedy in the limit, i.e. the probability the maximal reward action is selected tends to 1 as $t \to \infty$. Such action selection policies are called *greedy in the limit with infinite exploration (GLIE)* (Singh et al., 2000). Chapman et al. (2009) show that fictitious play with ϵ-decreasing exploration (ϵ decaying at rate 1/t) will converge to a Nash equilibrium in any game with the fictitious play property (FPP). In contrast, Claus and Boutilier (1998) show that both ILs and JALs (using Boltzmann exploration) will converge to a Nash equilibrium in any cooperative game.

Our explorative strategies, ϵ-FP and ϵ-greedy, guarantee infinite exploration for $\epsilon > 0$, but are not greedy in the limit as the exploration parameter remains constant. We deliberately keep this parameter constant to maximise reward in finite time – refer to the previous section where we showed that an explorative agent can maximise reward by explicitly managing its mixed strategy profile throughout the game. In addition, decaying ϵ in finite time requires an additional decay parameter or function. When both agents explore and the game is sufficiently long, decaying ϵ makes sense, as the joint action space becomes thoroughly explored.

Nevertheless, even without a decaying exploration parameter, there is fast convergence towards the Nash equilibrium. See, for example, Figure 10 where 2 JALs (left) both using ϵ-FP (Type III) and 2 ILs (right) using ϵ-greedy (Type IV) both converge towards the mixed Nash equilibrium in the matching pennies game (Case Game 1). Without exploration (see Figure 1 (bottom left) and Figure 2 (left)), there exists only occasional convergence to the mixed Nash for JALs and none for ILs. Note that, as expected, the convergence of ILs appears slower than with JALs. There is clear evidence in each corner of the plot, that the agents are learning at a slower rate to move away from pure strategies and towards the mixed strategy Nash equilibrium.

8 The expected rewards of the strategies

The expected rewards of the strategies have thus far been approximated by Monte Carlo simulations. In this section we derive the theoretical expected reward as a function of the unknown rewards and noise variance. Expected rewards can then be computed more precisely and theoretical properties of the strategies can be found. In more detail, the expected reward at time t for all strategies is calculated by multiplying the probability of the next joint action being selected

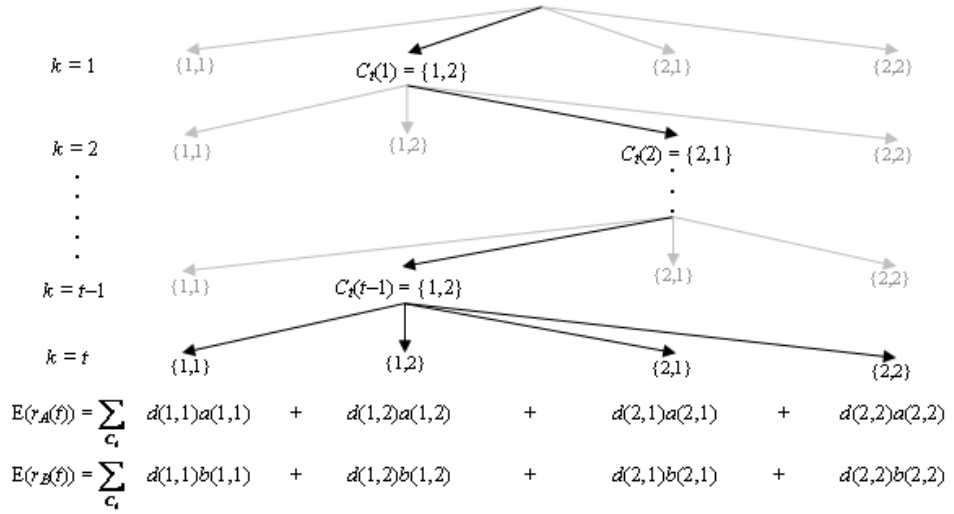

Figure 11: Computing the expected reward: The calculation cycles through all past sequences of joint actions and finds the probability of selecting each branching sequence. In our approximation, we prune the decision tree for unlikely sequences of joint actions.

with the corresponding expected reward, given all possible sequences of past joint actions i.e.:

$$\mathrm{E}(r_A(t)) = \sum_{C_t} D_{C_t} R_A \tag{11}$$

$$\mathrm{E}(r_B(t)) = \sum_{C_t} D_{C_t} R_B, \tag{12}$$

where,

$$D_{C_t} = \begin{bmatrix} d_{C_t}(1,1) & d_{C_t}(1,2) \\ d_{C_t}(2,1) & d_{C_t}(2,2) \end{bmatrix}, \qquad R_A = \begin{bmatrix} a(1,1) & a(1,2) \\ a(2,1) & a(2,2) \end{bmatrix}, \qquad R_B = \begin{bmatrix} b(1,1) & b(1,2) \\ b(2,1) & b(2,2) \end{bmatrix}.$$

C_t is the sequence of joint actions played up to time t. $d_{C_t}(i,j)$ represents the probability of joint action $\{i,j\}$ being played at time t along with the sequence of past joint actions C_t. This probability is dependent on the strategies played by both opponents, and its various derivations are detailed in this section.

The calculation of the expected reward therefore has to cycle through all permutations of C_t and for each one find the probability of selecting C_t and the next joint action – this process is illustrated in Figure 11. Notice that this computation scales exponentially in t, due to the exponentially growing number of past joint action sequences C_t. The results are therefore only useful for short games where the computation is still tractable.

To proceed, we derive the probabilities of selecting a joint action for the initial time-steps, and then induct forwards. At the start of the game, the agents initialise their strategies. As discussed earlier, we consider the simplest initialisation strategy where both players choose each action once, in a random order. As a result, with probability 0.5 the joint actions $\{1,1\}$ and $\{2,2\}$ have been selected (in no particular order) for the first two time-steps, and with probability 0.5, $\{1,2\}$ and $\{2,1\}$ have been selected (as is the case in Figure 11). We refer to these initialisation possibilities as *Initial1* and *Initial2* respectively. Therefore the expected reward to each agent for the first 2 time-steps, for all combinations of strategies, is trivially:

$$\mathrm{E}\left(r_A(1)\right) = \mathrm{E}\left(r_A(2)\right) = \frac{1}{4}\left(a(1,1) + a(1,2) + a(2,1) + a(2,2)\right), \tag{13}$$

$$\mathrm{E}\left(r_B(1)\right) = \mathrm{E}\left(r_B(2)\right) = \frac{1}{4}\left(b(1,1) + b(1,2) + b(2,1) + b(2,2)\right). \tag{14}$$

For both time-steps each joint action has equal weighting due to the randomised initialisation procedure.

Once initialisation is completed, the agents start selecting actions using the various strategies constructed in Section 3. We begin by considering both agents using fictitious play with no added exploration. The derivation of expected rewards for the other strategies is similar and follows afterwards.

8.1 Type I: fictitious play

With the fictitious play strategy, an agent's estimate of the reward attributed for a joint action is performed using recursive averaging and is formulated in (5). This estimate is an unbiased estimator of the true expected reward. As the noise process is a zero-mean Gaussian, the estimates of $a(i,j)$ also follow a Gaussian distribution:

$$\widehat{a}(i,j) \sim \mathcal{N}\left(a(i,j), \sqrt{\frac{\sigma_\eta^2}{n_{i,j}}}\right) \tag{15}$$

(and similarly for $b(i,j)$) where σ_η^2 is the variance of η_t. As the noise processes are i.i.d., the distributions of $\widehat{a}(i,j)$ and $\widehat{b}(i,j)$, for $i,j = 1,2$, will also be independent of each other.

Once the initialisation sequence has been completed, both agents start using the predicted best response to select actions. The probability of selecting the next action will no longer be equally weighted across the set of joint actions (unless all rewards are equal). At the third time-step, each agent will simply select the action which yielded the highest reward during initialisation. Consequently, the probability that agents A and B respectively select action 1 at the third time-step are:

$$\begin{aligned} P_1(3) =& \Phi\left(0, a(2,2) - a(1,1), 2\sigma_\eta^2\right), \\ Q_1(3) =& \Phi\left(0, b(2,2) - b(1,1), 2\sigma_\eta^2\right), \end{aligned} \tag{16}$$

with *initial1* and:

$$\begin{aligned} P_2(3) =& \Phi\left(0, a(2,1) - a(1,2), 2\sigma_\eta^2\right), \\ Q_2(3) =& \Phi\left(0, b(2,1) - b(1,2), 2\sigma_\eta^2\right), \end{aligned} \tag{17}$$

with *initial2*, where Φ is the Gaussian cumulative distribution function (CDF). These densities are Gaussian because the sum (or difference) of independent Gaussians is still Gaussian, i.e.:

$$\widehat{a}(2,2) - \widehat{a}(1,1) \sim \mathcal{N}\left(a(2,2) - a(1,1), 2\sigma_\eta^2\right), \tag{18}$$

due to the independence of the estimators and (15), and similarly for other densities. Notice that the probabilities of selecting each action are the same for the bandit greedy strategy. In fact, the probabilities and expected rewards of the two strategies will be identical only up to the third time-step. Beyond which the alternative learning policies can lead to different decisions. The expected reward to agent A at the third time-step, using both fictitious play and bandit greedy, therefore is:

$$\mathrm{E}\left(r_A(3)\right) = \sum_{i=1}^{2} D_i R_A \tag{19}$$

where,

$$D_i = \frac{1}{2}\begin{bmatrix} P_i(3)Q_i(3) & P_i(3)(1-Q_i(3)) \\ (1-P_i(3))Q_i(3) & (1-P_i(3))(1-Q_i(3)) \end{bmatrix}, \tag{20}$$

and similarly for agent B.

The probabilities of selecting the action sequences, D_{C_t}, become more difficult to calculate as the game continues beyond the third time-step. This is due to dependence on the decisions made and rewards observed at previous time-steps. The number of previous observations from

each joint action explicitly affect the probability of the next fictitious play decision, see (6). The probability of the next joint action, given the past sequence, is still a product of two independent Gaussians, due to the independence of the noise processes. This probability, however, is not independent on the past sequence of observations C_t. The entries of D_{C_t} are therefore a product of 2 independent multivariate Gaussian distributions (dimension $t-2$, as the first 2 joint actions follow the initialisation procedure outlined earlier):

$$D_{C_t} = \begin{bmatrix} P_i(t)Q_i(t) & P_i(t)\left(Q^*(t-1) - Q_i(t)\right) \\ \left(P^*(t-1) - P_i(t)\right)Q_i(t) & \left(P^*(t-1) - P_i(t)\right)\left(Q^*(t-1) - Q_i(t)\right) \end{bmatrix}, \tag{21}$$

where,

$$P_i(t) = \Phi_{t-2}(0, \mu_A(C_t), \Sigma_A(C_t)), \tag{22}$$

$$Q_i(t) = \Phi_{t-2}(0, \mu_B(C_t), \Sigma_B(C_t)), \tag{23}$$

and similarly for agent B, where Φ_{t-2} refers to the multivariate Gaussian CDF (dimension $t-2$). The derivation of the mean $\mu_A(C_t)$, covariance $\Sigma_A(C_t)$ and also $P^*(t-1)$ and $Q^*(t-1)$ can be found in Appendix I. Note that the entries of D_{C_t} can be calculated from $d_{C_t}(1,1)$ (and $D_{C_{t-1}}$) without having to recompute Gaussian densities.

The calculation of the expected reward at any time-step t is now formulated using equations (11) and (21). The calculation cycles through all possible (ordered) joint action sequences, C_t, and calculates the probability of this sequence together with the next joint action, D_{C_t}. This operation needs to be performed twice, once with each initialisation procedure. This summation therefore scales exponentially over time ($\mathcal{O}(4^{t-2.5})$) and is intractable for large t. Furthermore, for each joint action sequence, two $t-2$ multivariate Gaussian CDFs must be calculated. The multivariate Gaussian CDF has no closed-form solution and must be approximated numerically, and this also scales $\mathcal{O}(p^t)$ (where p is a constant, see Genz and Kahaner (1986)). Nevertheless, for short-length games, this calculation can be used to find the expected reward for the fictitious play strategy (Type I), for all 2×2 games and reward observation variances.

8.2 Other strategies

The calculation for the bandit greedy strategy (Type II) is identical, apart from the means and covariances of the multivariate Gaussian distributions, which differ due to the action selection rule. Their derivations are given in Appendix II. This calculation hence also scales with the same complexity.

The calculation of the expected rewards for ϵ-FP (Type III) and ϵ-greedy (Type IV) requires even heavier computation. This is because there are several possible decision processes that result in each sequence of past actions, C_t, due to the added exploration. At each stage, an agent might select action 1 (for example) as it is the best response (for JALs) or greedy choice (for ILs). Alternatively action 2 will be the best response/greedy choice but the agent explores and selects action 1 instead. There are therefore 2^{t-2} possible sequences of exploration/exploitation decisions, U, that lead to a sequence of past actions, C_t. The probability D_{C_t} is therefore the summation of the probabilities for each individual exploration/exploitation sequence. The exploration parameter is constant, therefore the number of explorative steps in a sequence of past actions follows a binomial distribution. For calculating D_{C_t} and hence the expected reward for both ϵ-FP and ϵ-greedy, replace (22) and (23) with the following:

$$P_i(t) = \sum_{i=0}^{t-2} \left(\epsilon^i (1-\epsilon)^{t-2-i} \sum_{S_U^{\text{unique}}} \Phi_{t-2}(0, \mu_A(C_t), \Sigma_A(C_t)) \right), \tag{24}$$

$$Q_i(t) = \sum_{i=0}^{t-2} \left(\epsilon^i (1-\epsilon)^{t-2-i} \sum_{S_U^{\text{unique}}} \Phi_{t-2}(0, \mu_B(C_t), \Sigma_B(C_t)) \right), \tag{25}$$

where S_U^{unique} refers to the unique set of permutations of U. Details of how $\mu_A(C_t)$ and $\Sigma_A(C_t)$ are calculated are given in Appendix III.

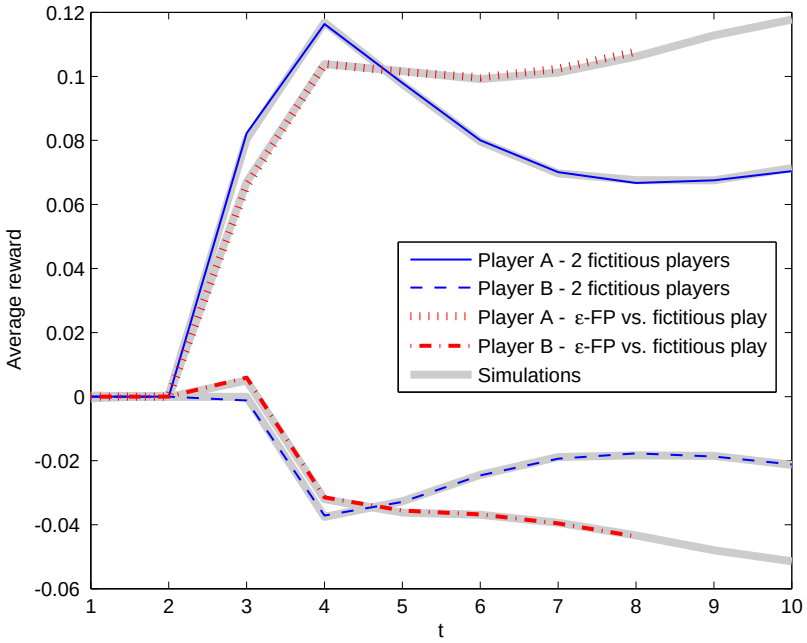

Figure 12: Average theoretical rewards for JALs in a 2×2 game. Average rewards from simulations are included for comparison.

We have derived the expected reward for all four types of strategies in self-play. The calculation of the expected rewards to each agent for any combination of these strategies is immediate. For example, for ϵ-FP against fictitious play then only replace $P_i(t)$ with (24) and use (23) for $Q_i(t)$. With all these results, we can compute exact expected rewards for any combination of our strategies, given the game configuration.

8.3 Examples

Figure 12 displays the expected rewards (up to $t = 10$) for JALs playing the following non-zero sum game with $\sigma_\eta^2 = \sigma_\nu^2 = 0.1$:

		Agent B	
		Action 1	Action 2
Agent A	Action 1	$-0.3, 0.2$	$0.5, -0.2$
	Action 2	$0.3, -0.4$	$-0.5, 0.4$

We consider this new game for better illustration, as our two case games have a symmetric reward structure and are hence going to yield identical rewards to each agent in self-play. Moreover, this game configuration has an unusual reward dynamic to each player, which is more challenging to approximate. In this asymmetric game, there is a mixed strategy Nash equilibrium at $\left\{\frac{2}{3}, \frac{5}{8}\right\}$, which yields a reward of 0 to both agents. We include results for JALs with both agents using fictitious play, and also agent A using ϵ-FP. Figure 13 displays results for ILs with both agents using bandit greedy and also agent A using ϵ-greedy. The average reward over 100,000 simulations is included in both figures for comparison. This is another example of a game where agent A can exploit agent B by using exploration, for both ILs and JALs, to gain a higher reward.

The theoretical analysis in this section has constructed calculations of the expected reward for our strategies without the need for averaging repeated simulations, although the solution is not closed-form as the multivariate Gaussian CDF can only be approximated numerically. The results only go as far as $t = 10$ (or $t = 8$ for explorative strategies) due to the computational demands of the calculation. Nevertheless, the formulation could be used to find the optimal ϵ,

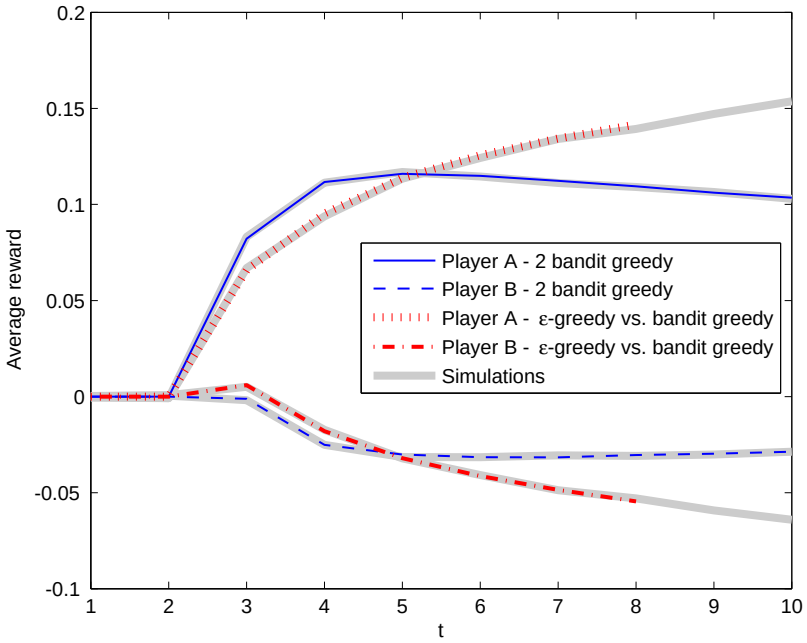

Figure 13: Average theoretical rewards for ILs in a 2×2 game. Average rewards from simulations are included for comparison.

for short length games, with which agent A most exploits agent B. For longer games, we require more computationally efficient approximations, which we construct in the next section.

9 Approximating the Expected Rewards

In the previous section, we formulated the exact expected rewards for both explorative and non-explorative strategies. These calculations scale exponentially in the length of the game. This section outlines proposed approximations of the expected rewards which scale linearly in the length of the game. This will allow efficient computation of optimal strategies for games of a much longer length. We make use of the structure of the multivariate Gaussian covariance matrix, for both ILs and JALs, to efficiently estimate cumulative probabilities. In addition, we prune nodes of the decision tree (see Figure 11 for reference) to bound the number of action paths, by using a technique analogous to particle filtering (Gordon et al., 1993). This removes both exponential components of the calculation to form a linear approximation.

9.1 Independent Learners

First consider ILs using the bandit greedy strategy (Type II). Notice in (45) of Appendix II that $\sigma^2_{k,t} = \sigma^2_{t-1,t}$ for $k \leq t$. We say that a symmetric matrix 'factors' when there exist vectors $[a_1, \ldots, a_t]$ and $[b_1, \ldots, b_t]$ such that $\Sigma_{i,j} = a_i b_j$ for $i \leq j$. Notice that if we set $a_i = 1$ and $b_j = \sigma^2_{j,j}$ then $\Sigma_A(C_t)$ does indeed factor. It follows that this is a sufficient condition for $\Sigma_A(C_t)$ to have the 'triangle property'. Specifically, a positive definite matrix Σ has the triangle property (Barrett and Feinsilver, 1978) if:

$$\Sigma_{i,j} = \frac{\Sigma_{i,k}\Sigma_{kj}}{\Sigma_{k,k}} \text{ for all } i < j. \tag{26}$$

This means all elements $\Sigma_{i,j}$ $(i < j)$ are determined by the main diagonal and the superdiagonal (i.e. the diagonal spanning $\Sigma_{1,2}, \ldots, \Sigma_{t-1,t}$). Furthermore, Barrett and Feinsilver (1978) proves

that Σ has the triangle property if and only if its inverse (or precision matrix) Σ^{-1} is tridiagonal. A tridiagonal matrix A is a matrix where $A_{i,j} = 0$ for $i < j - 1$ or $i > j + 1$.

The tridiagonal structure of $\Sigma_A(C_t)$ implies conditional independence between components of the multivariate Gaussian density that are 2 or more time-steps apart. Notice that the conditional independence between estimates resulting from recursive averaging was also exploited in Section 5 to find convergence properties of non-explorative strategies (by formulating the recursive averages as discretised Brownian Motion).

The conditional independence property can be used to compute cumulative probabilities more efficiently. Specifically, we do this linearly in the dimension of the distribution using a technique developed in Van Horn (2009), which outlines a method of sampling draws from the pdf of a t-variate Gaussian distribution linearly in t if the inverse covariance matrix is tridiagonal. Repeated samples can then be used to form a Monte Carlo approximation of the cumulative probabilities. This removes one exponential component from the calculation. In more detail, the draws from the t-variate pdf require knowledge of the inverse covariance matrix $\Sigma_A^{-1}(C_t)$. Rather than inverting the whole matrix ($\mathcal{O}(t^3)$), we can update the i-th row and column of $\Sigma_A^{-1}(C_i)$ using the following formula:

$$\sigma_{i,i}^{-1} = \frac{\sigma_{i-1,i-1}^2}{\sigma_{i-1,i-1}^2\sigma_{i,i}^2 - (\sigma_{i,i-1}^2)^2}, \tag{27}$$

$$\sigma_{i-1,i}^{-1} = \sigma_{i,i-1}^{-1} = \frac{-\sigma_{i,i-1}^2}{\sigma_{i-1,i-1}^2\sigma_{i,i}^2 - (\sigma_{i,i-1}^2)^2}, \tag{28}$$

where $\sigma_{i,j}^{-1}$ is the ijth component of $\Sigma_A^{-1}(C_i)$. Recall that $\sigma_{i,j}^{-1} = 0$ for $i < j - 1$ or $i > j + 1$. This update is therefore linear in t. We can then use these inverse covariance matrix components to iteratively generate densities $x_1, x_2, ...$ where $x_1, \ldots, x_{t-2}$ is a sample from the Gaussian pdf defined by $\mu_A(C_t)$ and $\Sigma_A(C_t)$ (given in (36) and (37) respectively). The iteration is given by:

$$x_i = \mu_A(t) + \mu_i + \frac{1}{\sqrt{\sigma_{i,i}^{-1}}}N \tag{29}$$

$$\mu_i = \begin{cases} 0 & \text{if} \quad i = 1 \\ -\frac{\sigma_{i,i}^{-1}}{\sigma_{i-1,i}^{-1}}x_{i-1} & \text{otherwise} \end{cases} \tag{30}$$

and similarly for B, where N is a draw from the standard Gaussian distribution. We can generate thousands of samples to formulate an estimate of the Gaussian CDFs defined in (22) and (23). The overall calculation however, even with this approximation, remains exponential – as we still have to search through the entire set of possible sequences of joint actions C_t. To counteract this issue, we restrict the number of Gaussian samples generated at each time-step. We do this by dividing the samples at time i between the four next possible joint actions (refer to Figure 11 to see the branching process). The division is weighted proportionately by the estimated probability of each joint action occurring. This is analogous to techniques used in particle filtering, where Monte Carlo samples are propagated sequentially and the particles are weighted according to their posterior probability. The difference being that rather than weighting particles, we divide them into distinct groups and propagate them on different action paths. The division of samples as the action space grows, means that certain joint action paths are not visited in the approximation, and others become approximated with few samples. This is effectively pruning nodes in the decision tree to reduce the search space. As t increases, the number of action sequences C_t will eventually exceed the number of Gaussian samples, at this stage the total number of joint action paths that can be considered is bounded above by the number of Gaussian samples. As the Gaussian samples are generated linearly in t, the overall calculation also scales linearly.

Figure 14 shows the approximation of rewards for ILs playing the game considered in Section 8. The efficient computation allows expected rewards to be approximated for longer games. Average simulation results are included for comparison. Note that the approximation for the ϵ-greedy strategy (Type IV) is a simple extension of bandit greedy as $\epsilon\%$ of particles need to be directed to the alternative action choice, independently at each time-step.

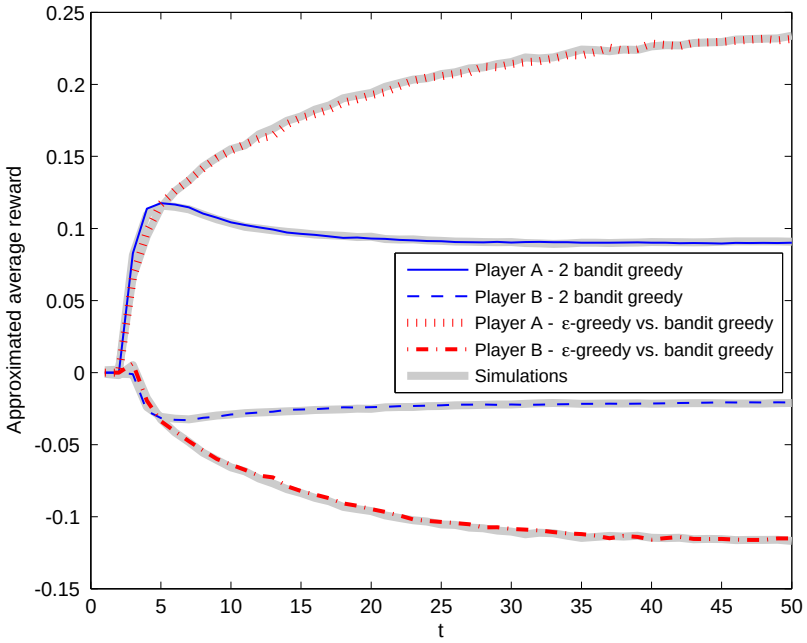

Figure 14: Approximated average rewards for ILs in a 2×2 game. Average rewards from simulations are included for comparison.

9.2 Joint Action Learners

We now consider formulating an approximation for JALs. Notice that if an opponent plays a random mixed strategy then the predicted best response using fictitious play is equivalent to the greedy choice using the bandit greedy strategy. Moreover, notice in (6) that if the opponent plays a pure strategy, the predicted best response is also equivalent to the bandit greedy selection. With all other opponent strategy profiles, the predicted best response and bandit greedy selection differ. This is because the JAL weights the reward estimates dependent on the past actions selected by the opponent, whereas the IL weights observations from each action equally. This weighting between rewards estimates ensures that the multivariate Gaussian density does not have the conditional independence structure found with ILs and hence does not always have a tridiagonal inverse. For pure and random mixed strategies however, the inverse is tridiagonal (from the equivalence to bandit problems). It can also be seen that for large t, the weightings attributed to past actions do not change significantly between time steps. This implies that, as t grows, each marginal density of the Gaussian moves in the direction of conditional independence with marginal densities that are 2 or more time steps away.

For these reasons, it is not unreasonable to approximate the multivariate densities by banding the inverse covariance matrix so that it is tridiagonal. Notice that, in general, inverting the covariance matrix and setting non-tridiagonal elements to zero will not always ensure positive-definite matrices (required for multivariate Gaussians). Nevertheless, recursively updating the tth row and column of the inverse matrix $\Sigma_A^{-1}(C_t)$ using (27) and (28) will ensure an approximated tridiagonal inverse that is positive-definite (see Bickel and Levina (2008) for details). The approximation is then exactly the same as before with ILs, using (29) and (30) to generate the Gaussian samples.

Figure 15 displays approximated expected rewards for JALs and also averaged simulation results for comparison. There is some bias in the approximation due to the banding of the inverse covariance matrix – however this error is bounded over time as the predicted best response update becomes more conditionally independent of decisions made two or more time-steps apart, as t grows. Therefore this approximation technique, which recursively updates Gaussian samples using only the previous values, can be used to approximate rewards for large t, in linear

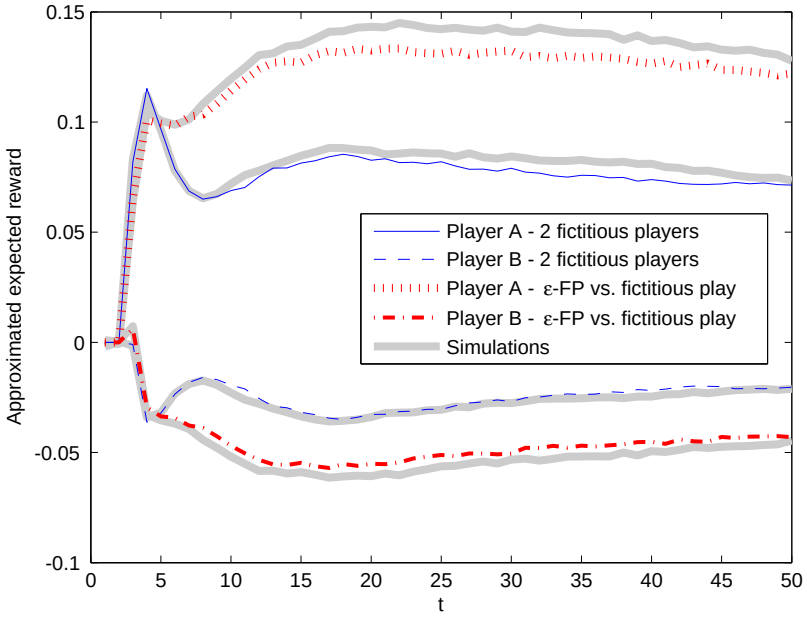

Figure 15: Approximated average rewards for JALs in a 2×2 game. Average rewards from simulations are included for comparison.

computation time.

10 Conclusions

In this work we have studied 2-player repeated 2×2 games where expected rewards are unknown *a priori*. The agents must learn as they play, and hence must simultaneously estimate rewards and adapt to the opponent. We investigated two fundamental learning techniques: Individual Learning (IL) and Joint Action Learning (JAL). Both ILs and JALs, when used with suitable exploration strategies, have been shown in Claus and Boutilier (1998) and Chapman et al. (2009) to converge to a Nash equilibrium for certain games. In this report, however, we showed that ILs (using the greedy strategy) and JALs (using fictitious play) with no exploration, have no such guarantee of converging to a Nash equilibrium – and hence are often suboptimal strategies. We sketch a proof demonstrating possible convergence to non-Nash pure strategies or any of the Nash equilibria strategies. We then constructed exploration strategies, based on the ϵ-greedy strategy from bandit literature, and showed that an agent can use this strategy to exploit a non-explorative opponent. We found surprisingly high optimal values of ϵ, for games with a mixed strategy Nash equilibrium, as the agent could *exploit by exploring* – or in other words explicitly manage its mixed strategy, to keep its opponent on a favourable pure strategy, and hence maintain long term rewards.

We also constructed theoretical approximations of the expected reward, in finite time, for both ILs and JALs. We first formulated the exact expected reward which scales exponentially, and is hence intractable for long-length games. We then constructed linear approximations by making use of conditional independencies inherent in the learning mechanisms. The rewards for IL strategies, could be approximated in an unbiased way as the conditional independencies are exact. The structure of JALs however, is such that the learning moves in the direction of conditional independence, resulting in approximations that are biased but perform better the longer the game runs.

Future work involves considering games with more actions or more agents, to try and generalise some of our findings. The need to explore will increase with a wider array of actions, and the opportunity to exploit opponents will be more available with multiple agents – although the

analysis and theoretical derivations are likely to be more complex. Some of our findings, such as the proof of convergence to non-Nash equilibria with non-explorative strategies, will extend naturally to a wider setting – where the number of possible joint strategies that agents can converge to is dependent on the number of agents and actions present. Specifically, this number (under the genericity assumption) is M^N plus the number of mixed strategy Nash equilibria, where M is the number of actions to an agent and N is the number of agents. The proof is structured using the same methods sketched in this report.

In addition, we could use the approximated expected rewards to construct on-line tuning of the exploration parameter ϵ – this is of particular interest as the optimal values of ϵ for our case games were found to be wide-ranging dependent on the structure of the game. Finally, we could used the theory of discretised Brownian Motion to find the exact probability of convergence to each non-Nash equilibrium in the absence of exploration. This probability will be dependent on the noise variance parameter, as well as the structure of the reward matrix. Furthermore, this probability would determine the need for the agents to learn by exploring their action decisions, for different classes of games.

In the context of Aladdin, this report provides theoretical insights into how agents should select strategies in an environment with multiple agents and unknown rewards, such as a disaster management scenario. In particular, we have proved the sub-optimality of not exploring and as a result have constructed strategies that explore the action space. Furthermore, we have demonstrated the benefit of exploring in certain environments using these strategies and moreover, constructed techniques for approximating rewards to each agent based on the exploration parameter, such that it can be optimised by an agent to gain a high utility.

In a wider context, our findings potential application in many Aladdin technologies – in particular technologies that are used in environments with a dynamic or uncertain nature. In addition, Aladdin systems are usually decentralised or distributed, where agents are uncertain of the optimality of their future action choices. In this report we use a similar framework which is also decentralised, uncertain and noisy. Consequently our findings may be of use to Aladdin technologies, in particular because it highlights the importance of agents exploring their actions. The effects of exploration in Aladdin systems are not yet fully understood, hence our findings could be used to improve the performance of these systems as a whole, and not just the performance of individual agents.

To this end, the techniques developed in this report can be used to model interactions between agents in decentralised systems, such that the optimal level of exploration can be found. Consider, for example, two ambulances (the agents) having to simultaneously choose which building to attend to, in a disaster management scenario. In a decentralised system, the agents might have little or no prior knowledge about the number of casualties in each building. Furthermore, each agent is unaware of the decisions made by the other, such that there is a need for coordinated actions. If an agent does not perform any exploration (ie. restrict attention to a particular building), then (all other things equal) it is likely this agent will perform worse than an agent that does explore – as demonstrated by the results in this report.

The techniques used in this report are based on game theoretic principles. Traditional game theoretic analysis considers non-noisy predictable environments where exploration of actions is not so important. Our setting does not make these assumptions however and therefore allows the game theoretic analysis of multi-agent interactions to be applicable in wider and more practical settings. In conclusion, the games with unknown rewards framework, and the strategies and techniques constructed in this report, could greatly benefit many Aladdin technologies.

References

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2-3):235–256, 2002.

M. Babes, M. Wunder, and M. Littman. Q-learning in Two-Player Two-Action Games. *Proceedings of 8th International Conference on Autonomous Agents and Multiagent Systems*, 2009.

W. Barrett and P. Feinsilver. Gaussian Families and a Theorem on Patterned Matrices. *Journal of Applied Probability*, 15(3):514–522, 1978.

U. Berger. Fictitious Play in $2 \times N$ Games. *Journal of Economic Theory*, 120(2):139–154, 2005.

P. Bickel and E. Levina. Regularized Estimation of Large Covariance Matrices. *Annals of Statistics*, 36(1):199–227, 2008.

G. Brown. Iterative Solution of Games by Fictitious Play. *Activity Analysis of Production and Allocation*, 13:374–376, 1951.

J. Chang. *Lecture Notes on Stochastic Processes*. 1999.

A. Chapman, D. Leslie, and D. Flores. Convergent Learning Algorithms for Potential Games with Unknown Perturbed Rewards. *Technical Report*, 2009.

C. Claus and C. Boutilier. The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems. *Proceedings of the National Conference on Artificial Intelligence*, pages 746–752, 1998.

A. Dixit, S. Skeath, and J. Repcheck. *Games of Strategy*. Norton New York, 2004.

D. Fudenberg and E. Maskin. The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. *Econometrica*, 54(3):533–554, 1986.

A. Genz and D. Kahaner. The Numerical Evaluation of Certain Multivariate Normal Integrals. *Journal of Computational and Applied Mathematics*, 16(2):255–258, 1986.

N. Gordon, D. Salmond, and A. Smith. Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation. 140(2):107–113, 1993.

S. Govindan and R. Wilson. A Decomposition Algorithm for N-player Games. *Economic Theory*, pages 1–21, 2010.

L. Kaelbling. *Learning in Embedded Systems*. MIT press, 1993.

R. Luce. *Individual Choice Behavior*. Wiley New York, 1959.

J. Marden, G. Arslan, and J. Shamma. Regret Based Dynamics: Convergence in Weakly Acyclic Games. *Proceedings of the 6th international joint conference on Autonomous Agents and Multiagent Systems*, page 42, 2007.

J. Marden, H. Young, G. Arslan, and J. Shamma. Payoff Based Dynamics for Multi-Player Weakly Acyclic Games. *SIAM Journal on Control and Optimization*, 48(1):373–396, 2009.

D. Monderer and L. Shapley. Potential games. *Games and Economic Behavior*, 14(1):124–143, 1996.

J. Nachbar. Evolutionary Selection Dynamics in Games: Convergence and Limit Properties. *International Journal of Game Theory*, 19(1):59–89, 1990.

N. Pavlidis, D. Tasoulis, and D. Hand. Simulation Studies of Multi-Armed Bandits with Covariates. *Proceedings of the 10th International Conference on Computer Modeling and Simulation*, pages 493–498, 2008.

V. Pruzhansky. On Finding CURB Sets in Extensive Games. *International Journal of Game Theory*, 32(2):205–210, 2003.

S. Ramchurn, A. Rogers, K. Macarthur, A. Farinelli, P. Vytelingum, I. Vetsikas, and N. Jennings. Agent-Based Coordination Technologies in Disaster Management. *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems: Demo Papers*, pages 1651–1652, 2008.

A. Rogers, R. Dash, S. Ramchurn, P. Vytelingum, and N. Jennings. Coordinating Team Players within a Noisy Iterated Prisoners Dilemma Tournament. *Theoretical Computer Science*, 377 (1-3):243–259, 2007.

S. Singh, T. Jaakkola, M. Littman, and C. Szepesvari. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38(3):287–308, 2000.

R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT press, 1998.

K. Van Horn. Efficient Computation of Statistics for Banded Multivariate Normal Distributions. *Technical Report*, 2009.

J. Vermorel and M. Mohri. Multi-Armed Bandit Algorithms and Empirical Evaluation. *Lecture Notes in Computer Science*, 3720:437–448, 2005.

C. Watkins. *Learning from Delayed Rewards*. Cambridge University, 1989.

C. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3):279–292, 1992.

H. Young. The Evolution of Conventions. *Econometrica*, 61(1):57–84, 1993.

Appendix

Appendix I

To calculate $P^*(t-1)$ and $Q^*(t-1)$:

$$P^*(t-1) = \frac{E_{t-1}(C_t^A(t-1), 1)}{Q_i(t-1)}, \tag{31}$$

$$Q^*(t-1) = \frac{E_{t-1}(1, C_t^B(t-1))}{P_i(t-1)}. \tag{32}$$

This ensures that the Gaussian CDFs for these action sequences can be calculated from the CDFs used to calculate $d_{C_t}(1,1)$. This involves storing the last probability matrix $D_{C_{t-1}}$, which is more efficient than recomputing CDFs.

To calculate the Gaussian CDFs $P_i(t)$ and $Q_i(t)$:

$$P_i(t) = \Phi_{t-2}(0, Z_t.^*\mu_A(C_t), (Z_t^T Z_t).^*\Sigma_A(C_t)), \tag{33}$$

$$Q_i(t) = \Phi_{t-2}(0, Z_t.^*\mu_B(C_t), (Z_t^T Z_t).^*\Sigma_B(C_t)), \tag{34}$$

$$\mu_A(C_t) = \begin{bmatrix} \mu_A(C_{t-1}) \\ \sum_{i,j} \text{sign}(n_{i,j}^t)(2i-3)a(i,j)(n_{i,j}^t + n_{i^c,j}^t) \end{bmatrix}, \tag{35}$$

$$\Sigma_A(C_t) = \begin{bmatrix} \Sigma_A(C_{t-1}) & W_t^T \\ W_t & \sigma_{t,t}^2 \end{bmatrix}, \tag{36}$$

$$\sigma_{t,t}^2 = \sigma_\eta^2 \left(\sum_{i,j} \text{sign}(n_{i,j}^t)^2 \frac{(n_{i,j}^t + n_{i^c,j}^t)^2}{n_{i,j}^t} \right), \tag{37}$$

$$W_t = \begin{bmatrix} \sigma_{t,3}^2 & \dots & \sigma_{t,t-1}^2 \end{bmatrix}, \tag{38}$$

$$\sigma_{t,k}^2 = \sigma_\eta^2 \left(\sum_{i,j} \text{sign}(n_{i,j}^k)^2 \frac{(n_{i,j}^k + n_{i^c,j}^k)(n_{i,j}^t + n_{i^c,j}^t)}{n_{i,j}^t} \right), \tag{39}$$

$$Z_t(t-1:t) = [1\ 1] \text{ if } C_t^A(t-1) = 1 \tag{40}$$

$$Z_t(t-1:t) = [-1\ 1] \text{ if } C_t^A(t-1) = 2, \tag{41}$$

and similarly for agent B, where if $i = 1$ then $i^c = 2$ and vice-versa. Note that the $\text{sign}(n_{i,j}^k)$ term is required in the mean and covariance matrix calculation to ensure components of the mean and covariance where $n_{i,j}^k = 0$ (an unexplored joint action) are calculated as zero. The Z_t vector is required to find the correct portion of the cumulative distribution for each Gaussian, this is set as -1 when $C_t^A(t-1) = 2$ as agent A has last selected action 2 (and the inequality in (6) reverses). The equations are setup such that the mean and covariance can be updated recursively, when using a depth-first search of the joint action space. Furthermore, the mean and covariances of previous joint action sequences do not need to be stored for new sequences, using this method.

Appendix II

To calculate the expected reward for the bandit greedy strategy, replace (36), (38) and (40) with:

$$\mu_A(C_t) = \begin{bmatrix} \mu_A(C_{t-1}) \\ \sum_{i,j} (2i-3) n_{i,j}^t a(i,j) / (n_{i,j}^t + n_{i,j^c}^t) \end{bmatrix}, \tag{42}$$

$$\sigma_{t,t}^2 = \sigma_\eta^2 \left(\sum_i 1/(n_{i,1}^t + n_{i,2}^t) \right), \tag{43}$$

$$\sigma_{t,k}^2 = \sigma_{t,t}^2. \tag{44}$$

This reflects the different way that past actions are weighted in the decision process.

Appendix III

To calculate the expected reward for explorative strategies, replace (34) and (35) with:

$$P_i(t) = \sum_{i=0}^{t-2} \left(\epsilon^i (1-\epsilon)^{t-2-i} \sum_{S_U^{\text{unique}}} \Phi(0, V.^* \mu_A(C_t), (V^T V).^* \Sigma_A(C_t)) \right), \tag{46}$$

$$Q_i(t) = \sum_{i=0}^{t-2} \left(\epsilon^i (1-\epsilon)^{t-2-i} \sum_{S_U^{\text{unique}}} \Phi(0, V.^* \mu_B(C_t), (V^T V).^* \Sigma_B(C_t)) \right), \tag{47}$$

$$V = U.^* Z_t \tag{48}$$

$$U = [U_1 U_2], \tag{49}$$

$$U_1 = [-1, \ldots, -1] \text{ (length } i), \tag{50}$$

$$U_2 = [1, \ldots, 1] \text{ (length } t-2-i). \tag{51}$$

S_U^{unique} refers to the unique set of permutations of U, of which there are $\binom{t}{i}$ for each i and 2^{t-2} in total. The vector U changes the sign of the mean and covariances for entries that are -1. These entries represent explorative time-steps where the predicted best response (or the greedy selection for ILs) indicated that the alternative action should be selected.