CONVERGENT LEARNING ALGORITHMS FOR UNKNOWN REWARD GAMES *

ARCHIE C. CHAPMAN†, DAVID S. LESLIE‡, ALEX ROGERS§, AND NICHOLAS R. JENNINGS¶

Abstract. In this paper, we address the problem of convergence to Nash equilibria in games with rewards that are initially unknown and must be estimated over time from noisy observations. These games arise in many real–world applications, whenever rewards for actions cannot be prespecified and must be learnt on–line, but standard results in game theory do not consider such settings. For this problem, we derive a multi–agent version of Q–learning to estimate the reward functions using novel forms of the ε–greedy learning policy. Using these Q–learning schemes to estimate reward functions, we then provide conditions guaranteeing the convergence of adaptive play and the better–reply processes to Nash equilibria in potential games and games with more general forms of acyclicity, and of regret matching to the set of correlated equilibria in generic games. A secondary result is that we prove the strong ergoditicity of stochastic adaptive play and stochastic better–reply processes in the case of vanishing perturbations. Finally, we illustrate the efficacy of the algorithms in a set of randomly generated 3-player coordination games, and show the practical necessity of our results by demonstrating that violations to the derived learning parameter conditions can cause the algorithms to fail to converge.

Key words. Potential games, distributed optimisation, reinforcement learning, strong ergodicity.

AMS subject classifications. 91A10, 68T05, 68W15

1. Introduction.

The design and control of large, distributed systems is a major engineering challenge. In particular, in many scenarios, centralised control algorithms are not applicable, because limits on the system's computational and communication resources make it impossible for a central authority to have complete knowledge of the environment and direct communication with all of the system's components [Jennings, 2001]. In response to these constraints, researchers have focused on decentralised control mechanisms for such systems.

In this context, a class of noncooperative games called *potential games* [Monderer and Shapley, 1996] have gained prominence as a design template for decentralised control in the distributed optimisation and multi–agent systems research communities. Potential games have long been used to model congestion problems on networks [Wardrop, 1952; Rosenthal, 1973]. However, more recently, they have been used to design decentralised methods of solving large–scale distributed problems, such as power control and channel selection problems in ad hoc wireless networks [Scutari et al., 2006], task allocation, coverage and scheduling problems [Arslan et al., 2007; Chapman et al., 2010] and distributed constraint optimisation problems [Chapman et al., 2011]. In more detail, given a global target function, a potential game is constructed by distributing the system's control variables among a set of *agents* (or players), and each agent's reward function is derived so that it is *aligned* with the system–wide goals. That is, an agent's reward increases only if the global reward increases (as in Wolpert and Tumer [2002]). If the agents' rewards are perfectly aligned with the global target function, then the global target function is a *potential* for the game, which, in turn, implies that the (pure) Nash equilibria of the game are local optima of the global target function.

*Part of this work was carried out as part of the ALADDIN (Autonomous Learning Agents for Decentralised Data and Information Systems) project, which was jointly funded by a BAE Systems and EPSRC (Engineering and Physical Sciences Research Council, UK) strategic partnership (EP/C548051/1). A. Chapman was supported by a University of Sydney Business School Postdoctoral Fellowship.

† School of Electrical and Information Engineering, and Discipline of Business Analytics, University of Sydney Business School, Sydney, NSW 2006, Australia (archie.chapman@sydney.edu.au).

‡School of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, United Kingdom (david.leslie@bristol.ac.uk).

§Electronics and Computer Science, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom (acr@ecs.soton.ac.uk).

¶Electronics and Computer Science, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom (nrj@ecs.soton.ac.uk).

Given this framework for distributing an optimisation problem, the second problem facing a designer of a decentralised optimisation method is specifying a distributed algorithm for computing a solution. This is addressed by the literature on *learning in games*; the dynamics of learning processes in repeated games is a well investigated branch of game theory (see Fudenberg and Levine [1998], for example). In particular, the results that are relevant to this work are the guaranteed convergence to a Nash equilibrium in potential games of adaptive play and the broad class of finite–memory better–reply processes [Young, 2004, 1993], and of regret matching to correlated equilibrium [Hart and Mas-Colell, 2000, 2001]. Thus, decentralised solutions to an optimisation problem can be found by, first, deriving a potential game, and then using one of these algorithms to compute a Nash or correlated equilibrium.

There is, however, one major shortcoming to this model. As is standard in game theory, there is an assumption that the value of each configuration of variables, or the agents' rewards for different joint action profiles, is known from the outset. Although this is a sound assumption in some domains, in many of the large, distributed control application domains to which the decentralised control methods described above are targeted, it is not realistic to assume that the rewards for different variable configurations can be prespecified. For example, in many monitoring and coverage problems, the system's task is to learn about the phenomena under observation, but the rewards earned by the agents in the system are a function of the phenomena detected, so cannot be known before they are deployed. Similarly, latencies on a newly constructed or reconfigured ad–hoc network, which drive users' routing policies, may be initially unknown, and can be estimated only after observing the network's traffic flows.

Against this background, we address the problem of distributed computation of equilibria in games with rewards that are initially unknown and which must be estimated online from noisy observations. The algorithms derived allow agents to effectively learn their reward functions while coordinating on an equilibrium. Because of their links to distributed optimisation, we place particular focus on convergence to pure Nash equilibria in potential games with unknown noisy rewards, but also derive an algorithm that converges to correlated equilibrium in generic games. The adaptive processes we derive simultaneously perform: (i) the recursive estimation of reward function means using Q–learning [Sutton and Barto, 1998] employing a novel greedy–in–the–limit–with–infinite–exploration (GLIE) randomised learning policy [Singh et al., 2000] constructed for multi–agent problems, and (ii) the adjustment to the strategies of others in the game using one of the learning processes enumerated above, namely, adaptive play, better–reply processes, or regret matching.

Although Q–learning and the action adaptation processes above are well understood independently, the combined problem of learning the equilibria of games with unknown noisy reward functions is less well understood, and it is this shortcoming that we address. Specifically, the main theoretic results in this paper are:

1. We derive a novel multi–agent version of Q–learning with GLIE ε–greedy learning policies for which reward estimates converge to their true mean value. We use this as a component of our novel action adjustment processes below.
2. As a preliminary step to our main results, we prove, for the first time, the strong ergodicity of stochastic adaptive play and better–reply processes with vanishing choice perturbations in games with known rewards.
3. We prove the convergence of novel variants of adaptive play and the better–reply processes, employing Q–learnt rewards and a GLIE ε–greedy learning policy rule, to Nash equilibrium in repeated potential games with unknown noisy rewards (and in other more general classes of acyclic games specific to each process).
4. We prove the convergence of a novel variant of the regret matching algorithm, employing Q–learnt rewards and a GLIE ε–greedy learning policy rule, to the set of

correlated equilibrium in generic games with unknown noisy rewards.

One drawback of the Q–learning scheme we derive is that the size of the learning problem faced by the agents grows exponentially with the number of players, thereby reducing the usefulness of our algorithms in large games. To tackle this, we provide similar results to those above for games that can be encoded in two common compact graphical representations, *graphical normal form* and *hypergraphical normal form* [Kearns et al., 2001; Gottlob et al., 2005; Papadimitriou and Roughgarden, 2008]. Specifically, we show how the sparse interaction structure that these representations encode can be exploited to derive efficient exploration policies for Q–learning, such that the learning problem facing the agents is significantly reduced.

In addition to the main theoretical contribution of the paper, we empirically evaluate the algorithms in a simple three player potential game with unknown noisy rewards. By so doing, we seek to demonstrate the efficacy of the algorithms in solving these problems, and their advantages over other distributed methods previously proposed for these problems (such as Claus and Boutilier [1998], Cominetti et al. [2010]). We also demonstrate the necessity of the conditions on the learning parameters derived in our convergence results, by showing that if they are violated, the algorithms are more likely to fail to converge.

The paper progresses as follows: The next section contrasts our contributions to existing work in the area of algorithms for games with unknown and/or noisy rewards. Section 3 covers the necessary game–theoretic background material and formally states the problem addressed by the paper. In Section 4 we derive our multi–agent versions of Q–learning and the ε–greedy policy. Section 5 presents the main action adaptation process convergence results. Following these theoretical results, in Section 6 we compare the performance of the algorithms in a test domain. Section 7 summarises the paper and discusses how our results may be extended to further algorithms.

2. Related work. Several authors have previously tackled the problem of learning Nash equilibria in games with unknown noisy rewards by applying Q–learning based approaches. Most closely related to our work is that of Claus and Boutilier [1998], who specify a *joint action learner* (JAL), in which each agent keeps track of the frequency of other agents' actions, as in fictitious play [Brown, 1951; Fudenberg and Levine, 1998; Leslie and Collins, 2006], while updating the reward estimate for the joint action played. However, the authors do not provide convergence conditions for their algorithm, in that they do not investigate the sampling probabilities required to ensure that the reward function estimates converge, nor do they make the link between the convergence of these estimates and convergence of the actions played to Nash equilibrium. Their investigation relies instead on experimental evidence of convergence, and, furthermore, it is restricted to team games (games with a common payoff function), whereas we consider several further classes of games. Additionally, other authors consider *independent action learners* (IAL), in which agents use variants of the Q–learning procedure independent of each other, oblivious of the effects of changes in other agents' actions on their own payoffs. In particular, under the IAL processes of Claus and Boutilier [1998], Leslie and Collins [2005] and Cominetti et al. [2010], the agents update their estimate of the reward they receive for each of their actions, independent of the other agents, using Q–learning. These algorithms all use a Boltzmann distribution to select actions, but differ in the specific manner in which this is used, with Claus and Boutilier [1998] specifying an annealing schedule for the temperature coefficient and Leslie and Collins [2005] and Cominetti et al. [2010] using a constant temperature. However, none of these works prove convergence to Nash equilibrium; indeed with a constant temperature it is impossible to generically achieve convergence to Nash equilibrium. In Section 6, we demonstrate the superior performance of our algorithms over the JAL and IAL processes discussed above.

Single–agent learning in unknown noisy game environments has also been investigated in the context of zero–sum games. In particular, Baños [1968] considers two–player zero–sum games, in which one agent does not know the payoffs and receives only a noisy observation of the mean payoff for the action it plays each time a move is made. The author derives a class of strategies for this player that perform as well asymptotically as if the player had known the mean payoffs of the games from the outset. Auer et al. [1995] consider an adversarial multi–armed bandit (MAB) problem, in which an adversary has control of the payoffs of each of the MAB's arms and aims to minimise the player's payoff (these games contain the zero–sum games studied by Baños [1968] as a subclass). The authors provide an algorithm for general multi–player games that asymptotically guarantees a player its maximin value. For two–player zero–sum games, this is the same guarantee as the strategy derived by Baños, however the authors also show that their algorithm is more efficient than that of Baños. Both of these approaches converge to a Nash equilibrium only in 2–player zero–sum games (where the Nash equilibrium, minimax, and maximin concepts give the same solution). Thus, they do not apply to multi–player and/or potential games.

Evolutionary approaches to learning in games with noisy reward functions have also been investigated, which draw conclusions similar to ours regarding the long–run stability of Nash equilibria. For example, Mertikopoulos and Moustakas [2010] consider a continuous–time evolutionary learning procedure in a noisy game, reminiscent of JAL (discussed above), and show that under this process, the game's strict Nash equilibria is asymptotically stable. Similarly, Hofbauer and Sandholm [2007] consider evolutionary better–reply learning in population games with noisy payoffs, and derive a process that converges to approximate Nash equilibrium in stable games, potential games, and supermodular games.

Several algorithms have been proposed for games where agents cannot monitor their opponents' actions, so the payoffs that they receive appear to be randomised as the other players' actions change. This is a different scenario to the situation we consider: agents' payoffs are corrupted by noise that is induced by their opponents' unobservable switches in actions, whereas our work considers noise in rewards that is caused by some exogenous random perturbation under the assumption that opponents' actions can be observed. One approach to games with unobservable actions is *modified regret matching* [Hart and Mas-Colell, 2000], for situations where the agents do not know the payoffs and cannot observe their opponents' actions. Asymptotic play of this algorithm is guaranteed to be in the set of correlated equilibria in all generic games. A second relevant approach to games with unknown rewards and unobserved opponent actions is given in Marden et al. [2009], who investigate payoff–based dynamics that converge to pure–strategy Nash equilibria in weakly acyclic games, one of which, *sample experimentation dynamics*, can admit perturbations in agents' rewards. This algorithm alternates between two phases — exploration and exploitation. However, it requires that several parameters are set in advance, which control the exploration phase length, exploration rates, and tolerances on payoff difference and switching rates for deciding when to change strategies. These parameters depend on the problem at hand, and if they are incorrectly set, then the algorithm may fail to converge. This means that a user must have sufficient *a priori* knowledge of the problem at hand or set them in a conservative manner, which slows the rate of convergence.

The only algorithms proven to converge, in some sense, to a Nash equilibrium in all games are the *regret–testing* algorithms of Foster and Young [2006] (see also Young [2009]). These algorithms will stay near a Nash equilibrium for a long time once it has been reached, but perform what is essentially a randomised exhaustive search to find an equilibrium in the first place. We sacrifice this convergence in all games in order to improve the rate of convergence in the games we are interested in (i.e. classes of games directly associated with

distributed optimisation problems).

Finally, while our results rely on conditions for products of stochasitc matrices to be strong ergodic derived by Anily and Federgruen [1987], we note that recent results by Touri and Nedić [2010] may also be employed to the same end.

3. Game theory preliminaries. This section covers noncooperative games, acyclicity properties for games, and games with unknown and noisy rewards. Throughout, we use $\mathbb{P}(\cdot)$ to denote the probability of an event occurring.

3.1. Noncooperative games. We consider repeated play of a finite noncooperative game $\Gamma = \langle N, \{A_i, r_i\}_{i\in N}\rangle$, where $N = \{1,\dots,n\}$ is a set of agents, A_i is the set of actions of agent i, and $r_i : \times_{i\in N} A_i \to \mathbb{R}$ is i's reward function. Let $A = \times_{i\in N} A_i$ be the set of all joint actions (also called outcomes), and $\mathbf{a} \in A$ be a particular joint action. Agents can choose to play according to a distribution over pure actions, $\sigma_i \in \Delta(A_i)$, known as a *mixed strategy*, where each element $\sigma_i(a_i)$ is the probability i plays a_i. The rewards of the mixed extension of the game are given by the expected value of r_i under the joint mixed strategy $\sigma \in \times_{i\in N}\Delta(A_i)$ over outcomes:

$$r_i(\sigma) = \sum_{\mathbf{a}\in A}\left(\prod_{j\in N}\sigma_j(a_j)\right) r_i(\mathbf{a}). \tag{3.1}$$

We use the notation $\mathbf{a} = (a_i, \mathbf{a}_{-i})$, where $\mathbf{a}_{-i}$ is the joint action chosen by all agents other than i and, similarly, $\sigma = (\sigma_i, \sigma_{-i})$ where σ_{-i} is the joint independent lottery. In this paper, we are interested two solutions, namely *Nash* and *correlated equilibrium*.

DEFINITION 3.1. *A joint strategy, $\sigma^* \in \times_{i\in N}\Delta(A_i)$, is a* Nash equilibrium *if it satisfies:*

$$r_i(\sigma^*) - r_i(\sigma_i, \sigma^*_{-i}) \geq 0 \quad \forall\, \sigma_i \in \Delta(A_i),\ \forall\, i \in N.$$

If there is no player, i, that is indifferent between σ^*_i and another strategy (i.e. the inequality above is strict), then σ^* is a strict Nash equilibrium*. All strict Nash equilibria are pure.*

DEFINITION 3.2. *A distribution $\psi \in \Delta(A)$ is a* correlated equilibrium *if it satisfies:*

$$\sum_{\mathbf{a}\in A\,:\,a_i\neq k} \psi(\mathbf{a})\left(r_i(\mathbf{a}) - r_i(k, \mathbf{a}_{-i})\right) \geq 0 \quad \forall\, k \in A_i,\ \forall\, i \in N.$$

In a Nash equilibrium, every agent plays a best response to its opponents' strategies; while in a correlated equilibrium, every agent plays a best response to its opponents' strategies conditional on the correlating signal ψ, which, in a repeated game, may be the history of play. Note that every Nash equilibrium is a correlated equilibrium with ψ a product distribution (i.e. no correlation). Approximate δ–Nash and δ–correlated equilibria are defined by replacing the right hand side of the two expressions above with $\delta > 0$. A final technical definition is *generic games*: Γ is generic if a small change to any single reward does not change the number or location of the Nash equilibria of Γ. A sufficient condition for Γ to be generic is that a player is never indifferent between its pure actions. An important implication is that all pure Nash equilibria in generic games are strict.

3.2. Acyclicity properties for games. This section outlines a hierarchy of acyclicity properties for games. The first property is characterised by constructing a *best–reply graph* for a game Γ. This is a directed graph with vertices given by A, with an edge from $\mathbf{a}$ to $\mathbf{a}'$ if and only if there is exactly one player i that changes its action (i.e. $a_i \neq a'_i$ and $\mathbf{a}_{-i} = \mathbf{a}'_{-i}$) and $a'_i = \operatorname{argmax}_{a_i\in A_i}(r_i(a_i, \mathbf{a}_{-i}))$. Note that a pure Nash equilibrium of Γ is found at a *sink* of the best–reply graph, that is, a vertex with no outgoing edges. Next, a sequence of steps $(\mathbf{a}^0, \mathbf{a}^1, \dots, \mathbf{a}^t \dots)$ is called a *best–reply path* in Γ if each successive pair $\mathbf{a}^t, \mathbf{a}^{t+1}$ is joined

by an edge from $\mathbf{a}^t$ to $\mathbf{a}^{t+1}$ in the best–reply graph. A game Γ is *weakly acyclic* under best replies (or a *WAG*) if from every $\mathbf{a} \in A$, there exists a best–reply path that terminates in a Nash equilibrium in a finite number of steps [Young, 1993]. In a WAG, for each $\mathbf{a} \in A$, let $L_{\mathbf{a}}$ be the length of the shortest best–reply path from $\mathbf{a}$ to a pure Nash equilibrium, and let $L_\Gamma = \max_{\mathbf{a}\in A} L_{\mathbf{a}}$; we will need this constant in relation to the adaptive play algorithms.

A second, broader class is characterised by a similarly constructed *better–reply graph* for Γ. Again, this is a directed graph with vertices A, but with an edge from $\mathbf{a}$ to $\mathbf{a}'$ if and only if $a_i \neq a'_i$ and $\mathbf{a}_{-i} = \mathbf{a}'_{-i}$ and the change causes i's reward to improve: $r_i(a'_i, \mathbf{a}_{-i}) > r_i(a_i, \mathbf{a}_{-i})$. As for the best–reply graph, pure Nash equilibria are found at the sinks of the better–reply graph of Γ. A *better–reply path* is a sequence $(\mathbf{a}^0, \mathbf{a}^1, \ldots, \mathbf{a}^t \ldots)$ such that each successive pair is joined by a directed edge in the better–reply graph. A game Γ such that from every $\mathbf{a} \in A$ there exists a better–reply path that terminates in a sink in a finite number of steps is called a *weakly acyclic under better replies game* (*WABRG*) [Young, 2004].

A third form of acyclicity is characterised using a potential function. A *potential*, $\phi(\mathbf{a})$, is a function specifying the participants' joint preference over A [Monderer and Shapley, 1996], such that the difference in the potential induced by a unilateral change of action equals the change in the deviator's reward:

$$\phi(a_i, \mathbf{a}_{-i}) - \phi(a'_i, \mathbf{a}_{-i}) = r_i(a_i, \mathbf{a}_{-i}) - r_i(a'_i, \mathbf{a}_{-i}) \quad \forall\, a_i,\ a'_i \in A_i,\ \forall \mathbf{a}_{-i} \in A_{-i}.$$

A game that admits such a function is called a *potential game (PG)*.[1] Importantly, local optima of the $\phi(\mathbf{a})$ are Nash equilibria of Γ (analogous to the sinks of the best– or better–reply graph); that is, the potential is locally maximised by myopic self–interested players.

In order to highlight the connections to distributed optimisation, assume now that $\phi(\mathbf{a})$ represents the systems' global objectives. If the players' rewards are perfectly aligned with $\phi(\mathbf{a})$ (i.e. an increase in a player's reward improves the system reward by the same amount), then $\phi(\mathbf{a})$ is a potential for the game. Thus, the game's pure Nash equilibria are local optima of the global target function.

In summary, the key relationships between the classes above are given by: PG $\subset$ WAG $\subset$ WABRG [Monderer and Shapley, 1996; Young, 2004]. We will refer to these classes of games in Section 5 when considering our Q–learning algorithm variants.

3.3. Games with unknown noisy rewards. We now introduce the model of rewards received in a repeated learning situation that will be studied in the rest of this article. Much work on learning in games either assumes that the reward functions r_i are known in advance [e.g. Hart and Mas-Colell, 2000], or that the observed rewards are deterministic functions of the joint action selected [e.g Rosenthal, 1973; Cominetti et al., 2010]. However, as argued in Section 1, a more realistic scenario is that the observed rewards are noisy, and comprise of an expected value equal to the unknown underlying reward function $r_i(\mathbf{a})$ and a zero–mean random perturbation. We call this scenario *unknown noisy rewards*. This situation therefore requires the individuals to estimate their underlying reward functions, while also adapting their strategies in response to the actions of other agents.

DEFINITION 3.3. *A* game with unknown noisy rewards *is a game in which, when the joint action* $\mathbf{a} \in A$ *is played, agent* i *receives the reward*

$$R_i = r_i(\mathbf{a}) + e_i \tag{3.2}$$

[1] Potential games include *team games*, in which agents have the same reward function, which are often studied in distributed optimisation and artificial intelligence [e.g. Claus and Boutilier, 1998; Chapman et al., 2011]. Note that the team models here are built from the ground up, rather than imposed on existing agents with private motivations. This is a key point of difference from classic economic results on team decision–making, such as the works of Marschak and Radner [1972] and Groves [1973].

where $r_i(\mathbf{a})$ is the true expected reward to agent i from joint action $\mathbf{a} \in A$, and e_i is a random variable with expected value 0 and bounded variance.

To avoid unnecessary over–complication in this article, we assume that each realisation of each e_i is independent of all other random variables.[2] Note that a game with unknown noisy rewards is a generalisation of the bandit problem discussed by Sutton and Barto [1998], and we shall use similar reinforcement learning strategies to estimate the values of $r_i(\cdot)$.

3.4. Problem definition. We are now in a position to precisely describe the problem which we address. We imagine a game with unknown noisy rewards which is repeated over time. On each play of the game, the individuals select an action, receive rewards as per (3.2), and also observe the actions selected by the other players. Based on this information, the individuals update their estimates of the reward functions and adapt their actions.

For this scenario, we are interested in the evolution of strategies and, in particular, whether actions converge to equilibrium. Moreover, in the specific case of a potential game corresponding to a distributed optimisation problem, we want to prove convergence to Nash equilibrium, thereby providing a distributed method of computing (locally) optimal joint strategies with only noisy evaluations of the target function.

4. Convergence of reward function estimates using Q–learning. In this section we show that, in a game with unknown noisy rewards, agents can form estimates of the true reward functions that are sufficiently accurate to ensure that a Nash equilibrium can be found. In noisy environments, reinforcement learning is often used to estimate the mean value of a perturbed reward function [Sutton and Barto, 1998], so this is the method we adopt here. In particular, if the agents update their estimates of the expected rewards for joint actions using Q–learning, and select actions using an appropriate ε–greedy action selection policy, then with probability 1 the reward function estimates will converge to their true mean values. Now, Q–learning can be applied independently by each player of a game, who learns the expected reward for each action $a_i \in A_i$, ignoring the actions selected by the other agents (see Section 2). However this can result in very slow adaptation of actions towards Nash equilibrium. Instead, in this paper, we allow the learning of reward functions of joint actions, and simultaneous explicit reasoning about the action selection of the other agents. This is the JAL approach suggested (without analysis) in the context of fictitious play by Claus and Boutilier [1998]. The learning scheme we derive here will be used by all of the algorithms considered in Section 5.

In particular, we consider a multi–agent version of Q–learning for single–state problems, in which the agents select a joint action and each receives an individual reward. This algorithm operates by each individual recursively updating an estimate of its value of a joint action $\mathbf{a}$. Specifically, after playing action a_i^t, observing actions $\mathbf{a}^t_{-i}$, and receiving reward R_i^t, each individual i updates estimates Q_i^t using the equation:

$$Q_i^{t+1}(\mathbf{a}) = Q_i^t(\mathbf{a}) + \lambda(t) I\{\mathbf{a}^t = \mathbf{a}\}\left(R_i^t - Q_i^t(\mathbf{a})\right) \quad \forall \mathbf{a} \in A. \tag{4.1}$$

where the indicator $I\{\mathbf{a}^t = \mathbf{a}\}$ takes value 1 if $\mathbf{a}^t = \mathbf{a}$ and 0 otherwise, and $\lambda(t) \in (0,1)$ is a learning parameter. In general, $Q_i^t(\mathbf{a}) \to \mathbb{E}[R_i^t \,|\, \mathbf{a}^t = \mathbf{a}]$ with probability 1 if the conditions:

$$\sum_{t=1}^{\infty} \lambda(t) I\{\mathbf{a}^t = \mathbf{a}\} = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} (\lambda(t))^2 < \infty \tag{4.2}$$

[2] We believe that this assumption can be significantly relaxed without comprising our results, but requires significant effort to explain how estimation is adapted to handle correlated errors, which is beyond the scope of this paper.

hold for each $\mathbf{a} \in A$ [Jaakkola et al., 1994]. This can be achieved, under the condition that all $Q_i(\mathbf{a})$ are updated infinitely often, if:

$$\lambda(t) = \left(C_\lambda + \#^t(\mathbf{a})\right)^{-\rho_\lambda} \tag{4.3}$$

where $C_\lambda > 0$ is an arbitrary constant, $\rho_\lambda \in (1/2, 1]$ is a learning rate parameter, and $\#^t(\mathbf{a})$ is the number of times the joint action $\mathbf{a}$ has been selected up to time t. We use the form of $\{\lambda(t)\}_{t\geq 1}$ given by (4.3) in the remainder of the paper.

The condition that all actions $\mathbf{a}$ are played infinitely often can be met with probability 1 by using a randomised *learning policy*, in which the probability of playing each action is bounded below by a sequence that tends to zero sufficiently slowly as t becomes large. Furthermore, this learning policy can be chosen so that it is greedy in the limit, in that the probability with which it selects maximal reward actions tends to 1 as $t \to \infty$. Such policies are called *greedy in the limit with infinite exploration (GLIE)* [Singh et al., 2000].

One common GLIE policy is known as ε–greedy, and the results derived in this paper depend on the use of this particular rule. Under this policy, an agent selects a greedy action at time t with probability $(1-\varepsilon(t))$ (although note that we have not yet defined what a greedy action should be in this context), and chooses an action at random with probability $\varepsilon(t)$. In the single agent case, if, for example, $\varepsilon(t) = c/t$ with $0 < c < 1$, then for any a:

$$\sum_{t=1}^{\infty} \mathbb{P}(a^t = a) \geq \sum_{t=1}^{\infty} \frac{\varepsilon(t)}{|A|} = C \sum_{t=1}^{\infty} \frac{1}{t} = \infty,$$

and so with probability 1 each action is selected infinitely often [Singh et al., 2000].

In contrast to single agent settings, in multi–player games, the choice of joint action is made by the independent choices of more than one agent. As such, for each Q–value to be updated infinitely often, the schedule $\{\varepsilon(t)\}_{t\to\infty}$ that the sampling sequence follows must reflect the fact that the agents cannot explicitly coordinate to sample specific joint actions.

LEMMA 4.1. *In a game with unknown noisy rewards, if agents select their actions using a learning policy in which, for all $i \in N$, $a_i \in A_i$ and $t \geq 1$,*

$$\mathbb{P}(a_i^t = a_i) \geq \varepsilon_i(t), \quad \textit{with} \quad \varepsilon_i(t) = c_\varepsilon t^{-1/|N|},$$

where $c_\varepsilon > 0$ is a positive constant, then $\forall i \in N$, $\forall \mathbf{a} \in A$,

$$\lim_{t\to\infty} |Q_i^t(\mathbf{a}) - r_i(\mathbf{a})| = 0 \quad \textit{with probability 1.} \tag{4.4}$$

Proof. If the probability that agent i selects an action is bounded below by $\varepsilon_i(t) = c_\varepsilon t^{-1/|N|}$, then the probability that any joint action $\mathbf{a}$ is played is bounded below by:

$$\left(c_\varepsilon t^{-1/|N|}\right)^{|N|} = (c_\varepsilon)^{|N|} t^{-1}.$$

Since $\sum_{t=0}^{\infty} (c_\varepsilon)^{|N|} t^{-1} = \infty$, by a generalised Borel–Cantelli lemma [Jaakkola et al., 1994], with probability 1 each joint action $\mathbf{a} \in A$ is selected infinitely often, and the result follows. □

This may result in a practical learning procedure if $|N|$ is sufficiently small. However, in large games, visiting each joint action infinitely often is an impractical constraint. To achieve sufficiently high exploration rates through independent sampling, as in the ε–greedy approach, would require the agents' ε sequences to decrease so slowly that in any practical sense the agents will never move into an exploitation phase. To address this limitation, in Appendix A we consider sparse games in which each agent interacts directly with only a small number of other agents, such that the number of reward values each individual estimates can be significantly reduced. Here, however, we now move on to the main results of the paper.

5. Action adjustment process with learnt reward functions. In this section, we consider the convergence to Nash equilibria of adaptive play and better–reply processes with inertia, and of regret matching to the set of correlated equilibrium, using the Q–learning approaches described above. Specifically, for the first two classes of algorithm, we show that if the agents (i) update their estimates of the expected rewards for joint actions using Q–learning, (ii) update their beliefs over their opponents' actions using an appropriate action adjustment algorithm, and (iii) select a new action using an appropriate ε–greedy learning policy, then their actions converge to a Nash equilibrium in potential games with unknown noisy rewards. For regret matching, we show that if the same conditions hold, then the agents' action frequencies converge to the set of correlated equilibria in all generic games. The first part of this section comprises a brief recap of Markov chains, before the three subsequent sections cover the main results of the paper for each of the algorithms listed above.

5.1. Markov chain basics. Our approach is to consider the ergodicity properties of Markov chains induced by time–inhomogeneous versions of adaptive play and better–reply processes with inertia. We then show that Q–learning variants of the processes eventually behave as if the reward functions have been correctly learnt, so that the same convergence results follow. We begin by recalling some definitions from Markov chain theory.

Consider a nonstationary Markov chain $\{X^0, X^1, X^2, \ldots\}$ on a finite state space $\mathcal{X}$, with a sequence of transition matrices $\{\mathbf{P}(t)\}_{t\geq 1}$ such that $\mathbf{P}(X^{t+1} = y \,|\, X^t = x) = (P(t))_{xy}$. Ergodicity of this chain corresponds to properties of the products $\mathbf{P}(s,t) = \prod_{\tau=s}^{t} P(\tau)$. Following Isaacson and Madsen [1976] we distinguish between *weakly ergodic* chains, where the effect of the initial state vanishes (i.e. the rows of $\mathbf{P}(s,t)$ are near to identical for sufficiently large t), and *strongly ergodic* chains, which converge to a steady state distribution (i.e. all rows of $\mathbf{P}(s,t)$ converge to a fixed distribution).

DEFINITION 5.1. *A nonstationary Markov chain is* weakly ergodic *if, for all $s \geq 1$, a sequence of vectors $\mu(s,t)$ exist such that*

$$\lim_{t\to\infty} \Big(P(s,t)_{xy} - \mu(s,t)_y \Big) = 0 \quad \forall\, x, y \in \mathcal{X}. \tag{5.1}$$

A nonstationary Markov chain is strongly ergodic *if a steady state distribution μ exists such that, for all $s \geq 1$,*

$$\lim_{t\to\infty} \Big(P(s,t)_{xy} - \mu_y \Big) = 0 \quad \forall\, x, y \in \mathcal{X}. \tag{5.2}$$

Anily and Federgruen [1987] demonstrate how to show that a Markov chain with a converging sequence of transition matrices satisfies these ergodicity properties. This, in turn, can be used to show that as t gets large, the distribution of X^t is approximately equal to the distribution that is the limit of the stationary distributions of the transition matrices $P(t)$. We will show that our learning processes generate a strongly–ergodic Markov chain with a converging transition matrix sequence, for which the limiting stationary distribution places all of its mass on the pure Nash equilibria of the game.

5.2. Q–learning adaptive play. For both Q–learning adaptive play and better–reply processes with inertia, each agent possesses a finite memory of length m, recalling the history of the previous m actions taken by its opponents. Let h be a joint history of length m, where $h = (\mathbf{a}^{t-m}, \ldots, \mathbf{a}^{t-1})$, and let H of size $|A|^m$ be the collection of all the possible joint histories. After observing an action profile $\mathbf{a}^t$, the joint history configuration moves from h to h' by removing the left–most element of h and adjoining $\mathbf{a}^t$ as the right–most element of h'. A *successor* to h is any history $h' \in H$ that can be obtained in this way. For example, imagine

a two–player game with $A_1 = \{a,b\}$ and $A_2 = \{A,B\}$. Let $m = 2$ and set $h = (aA,aA)$; the successors to h are $h' \in \{(aA,aA),(aA,bA),(aA,aB),(aA,bB)\}$. Note that in the first m plays of the game there will not be a full memory; since the whole point of the proofs is that the starting point is forgotten we do not address this issue here (one could assume that an arbitrary initial history is selected, for example).

DEFINITION 5.2 (Adaptive play [Young, 1993]). *At each time–step, each agent samples $k \leq m$ of the elements of its memory of length m, and plays a best response to the actions in the sample.*

If $k \leq m/(L_\Gamma + 2)$, then adaptive play converges to a Nash equilibrium in games that are weakly acyclic under best replies (WAGs) [Theorem 1, Young, 1993]. We consider the (stationary) Markov chain on the state space H generated by adaptive play. Let $p_i(a_i|h)$ be the best–reply distribution for agent i, with $p_i(a_i|h) > 0$ only if there exists a sample of length k from h to which a_i is a best reply for i. Then, if $\mathbf{a}$ is the right–most element of h', a successor to h, the probability of moving from h to h' is:

$$P^0_{hh'} = \prod_{i \in N} p_i(a_i|h), \tag{5.3}$$

while if h' is not a successor to h, then $P^0_{hh'} = 0$. The convergence proved by Young is that this Markov process converges to an absorbing state, and each absorbing state is a history consisting entirely of one strict Nash equilibrium.

However for Q–learning variants we have seen that it is important to play all (joint) actions infinitely often. Therefore we consider a perturbation, which we call *uniform sampling*, applied to the adaptive play process. We begin by considering the simpler case of games with known rewards, before extending the analysis to unknown noisy rewards.

DEFINITION 5.3 (Stochastic adaptive play). *At each time–step, each agent acts independently and uses the (unperturbed) adaptive play rule with probability $1-\varepsilon(t)$, or uniformly samples from A_i with probability $\varepsilon(t)$.*

For $\varepsilon(t) = \varepsilon$ fixed, this process is considered by Young [1993]. The transition matrix of stochastic adaptive play at time t with perturbations $\varepsilon(t)$ is $\mathbf{P}^{\varepsilon(t)}$, where $\mathbf{P}^\varepsilon$ has hh' entry

$$P^\varepsilon_{hh'} = (1-\varepsilon)^{|N|} P^0_{hh'} + \sum_{K \subseteq N,\, K \neq \emptyset} \varepsilon^{|K|} (1-\varepsilon)^{|N|-|K|} U^K_{hh'} \tag{5.4}$$

where

$$U^K_{hh'} = \begin{cases} \displaystyle\prod_{i \in K} \frac{1}{|A_i|} I\{a'_i = a_i\} \prod_{i \notin K} p_i(a'_i|h) & \text{if } h \text{ is a successor to } h' \text{ and } a_i \text{ (resp. } a'_i\text{) is} \\ & \text{the } i^{th} \text{ entry of the right–most element of } h \\ & \text{(resp. } h'\text{); and} \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, if $\varepsilon(t) = \varepsilon$ for all t then we have a stationary, irreducible and aperiodic Markov chain on a finite state space, which is then both weakly and strongly ergodic; denote the unique stationary distribution of $\mathbf{P}^\varepsilon$ by μ^ε; this result tells us that $\mathbb{P}(h^t = h) \to \mu^\varepsilon_h$ as $t \to \infty$.

Young's analysis of this process considers stochastic stability of states h; a state h is called *stochastically stable* with respect to a family of processes $\mathbf{P}^\varepsilon$ if $\lim_{\varepsilon \to 0} \mu^\varepsilon_h > 0$. Theorem 2 of Young [1993] states that the stochastically stable states of stochastic adaptive play with fixed ε are the histories consisting entirely of a single strict Nash equilibrium in generic WAGs. One trivial implication of this result is that for any δ, for sufficiently small fixed $\varepsilon > 0$, $\lim_{t \to \infty} \mathbb{P}(\mathbf{a}^t \text{ is a strict Nash equilibrium}) > 1 - \delta$.

It is also clear that with a fixed ε, all joint actions will be played infinitely often, so the Q–learned estimates of joint action rewards will converge (in particular the conditions of Lemma 4.1 are trivially satisfied). Hence we can consider the following process:

DEFINITION 5.4 (*Q*–learning adaptive play). *At each time–step, actions are selected according to Definition 5.3, with best responses calculated with respect to joint action reward estimates that are updated according to* (4.1) *with* $\{\lambda(t)\}_{t\geq 1}$ *following* (4.3) *with* $C_\lambda > 0$ *and* $\rho_\lambda \in (1/2, 1]$.

Since the reward estimates converge almost surely, the following is straightforward:

LEMMA 5.5. *Let* Γ *be a WAG with unknown noisy rewards and* $k \leq m/(L_\Gamma + 2)$. *For any* $\delta > 0$ *there exists an* $\varepsilon > 0$ *such that, under Q–learning adaptive play with fixed* ε, *for all sufficiently large t,* $\mathbb{P}(\mathbf{a}^t$ *is a strict Nash equilibrium*$) > 1 - \delta$.

Proof. Note that the only difference from standard stochastic adaptive play is that the best responses are calculated using the estimated *Q*–values instead of the true rewards. However since the reward functions are bounded in absolute value, the game is generic, and the action spaces and memory are finite, there exists an $\eta > 0$ such that if for all $i \in N$, and for all $\mathbf{a} \in A$,

$$|Q_i^t(\mathbf{a}) - r_i(\mathbf{a})| < \eta, \tag{5.5}$$

then the best responses are the same whether the individuals use r_i or Q_i^t.

We know that (4.4) holds, so that with probability 1 there exists a $T < \infty$ such that for all $t \geq T$ (5.5) holds. It follows that, after T time–steps the actions of agents evolve exactly as if they were using r_i instead of Q_i^t. Hence the result follows immediately from the result on standard stochastic adaptive play. □

However, we do not need to consider fixed $\varepsilon(t)$, and with decreasing $\varepsilon(t)$ we do not need to pre–guess a "suitably small" value. We will show that a schedule similar to that required for Lemma 4.1 also implies suitable ergodicity properties of stochastic adaptive play. Combining this result with the technique of Lemma 5.5 will give our result.

LEMMA 5.6. *Let* Γ *be a generic WAG. Consider stochastic adaptive play with* $\varepsilon(t) = ct^{-1/mN}$. *If* $k \leq m/(L_\Gamma + 2)$ *then* $\lim_{t\to\infty} \mathbb{P}(\mathbf{a}^t$ *is a Nash equilibrium*$) = 1$.

Note that the ε schedule above is slightly different to that for Lemma 4.1; however, if $\varepsilon(t)$ satisfies the conditions here then it necessarily satisfies those of Lemma 4.1.

Proof. This result is proved in three steps, first showing that the process is weakly ergodic, second, that it is strongly ergodic, and third, that the distribution of histories converges to $\mu^* = \lim_{\varepsilon\to 0} \mu^\varepsilon$, with the consequence that in generic WAGs, $\lim_{\varepsilon\to 0} \mu_h^\varepsilon > 0$ only if h consists only of a single pure Nash equilibrium.

Weak ergodicity is proved by examining the *ergodic coefficients* of the sequence of matrices $\{\mathbf{P}^{\varepsilon(t)}(t)\}_{t\geq 1}$. The ergodic coefficient of any stochastic matrix $\mathbf{P}$, denoted $\text{erg}(\mathbf{P})$, is given by:

$$\text{erg}(\mathbf{P}) = \min_{xy} \sum_w \min(P_{xw}, P_{yw}).$$

By Theorem V.3.2 from Isaacson and Madsen [1976], the nonstationary process $\{\mathbf{P}^{\varepsilon(t)}(t)\}_{t\geq 1}$ is weakly ergodic if the product $\mathbf{P}^{\varepsilon(1)}(1) \cdot \mathbf{P}^{\varepsilon(2)}(2) \cdots$ can be divided into blocks of matrices:

$$\begin{aligned}\mathbf{P}^{\varepsilon(1)}(1) \cdot \mathbf{P}^{\varepsilon(2)}(2) \cdots &= [\mathbf{P}^{\varepsilon(1)}(1) \cdots \mathbf{P}^{\varepsilon(t_1)}(t_1)] \cdots [\mathbf{P}^{\varepsilon(t_k+1)}(t_k+1) \cdots \mathbf{P}^{\varepsilon(t_{k+1})}(t_{k+1})] \cdots \\ &= \mathbf{P}^{\varepsilon(1,t_1)}(1,t_1) \cdots \mathbf{P}^{\varepsilon(t_k+1,t_{k+1})}(t_k+1,t_{k+1}) \cdots\end{aligned}$$

such that:

$$\sum_{k=0}^{\infty} \text{erg}\left(\mathbf{P}^{\varepsilon(t_k,t_{k+1}-1)}(t_k, t_{k+1}-1)\right) = \infty \quad \text{where } t_0 = 1. \tag{5.6}$$

Stochastic adaptive play is shown to satisfy (5.6) by considering blocks of length m.[3] Let $\prod_{i\in N}\varepsilon_i(t) = Ct^{-1/m}$ be the minimum probability that any joint action is played at time–step t. Consider blocks of length m, such that $t_k = \{1, m+1, 2m+1\ldots\}$, and observe that any state h' can be reached from an initial state h by an appropriate sequence of m random samples, which occurs with least probability $\prod_{t=km+1}^{(k+1)m} Ct^{-1/m}$. Evaluating (5.6) then gives:

$$\sum_{k=0}^{\infty} \operatorname{erg}(\mathbf{P}^{\varepsilon(t_k,t_k+m)}(t_k,t_k+m)) \geq \sum_{k=0}^{\infty}\prod_{t=km+1}^{(k+1)m} Ct^{-1/m} > C^m \sum_{k=0}^{\infty}\left(((k+1)m)^{-1/m}\right)^m = \infty.$$

Thus, stochastic adaptive play with $\varepsilon_i(t) = Ct^{-1/Nm}$ is weakly ergodic.

We now prove that stochastic adaptive play is strongly ergodic, by showing that it meets the conditions of Theorem 2 of Anily and Federgruen [1987], which gives the sufficient conditions for a weakly–ergodic nonstationary Markov chain to be strongly ergodic. This involves: (i) constructing an *extension* $\bar{\varepsilon}(x)$ of the sequence $\{\varepsilon(t)\}_{t\geq 1}$; (ii) constructing a *regular extension* $\bar{\mathbf{P}}^{\bar{\varepsilon}(x)}(x)$ of the nonstationary process $\mathbf{P}^{\varepsilon(t)}(t)$; and (iii) showing that all entries of the regular extension $\bar{\mathbf{P}}^{\bar{\varepsilon}(x)}(x)$ are members of a closed class of asymptotically monotone functions.

DEFINITION 5.7. *Let $\{a(t)\}_{t\geq 1}$ be a sequence with $a(t)\in\mathbb{R}^m$ for some $m\geq 1$. The function $\bar{a}(x):(0,1]\to\mathbb{R}^m$ is an extension of the sequence if $\bar{a}(x_t)=a(t)$ for some sequence $\{x_t\}_{t\geq 1}$, with $\lim_{t\to\infty} x_t = 0$.*

To construct an extension of the sequence of sampling probabilities given in the statement of Lemma 5.6, let $\bar{\varepsilon}(\cdot)$ be a vector function whose ith component is given by: $\bar{\varepsilon}_i(x) = cx^{1/Nl}\ \forall i\in N$. To verify this, set $x_t = t^{-1}$ so that $x_t\in(0,1]$ for all $t\geq 1$, $\lim_{t\to\infty}x_t = 0$, and the ith component of $\bar{\varepsilon}(\cdot)$ evaluated at x_t gives: $\bar{\varepsilon}_i(x_t) = cx^{1/Nl} = c(t^{-1})^{1/Nl} = \varepsilon_i(t)$.

DEFINITION 5.8. *Let $\bar{\mathbf{P}}(\cdot)$ be an extension of a nonstationary Markov chain $\{\mathbf{P}^{\varepsilon(t)}(t)\}_{t\geq 1}$: $\bar{\mathbf{P}}(\cdot)$ is said to be a* regular extension *of $\{\mathbf{P}^{\varepsilon(t)}(t)\}_{t\geq 1}$ if there exists a $x^*\in\mathbb{R}_{++}$ such that the collection of all subchains of $\bar{\mathbf{P}}(x)$ is identical for all $x<x^*$.*

Let $\bar{\mathbf{P}}^{\bar{\varepsilon}_i(x)}(x)$ be an extension to the Markov chain generated by stochastic adaptive play. This is regular if the set $\{(h,h') : \bar{P}_{hh'}^{\bar{\varepsilon}_i(x)}(x) > 0\}$ is identical for all $x<x^*$.

In the setting of known rewards, the values $P^0_{hh'}$ and $Q^K_{hh'}$ are independent of x (or t), so the first term on the right–hand side of (5.4) is positive for all x if and only if $P^0_{hh'}>0$,

[3] In contrast, if $l<m$ transitions are considered, it is always possible to pick two starting states h and g such that the set of reachable states from both h and g after l transitions is empty. Thus, for every column of the associated l–step transition matrix, $\mathbf{P}^{\varepsilon(t,t+l)}(t,t+l)$, if $l<m$, one or the other rows' entry is 0, so its ergodic coefficient is also 0. For example, consider a two–player game with $A_1=\{a,b\}$ and $A_2=\{A,B\}$ and rewards $r_i(a,A)=r_i(b,B)=1$ and $r_i(a,B)=r_i(b,A)=0$ for both players $i=\{1,2\}$. Let both adjust their actions using stochastic adaptive play with $m=2$. Now imagine two copies of this process, one starting at $h=(aA,aA)$ and the other at $g=(bBbB)$, where these starting points are chosen to contain no entries in common. The successors to h are $h'=\{(aA,aA),(aA,bA),(aA,aB),(aA,bB)\}$, while for g they are $g'=\{(bB,aA),(bB,bA),(bB,aB),(bB,bB)\}$, so $h'\cap g'=\emptyset$, The one–step state transition probabilities are given in the following rows of the transition matrix:

	aA, aA	bA, aA	aA, bA	bA, bA	aB, aA	bB, aA	aB, bA	bB, bA	aA, aB	bA, aB	aA, bB	bA, bB	aB, bB	bB, aB	aB, bB	bB, bB
aA, aA	$(1-\varepsilon)^2$	0	$(1-\varepsilon)\varepsilon$	0	0	0	0	0	$\varepsilon(1-\varepsilon)$	0	ε^2	0	0	0	0	0
⋮	⋮															⋮
bB, bB	0	0	0	0	0	ε^2	0	$(1-\varepsilon)\varepsilon$	0	0	0	0	0	$\varepsilon(1-\varepsilon)$	0	$(1-\varepsilon)^2$

All column–wise minimums of these two rows are 0, so the ergodic coefficient of the one–step transition matrix is 0. It may be the case that the state space of the chain for a specific game may be reduced, but we can say no more in general terms without knowing the game's rewards; which are what the players are trying to estimate. Only by considering blocks of length m or greater can this be avoided.

and similarly, the second term of (5.4) is positive for all x whenever $Q^K_{hh'} > 0$. Thus, setting $x^* = 1$, we see that the set of transitions through the memory configuration space with strictly positive probabilities is identical for all $x < x^*$.

Having constructed and verified a regular extension to $\mathbf{P}^{\varepsilon(t)}(t)$, in order to prove that stochastic adaptive play is strongly ergodic, we are now left to show that every entry function in $\bar{\mathbf{P}}^{\bar{\varepsilon}_i(x)}(x)$ belongs to a closed class of asymptotically monotone functions $\mathcal{F}$.

DEFINITION 5.9. *A class $F \subset C^1$ of functions defined on $(1,0]$ is a* closed class of asymptotically monotone (CAM) functions *if: (i) $f \in F \Rightarrow f' \in F$ and $-f \in F$; (ii) $f,g \in F \Rightarrow (f+g) \in F$ and $(f \cdot g) \in F$; and (iii) all $f \in F$ change signs finitely often in on $(0,1]$.*

DEFINITION 5.10. *Let $\mathcal{F}$ be the class of real valued functions such that every $f \in \mathcal{F}$ is of the form $\sum_{k=1}^K (V_k(x))^{1/c_k}$, with c_k a given integer (including negatives) and $V_k(\cdot)$ a given polynomial function that is positive on $(0,1]$.*

Observe that the class $\mathcal{F}$ is CAM on $(0,1]$ (although not necessarily everywhere on $\mathbb{R}$). Regarding $\mathbf{P}^{\varepsilon(t)}(t)$, in the first term of (5.4), the product $\prod_{i\in N}(1 - c_\varepsilon t^{-1/Nlx})$ is a member of $\mathcal{F}$, as are the two products in the second term. Thus, all elements of the transition matrix of stochastic adaptive play are functions in the class $\mathcal{F}$. Therefore, stochastic adaptive play is strongly ergodic.

Finally, we characterise the supports of μ^* by combining Theorem 2 of Young [1993] and these ergodicity properties. Specifically, Theorem 2 of Anily and Federgruen [1987] states that if a nonstationary Markov process is strongly ergodic, then for large enough t, each transition matrix $\mathbf{P}^{\varepsilon(t)}(t)$ is associated with a steady state distribution over actions $\mu^{\varepsilon(t)}(t)$ with $\lim_{t\to\infty}\mu^{\varepsilon(t)}(t) = \mu^*$ and $\lim_{t\to\infty}\mathbf{P}^{(s,t)}_{hh'} = \mu^*_{h'}$ for all $h,h' \in H$, $s \geq 1$. Theorem 2 of Young [1993] tells us that μ^* puts mass only on the stochastically stable states, which in generic WAGs are a subset of the strict Nash equilibria. Therefore $\lim_{t\to\infty}\mathbb{P}(h^t \in$ stochastically stable states of $\Gamma) = 1$, and since the unique best reply to a strict Nash equilibrium is to continue playing the same, the result follows. □

Note that the players will move between strict Nash equilibria — that is the meaning of ergodicity in this context — but will spend increasingly long durations of play at one strict Nash equilibrium as $\varepsilon \to 0$.

THEOREM 5.11. *Let Γ be a game with unknown noisy rewards with mean rewards that are a generic WAG (a noisy generic WAG). Consider Q–learning adaptive play with $\varepsilon(t) = ct^{-1/mN}$. If $k \leq m/(L_\Gamma + 2)$ then $\lim_{t\to\infty}\mathbb{P}(\mathbf{a}^t$ is a Nash equilibrium$) = 1$.*

Proof. The proof follows exactly the same logic as Lemma 5.5. After an almost surely finite time T the reward estimates will be sufficiently close to the true mean rewards that Q–learning adaptive play will make action selections with exactly the same probabilities as a stochastic adaptive play. Hence the result follows from Lemma 5.6. □

5.3. Q–learning better–reply processes with inertia. We now examine Q–learning better–reply processes with inertia.

DEFINITION 5.12 (Better–reply processes with inertia [Young, 2004]). *At each time step, with probability ξ_i an agent plays the same action as in the previous time step, $a_i^t = a_i^{t-1}$, while with probability $1-\xi_i$ the agent selects an action according to a distribution that puts positive probability only on actions that are better replies to its full memory of length m than a_i^{t-1}.*

By Theorem. 6.2 of Young [2004], if Γ is generic and weakly acyclic under better replies (a generic WABRG) and each $0 < \xi_i < 1$ for all $i \in N$, then the unperturbed better–reply processes with inertia converges almost surely to a homogeneous state consisting of one strict Nash equilibrium of Γ. Any algorithm that selects from the set of better replies to its (full, undiscounted) memory falls into this class of algorithms. Accordingly, it is a large class of algorithms that includes those that choose actions based on either their expected reward

(i.e. an improvement in expected reward over the current action), such as the better–response dynamics of Friedman and Mezzetti [2001] and the evolutionary–inspired process of Kandori et al. [1993], or based on *regrets* computed from a finite memory, as in Young [2004]. Like adaptive play, the better–reply processes with inertia generates a Markov chain on a state space H. Now let $p_i(a_i|h)$ be the better–reply distribution used by agent i, with $p_i(a_i|h) > 0$ only if a_i is a better reply to h for i. Let $L \subseteq N$ be a set of players having inertia and choosing not to update their action at the current time–step. The Markov process transition function for a finite memory better–reply process with inertia, given a vector of inertial constants $\xi = \{\xi_1, \ldots, \xi_n\}$, is then given by:

$$P^0_{hh'} = \prod_{i \in N} (1 - \xi_i)\, p_i(a_i|h) + \sum_{L \subseteq N,\, L \neq \emptyset} \left(\prod_{i \in L} \xi_i \right) \left(\prod_{i \notin L} (1 - \xi_i) \right) I^L_{hh'} \tag{5.7}$$

where

$$I^L_{hh'} = \begin{cases} \prod_{i \in L} I\{a'_i = a_i\} \prod_{i \notin L} p_i(a'_i|h) & \text{if } h \text{ is a successor to } h' \text{ and } a_i \text{ (resp. } a'_i\text{) is the } i^{th} \text{ entry of the right–most element of } h \text{ (resp. } h'\text{); and} \\ 0 & \text{otherwise.} \end{cases}$$

The first term in (5.7) is the product of the transition probability of the finite memory better–reply process without inertia and the probability that no agent has inertia, while the second term captures the probabilities of transitions arising from all partitions of the agents into those that do have inertia (L) and those that playing according to the unperturbed dynamics ($N \setminus L$). Note that although this looks like the stochastic adaptive play transition function, this is in fact the *unperturbed* dynamics of the better–reply process. We will introduce experimentation as well, to ensure sufficient exploration of the joint action space.

We can analyse stochastic better–replies with inertia in an identical way to the earlier treatment of stochastic adaptive play. Specifically, substitute $P^0_{hh'}(t)$ from (5.7) into (5.4) and also treat $p_i(a_i|h)$ as the better–reply distribution for agent i defined earlier in this section. This gives the perturbed transition probability for stochastic better–replies with inertia, which looks identical to (5.4) but has different values $P^0_{hh'}(t)$ and $p_i(a_i|h)$. Adding uniform sampling leads to the following definition.

DEFINITION 5.13 (Stochastic better–reply process with inertia). *At each time–step, each agent acts independently and follows the (unperturbed) better–replies with inertia process with probability* $1 - \varepsilon(t)$*, or uniformly samples from* A_i *with probability* $\varepsilon(t)$.

For fixed $\varepsilon(t) = \varepsilon$ we again have a stationary, irreducible and aperiodic finite Markov chain, which is, therefore, strongly ergodic; denote the transition matrix $\mathbf{P}^\varepsilon$, and its corresponding unique stationary distribution μ^ε, with $\mathbb{P}(h^t = h) \to \mu^*_h$ as $t \to \infty$.

The behaviour of stochastic better–replies with inertia has not been stated elsewhere to date, but can be analysed using Theorem 4 of Young [1993].

LEMMA 5.14. *The stochastically stable states of the stochastic better–reply process with inertia with fixed* ε *in generic WABRGs are the histories consisting entirely of a single strict Nash equilibrium.*

Proof. In order to show that stochastic better–replies with inertia satisfies the requirements of Theorem 4 of Young [1993], note that it is ergodic, and in the limit as $\varepsilon \to 0$, its transition probabilities converge to those of the unperturbed process; that is $\lim_{\varepsilon \to 0} \mathbf{P}^\varepsilon = \mathbf{P}^0$. Next, recall that if Γ is a generic WABRG and $0 < \xi < 1$, then the unperturbed process converges almost surely to a homogeneous state consisting of one strict Nash equilibrium of Γ. Given that the stochastically stable states of stochastic better–replies with inertia are a subset of the absorbing states of the unperturbed process, the result follows. □

Furthermore, it also follows that for any δ, there exists a sufficiently small fixed $\varepsilon > 0$ such that $\lim_{t\to\infty} \mathbb{P}(\mathbf{a}_t \text{ is a strict Nash equilibrium}) > 1-\delta$.

With a fixed ε the conditions of Lemma 4.1 are satisfied, so we can consider a version of better–replies with inertia using Q–learnt reward estimates.

DEFINITION 5.15 (Q–learning better–replies with inertia). *At each time–step, actions are selected according to Definition 5.13, with better–replies calculated with respect to reward estimates that are updated according to 4.1 with* $\{\lambda(t)\}_{t\geq 1}$ *following* (4.3) *with* $C_\lambda > 0$ *and* $\rho_\lambda \in (1/2, 1]$.

LEMMA 5.16. *Let* Γ *be a WABRG with unknown noisy rewards and* $0 < \xi < 1$. *For any* $\delta < 0$ *there exists an* $\varepsilon > 0$ *such that, under Q–learning better–replies with inertia with fixed* ε, *for all sufficiently large t,* $\mathbb{P}(a_t \text{ is a strict Nash equilibrium}) > 1-\delta$.

The proof of this result follows that for Lemma 5.5. However, we are not concerned with choosing a δ, since we consider decreasing $\varepsilon(t)$. We will show that a suitable schedule implies the strong ergodicity of stochastic better–replies with inertia. Combining this result with Lemma 5.14 gives the following:

LEMMA 5.17. *Let* Γ *be a generic WABRG. Consider stochastic better–replies with inertia with* $\varepsilon(t) = ct^{-1/mN}$. *If* $0 < \xi < 1$, *then* $\lim_{t\to\infty} \mathbb{P}(\mathbf{a}^t \text{ is a Nash equilibrium}) = 1$.

Note that, as before, the ε schedule above necessarily satisfies the conditions of Lemma 4.1.

Proof. The stochastically stable states of stochastic better–replies with inertia are a subset of the strict Nash equilibria in WABRGs. Regarding strong ergodicity of stochastic better–replies with inertia, since (5.7) is independent of the values of $\{\varepsilon(t)\}_{t\geq 1}$, the perturbed transition matrix for stochastic finite memory better–reply with inertia has the same properties as that for stochastic adaptive play with respect to ε, and, *mutatis mutandis*, the argument for strong ergodicity is the same. The proof is completed by combining this result with Lemma 5.16. □

THEOREM 5.18. *Let* Γ *be a game with unknown noisy rewards with mean rewards that are a generic WABRG (a noisy generic WABRG). Consider Q–learning better–reply process with inertia and* $\varepsilon(t) = ct^{-1/mN}$. *If* $0 < \xi < 1$ *then* $\lim_{t\to\infty} \mathbb{P}(\mathbf{a}^t \text{ is a Nash equilibrium}) = 1$.

Proof. Following Lemma 5.5, after an almost surely finite time T the reward estimates will be sufficiently close to the true mean rewards that Q–learning better–replies with inertia will make action selections with exactly the same probabilities as a stochastic better–replies with inertia. Hence the result follows from Lemma 5.17. □

REMARK 1. *We note that a class of* best–reply algorithms with inertia *may be defined and analysed in a similar way to the better–reply processes above. Under these processes, the set of best responses substitute for better replies, and convergence is guaranteed in games that are generic and weakly acyclic under best replies.*

We conclude this section with the following corollaries of Theorems 5.11 and 5.18 for these two learning algorithms in potential games:

COROLLARY 5.19. *Q–learning adaptive play and Q–learning better–replies with inertia both converge to a pure Nash equilibrium in games with noisy unknown rewards with mean rewards that are generic and admit a potential function.*

5.4. Q–learning regret matching. We now consider a third class of algorithms, called *regret matching* [Hart and Mas-Colell, 2000], and introduce a Q–learning regret matching variant. Regret matching uses measures of *average regret* rather than expected utility to evaluate action choices, and has a state space given by the set of joint actions A. Unlike adaptive play and finite better reply processes with inertia, which converge to a strict Nash equilibrium in various acyclic games, the regret matching algorithm converges in long–run frequency of play to the set of correlated equilibria in all generic games.

Formally, let

$$z^t_{\mathbf{a}} = {}^1\!/_t \sum_{\tau=0}^{t-1} I\{\mathbf{a}^\tau = \mathbf{a}\}$$

be the empirical frequency of play of joint action $\mathbf{a}$ at time t, and let $\mathbf{z}^t$ be the vector of length $|A|$ containing all of the components $\{z^t_{\mathbf{a}}\}_{\mathbf{a}\in A}$. Denote Ψ the set of correlated equilibrium distributions over A (i.e. Ψ contains all ψ satisfying Definition 3.2). Regret matching generates a sequence of distributions $\{\mathbf{z}^t\}_{t\geq 1}$ whose distance from the set Ψ converges to zero [Hart and Mas-Colell, 2000].

In order to calculate its action choice probabilities, an agent, i, using regret matching, computes the difference in its reward for switching to action a'_i every time that it played action a_i in the past. Agent i updates each $z^t_{\mathbf{a}}$ in its memory using the recursion:

$$z^t_{\mathbf{a}} = \frac{1}{t}\left(I\{\mathbf{a}^{t-1} = \mathbf{a}\} + (t-1)\, z^{t-1}_{\mathbf{a}}\right). \tag{5.8}$$

The values in $\mathbf{z}^t$ are interpreted as agent i's *belief* over the joint actions. For every pair of actions $j,k \in A_i$, the average difference in rewards for switching to k every time that j was played is:

$$\begin{aligned} D_{j,k}(t) &= \frac{1}{t}\sum_{\tau=1}^{t} I\{a^t_i = j\}\left(r_i(k,\mathbf{a}^t_{-i}) - r_i(j,\mathbf{a}^t_{-i})\right) \\ &= \sum_{\mathbf{a}\in A} \mathbf{z}^t_{\mathbf{a}}\, I\{a_i = j\}\left(r_i(k,\mathbf{a}_{-i}) - r_i(j,\mathbf{a}_{-i})\right). \end{aligned}$$

Now let the average *regret* for not making the switch to k on every play of j be given by:

$$R_{j,k}(t) = \max\{D_{j,k}(t), 0\}. \tag{5.9}$$

Finally, let:

$$\xi_i \geq (|A_i|-1)\max_{\mathbf{a},\mathbf{a}'}\{|r_i(\mathbf{a})) - r_i(\mathbf{a}')|\} \quad \forall\, i \in N \tag{5.10}$$

be an inertial constant. Action choice probabilities are calculated as follows.

DEFINITION 5.20 (Regret matching). *At each time step, $\mathbf{z}^t$ is updated by* (5.8)*, and the agent computes the regret for its actions by* (5.9)*. Then, the agent chooses an action with probability:*

$$\mathbb{P}(a^t_i = a'_i) = \begin{cases} \frac{1}{\xi_i} R_{a^{t-1}_i, a'_i}(t), & \text{for all } a'_i \neq a^{t-1}_i; \\ 1 - \frac{1}{\xi_i}\sum\limits_{a'_i \neq a^{t-1}_i} R_{a_i,a'_i}(t) & a'_i = a^{t-1}_i. \end{cases}$$

Note that the choice of ξ ensures that a^{t-1}_i is repeated with positive probability, and that any other action is chosen iff it has positive regret. If all agents playing a generic game use the procedure above, then sequence $\{\mathbf{z}^t\}_{t\geq 1}$ "approaches" the set of correlated equilibria, in the sense of Blackwell [1956], such that as $t \to \infty$, $\mathbb{P}(\mathbf{z}^t \in \Psi) = 1$; in other words, the distribution of the empirical history of play converges to the set of correlated equilibria [Hart and Mas-Colell, 2000]. Furthermore, for any $\delta > 0$, the algorithm enters the set of δ–correlated equilibria in a finite amount of time with probability 1.

DEFINITION 5.21 (Q–learning regret matching). *Each time–step, each agent uniformly samples from A_i with probability $\varepsilon(t)$, or follows the unperturbed regret matching dynamics with probability $1-\varepsilon(t)$ using reward estimates that are updated according to* (4.1) *in which:*

- $\{\lambda(t)\}_{t\geq 1}$ *follows* (4.3) *with* $C_\lambda > 0$ *and* $\rho_\lambda \in (1/2, 1]$, *and*
- $\varepsilon(t) = ct^{-1/N}$ *for all* $i \in N$.

THEOREM 5.22. *Let* Γ *be a generic game with unknown noisy rewards. If Q–learning regret matching with* $\varepsilon(t) = ct^{-1/mN}$ *is used by all players, then* $\mathbb{P}(\lim_{t\to\infty} R_{j,k}(t) = 0) = 1$*; therefore, as* $t \to \infty$ *the empirical distribution of play* z^t *converges almost surely to the set of correlated equilibrium distributions of the game* Γ*, that is,* $\mathbb{P}(\mathbf{z}_t \to \Psi) = 1$.

Proof sketch. They two key elements of the convergence proof of regret matching are that: (i) by Theorem A and the associated Proposition in Hart and Mas-Colell [2000], the set of correlated equilibria are exactly the set of distributions over joint actions with zero regret; and (ii) the set of correlated equilibria is non-empty and compact, and therefore the set of approximate δ–correlated equilibria always has positive measure. Building on this, Hart and Mas-Colell [2001] state three further properties of regret matching, which we use to show the convergence of a Q–learning variant of regret matching.

First, the standard regret matching procedure does not need to be employed by the agents from the outset of the game, so that, initially, any finite number of time–steps where play is arbitrary could precede the use of regret matching and play would still converge to the set of correlated equilibria. We can use this property in conjunction with Lemma 4.1 to analyse how Q–learning regret matching behaves in games with initially unknown rewards. Specifically, after an almost–surely finite time T, the reward estimates will have indistinguishable better reply sets to the game in true mean rewards, and so will produce behaviour that is an ε–perturbation of that produced by the true mean rewards (it is perturbed by the small differences in the true and estimated rewards on the probabilities in Definition 5.20). Thus, after this time we can treat Q–learning regret matching as an $\varepsilon(t)$–perturbation of standard regret-matching, and moreover, we can treat the T time–steps required to learn the rewards "accurately enough" as arbitrary initial play; all time steps subsequently discussed are beyond this T.

Second, when the play of regret matching is perturbed by applying uniform sampling with a fixed ε, all regret values approach $\delta(\varepsilon) > 0$, and consequently, the historical frequency of play, $\mathbf{z}^t$, approaches a $\delta(\varepsilon)$–correlated equilibrium. Furthermore, for different fixed values of ε, this distance $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$. Thus, in Q–learning regret matching, if from time t the sampling purturbation is fixed so that $\varepsilon(t+w) = \varepsilon(t)$, then the process will converge to a $\delta(\varepsilon(t))$–correlated equilibrium, as in the standard procedure.

Third, with a decreasing ε, the set that standard regret matching approaches can be "shrunk" as $\varepsilon(t) \to 0$. This is shown by considering the approachable regret value $\delta(\varepsilon(t))$ and its associate approximate correlated equilibria for a large–enough block of transitions $t+w$, then resetting $\delta(\varepsilon(t+w))$, and so on.

Indeed, this is actually the same technique employed in the convergence proof for the standard, unperturbed regret matching proceedure [Hart and Mas-Colell, 2000]. Specifically, in that proof, the action transition probabilities over a block of time-steps from t to $t+w$ are approximated with the w^{th} power of the transition matrix at t (i.e. a stationary process). Regrets in the approximating process are shown to move towards zero over the block's duration. Using a judicious choice of w, the difference between the approximation and the true process is shown to vanish as, at each reset of the block length, both t and w go to ∞. In this step of the proof, the presence of inertia in the players choice plays a key role, as the transition matrix always has strictly positive diagonal elements; the same holds for process like Q–learning regret matching with vanishing purturbations. Now, repeatedly applying the approximating process shows that the regrets approach zero as $t \to \infty$, and because the difference between the approximate and true processes can be shown to vanish, the true regrets also approach zero. Since the set of correlated equilibria are equivalent to the of set distributions over joint

actions with no positive regret, the historical frequency of play of the process converges to the set of correlated equilibria. □

6. Experimental validation. We now validate our theoretical results by comparing our algorithms' performances to other algorithms proposed for games with noisy rewards. In doing so, we investigate three different learning policies employed by the agents, which represent three different orders of magnitude on the rate at which the sampling probability anneals to zero. These three settings correspond to satisfying both the Q–learning and weak ergodicity conditions (Lemma 4.1 and Theorem 5.11 or 5.18), only the Q–learning convergence conditions (i.e. only Lemma 4.1), or neither condition. This is done in order to directly examine the usefulness of our theoretical understanding of how the algorithms explore action space, update their reward estimates and adapt their actions. Thus, the empirical results in this section are used to highlight the practical consequences of our algorithms' theoretical properties, in contrast to the benchmark algorithms' performances, rather than to explicitly evaluate the performance of various parameter settings.

6.1. Algorithms and benchmarks. We demonstrate adaptive play with memory length 8 and sample size 2 (AP(8,2)), better–replies with inertia using memory length 3 and $\xi = 0.3$ (BRI(3,0.3)), and regret matching (RM). We choose the memory lengths for APand BRI so that the effect of these values on the decay of the sampling probability can be observed. The inertia for BRI is chosen as it represents the middle of a region of relative equal performance, ranging from 0.1 to 0.5. For RM, ξ is set to satisfy (5.10), so it has no free parameters. For all the three algorithms, the sampling probabilities were $\varepsilon(t) = 1/8\, t^{-1/Nm}$ for all $i \in N$. We use the same Q–learning parameters for all algorithms and benchmarks throughout the experiments, with $\rho_\lambda = 1$ and $c_\lambda = 0$. We obtain similar results to those presented here for a range of configurations of these three algorithms.

We compare these algorithms to six benchmarks. In the following, the suffix "-A" means that the algorithm uses the ε–greedy policy with the schedule $\varepsilon(t) = 1/8t^{-1/N}$, which satisfies Lemma 4.1 (Q–value convergence), but does not satisfy our conditions guaranteeing weak ergodicity. The suffix "-B" means that the algorithm uses a Boltzmann learning policy:

$$\mathbb{P}(a_i^t = a_i') = \frac{e^{Q_i^t(a_i', \mathbf{z}_{-i})/\eta(t)}}{\sum_{a_i \in A_i} e^{Q_i^t(a_i, \mathbf{z}_{-i})/\eta(t)}}$$

with the temperature parameter following $\eta(t) = 16(0.9^t)$. This learning policy does not satisfy Lemma 4.1, so the the Q–values are not guaranteed to converge.

Two of the benchmarks are variants of BRI with the different sampling schedules, BRI-A and BRI-B. We examine these to directly test the effect of violating either the conditions for weak ergodicity or Q–value convergence.

The next two benchmarks are variants of the *joint action learner* (JAL) of Claus and Boutilier [1998], which is based on standard fictitious play (as discussed in Section 2). The JAL benchmarks use a belief update very similar to that of RM, except that an agent separately stores each other agents' joint action frequencies; that is, i updates each z_j, $j \neq i$ individually. As in our algorithms, the agents use Q–learning to estimate the rewards from each joint action, as in (4.1). Then, the expected value of i's action, $Q_i^t(a_i, z_{-i})$, given its (joint) beliefs z_{-i}, is computed by (3.1), where the Q_i^t takes the place of r_i and (a_i, z_{-i}) that of σ. The specific variants are called JAL-A and JAL-B and use the sampling schedules described above. Indeed, the schedule that $\eta(t)$ follows under JAL-B is chosen because it is the one used by Claus and Boutilier [1998] in their original JAL description.

The final two benchmarks are *independent action learners* (IAL-A and IAL-B), under which agents use Q–learning to estimate the their rewards for their own actions, oblivious to

		Matt			
		Colin		*Colin*	
		Left	Right	Left	Right
Rowena	Up	(5,5,5)	(0,1,0)	(0,0,1)	(1,0,0)
	Down	(1,0,0)	(0,0,1)	(0,1,0)	(2,2,2)

Fig. 1: Three–player potential game.

others' actions [Claus and Boutilier, 1998; Cominetti et al., 2010]. We investigate these two to demonstrate the effect of ignoring other agents in games with unknown noisy rewards. If the reward distributions were stationary or the setting was a single–agent learning problem, then the sampling schedule for IAL-A would converge slower than is necessary for the reward estimates to converge to their true values. However, an agent's rewards may well be nonstationary (as a result of other players changing their actions), and we wish to test if this schedule can account for any nonstationarity. In IAL-B, proposed by Claus and Boutilier [1998] at the same time as JAL-B, the sampling probability is driven to zero relatively quickly.

6.2. Test problem and results. We compare the algorithms in a three–player two–action potential game, so that their behaviours can be clearly contrasted and their differences can be transparently analysed, without complications from a complex game setting. Mean rewards for the game are given in Figure 1, in which Rowena selects the row, Colin the column and Matt the matrix. The agents receive rewards equal to these values plus uniform noise $e \in [-\bar{e}, \bar{e}]$, as in Equation 3.2, where $\bar{e}$ itself is uniformly drawn from $[5, 10]$ at the beginning of each scenario. The game in mean rewards has two strict Nash equilibria. The Nash equilibrium located at (U, L, L) is globally optimal, while the other at (D, R, R) is suboptimal; the same number and length of best response paths lead to each one. The metric of interest is the probability of converging to different Nash equilibria, and the mean frequencies of convergence to each equilibrium by the algorithms were recorded for 50 repetitions of 50 scenarios, generated randomly as described above. We consider a duration of 800 time–steps, not because all algorithms converge in this time, but because most interesting behaviour occurs during this period and the clearest differentiations can be made. Since the algorithms are only guaranteed to converge to a strict Nash equilibrium, and not necessarily to the optimum, we also use this game to informally investigate the quality of their solutions.

The results are given in the plots in Figure 2, which illustrate the proportion of play is the optimal Nash equilibrium (dark), suboptimal Nash equilibrium (light) or a non–equilibrium outcome (medium) is played, over time (standard errors were too small to plot). The bold dashed or dotted lines on the plots show the same proportions for the agents' intended play; that is, the actions given by following their unperturbed dynamics, rather than sampling (e.g. with probability $\varepsilon(t)$), at that time–step. The distance between the actual and intended play of a Nash equilibrium gives the proportion of non–equilibrium play that is due to the sampling induced by the learning policy.

Results for our three novel algorithms are on the top row of Figure 2. The most noticeable feature is that AP, BRI and RM all converge towards a Nash equilibrium in a high proportion of simulations: 83%, 86% and 83% of actual play at $t = 800$, respectively, and more than 96% of intended play for both AP and BRI, and 90% for RM. These very high proportions validate our convergence results for these algorithms. Additionally, the proportion of the globally optimal play is high, at greater than 80% for all three. Compared to BRI, the (slightly) lower proportion converged to equilibrium for AP is expected, as the ε schedule goes to zero more slowly because of its longer memory. The relatively lower proportion for RM is expected because it converges to the set of correlated equilibria, which are a superset of the Nash equilibria. A fairer measurement of the convergence of RM is the proportion of runs in which it finds a correlated equilibrium; analysis of the results show that RM is within 1% of the

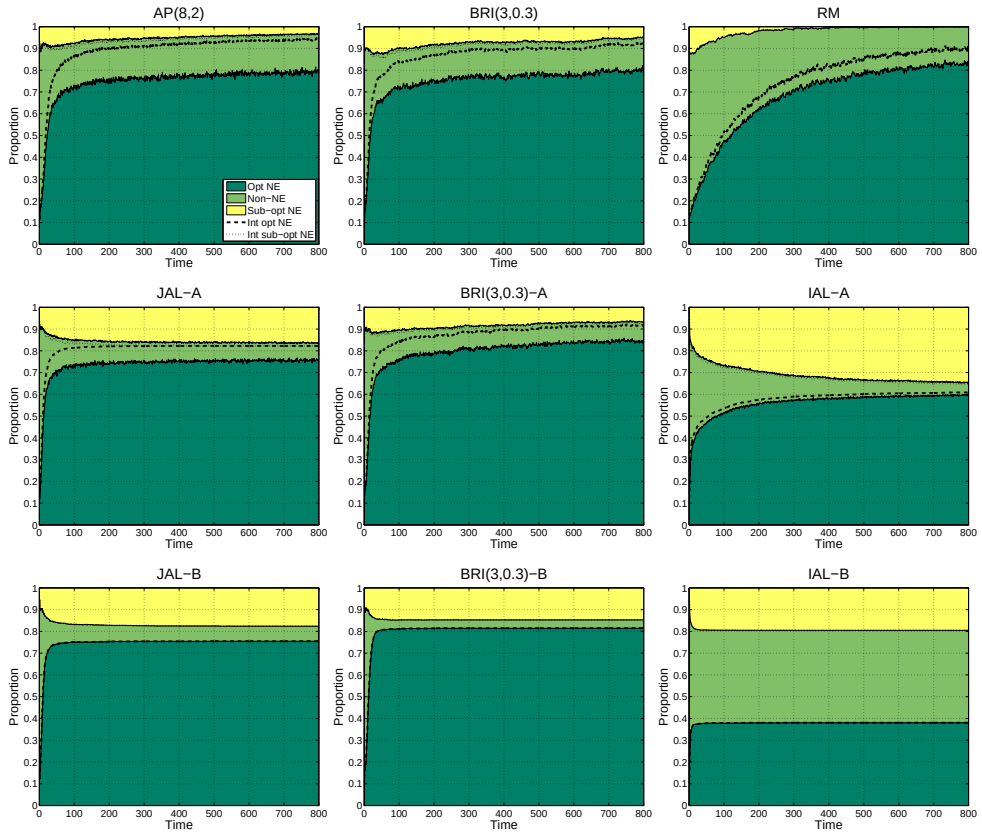

Fig. 2: Action proportions over time, showing the optimal, suboptimal and non–Nash equilibrium play, and intended play (without sampling) superimposed.

payoff of a correlated equilibrium over the period $t = 600$ to 800 in more than 95% of runs. This is consistent with convergence to an approximate correlated equilibrium, and validates our convergence results for RM.

The remaining 6 plots regard benchmarks. The results for JAL-A and BRI-A are particularly interesting, since they show almost as good performance as that of our algorithms. The -A sampling schedule satisfies Lemma 4.1, so the algorithms' Q–values should converge. This seems to be sufficient for JAL-A to converge (recalling that fictitious play converges to a Nash equilibrium in potential games), although it does converge to the optimal Nash equilibrium in only 76% of runs, which is significantly less often than AP, BRI or RM. On the other hand, the convergence of BRI-A may be a result of the game in question admitting a more compact reduction in the states over which its weak ergodicity can be shown (as noted in the discussion of weak ergodicity in Lemma 5.6); however, this could not have been known to the agents before they began to play the game.

Regarding JAL-B and BRI-B, the fact that the Q–values do not converge under the -B sampling schedule is illustrated in their quick "freezing" into fixed proportions of play, with actual and intended play taking almost the same values. This is because they become mired with incorrect reward estimates in non–equilibrium outcomes, and do not sample new actions frequently enough to learn the true better replies in the game. Even though the algorithms do reach Nash equilibrium outcomes in a good proportion of runs (approximately 92% and

96%, respectively), a set of simulations with a duration of 2500 time–steps showed that these proportions do not improved noticeably beyond these levels.

Finally, both IAL variants suffer from not incorporating the actions of other agents. IAL-A does, in fact, continue to improve its proportion of Nash equilibrium convergence over the longer term: After 2500 time–steps it had a total Nash–converged proportion similar to JAL-B over the same time, and over 10,000 time–steps, it had further reduced this to about 97%, which is comparable to the intended play of AP, BRI and RM, albeit over an order of magnitude longer duration. Nonetheless, the -A sampling schedule does appear to be sufficient for its Q–values to eventually converge. On the other hand, JAL-B freezes into fixed proportions of play within 100 time–steps at a very low proportion of Nash equilibrium convergence ($< 60\%$), and does not improve over longer durations.

These benchmark results are consistent with our theoretical analysis, and correspond with our understanding of the conditions under which the algorithms should converge. In particular, the fact that, by Lemma 4.1, the -A schedule is sufficient for the Q–values to converge appears to enable the algorithms not directly covered by our theoretical results to perform well. In contrast, the -B schedule is not sufficient for the convergence of reward estimates, which prevents the associated algorithms from converging. Collectively these results indicate that accurate learning of rewards, coupled with principled reasoning over and appropriate responses to opponents' actions, are sufficient to drive convergence to equilibrium in practice. On the other hand, dropping either Q–value convergence (i.e. using schedule -B) or explicit reasoning over opponent actions (as in the IAL variants) prevent play from converging at an acceptable rate.

7. Conclusions. In this paper, we proved the convergence to Nash equilibria of variants of adaptive play and the better–reply processes in potential games and other more general acyclic games with rewards that are initially unknown and which must be estimated over time from noisy observations. We also derived a Q–learning variant of regret matching, and proved its almost sure converge to the set of correlated equilibria. Finally, the necessity of the conditions on the algorithms' sampling rates that we derived were empirically verified. Our results guarantee the convergence of several distributed optimisation methods, for settings where reward functions cannot be prespecified and that have constraints on communication between system components.

There are a number of ways in which this work may be taken forward. One particularly interesting direction is to put finite–time bounds on the algorithms' performance, by employing different frameworks for analysing online learning of noisy rewards and the consequent convergence of an algorithm to Nash equilibrium, such as PAC or KWIK learning [Valiant, 1984; Li et al., 2008]. A second opportunity is to extend the convergence of the algorithms to more complicated settings, and, in particular, we are interested in settings where the payoffs in the game vary according to some state variable, such as is addressed for individual agents in the growing literature on contextual multi–armed bandits and multi–armed bandits with co-variates [Lu et al., 2010].

References.

Anily, S. and Federgruen, A. (1987). Ergodicity in parametric nonstationary Markov chains: An application to simulated annealing methods. *Operations Research*, 35(6):867–874.

Arslan, G., Marden, J. R., and Shamma, J. S. (2007). Autonomous vehicle-target assignment: A game theoretical formulation. *Journal of Dynamic Systems, Measurement, and Control*, 129(5):584–596.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science (FOCS '95)*, pages 322–331, Washington, DC, USA. IEEE Computer Society.

Baños, A. (1968). On pseudo–games. *The Annals of Mathematical Statistics*, 39:1932–1945.

Blackwell, D. (1956). An analogue of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8.

Brown, G. W. (1951). Iterative solution of games by fictitious play. In Koopmans, T. C., editor, *Activity Analysis of Production and Allocation*, pages 374–376. John Wiley & Sons, Inc., New York.

Chapman, A. C., Micillo, R. A., Kota, R., and Jennings, N. R. (2010). Decentralised dynamic task allocation using overlapping potential games. *The Computer Journal*, 53(9):1462–1477.

Chapman, A. C., Rogers, A., Jennings, N. R., and Leslie, D. S. (2011). A unifying framework for iterative approximate best response algorithms for distributed constraint optimisation problems. *The Knowledge Engineering Review*, 26(4):411–4.

Claus, C. and Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multiagent systems. In *In Proceedings of the 15th AAAI National Conference on Artificial Intelligence*, pages 746–752. AAAI Press.

Cominetti, R., Melo, E., and Sorin, S. (2010). A payoff-based learning procedure and its application to traffic games. *Games and Economic Behavior*, 70(1):71–83. Special Issue In Honor of Ehud Kalai.

Foster, D. P. and Young, H. P. (2006). Regret testing: learning to play Nash equilibrium without knowing you have an opponent. *Theoretical Economics*, 1:341–367.

Friedman, J. W. and Mezzetti, C. (2001). Learning in games by random sampling. *Journal of Economic Theory*, 98(1):55–84.

Fudenberg, D. and Levine, D. K. (1998). *The Theory of Learning in Games*. MIT Press, Cambridge, MA.

Gottlob, G., Greco, G., and Scarcello, F. (2005). Pure Nash equilibria: Hard and easy games. *Journal of Artificial Intelligence Research*, 24:357–406.

Groves, T. (1973). Incentives in Teams. *Econometrica*, 41(4):617–631.

Hart, S. and Mas-Colell, A. (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68:1127–1150.

Hart, S. and Mas-Colell, A. (2001). *A reinforcement procedure leading to correlated equilibrium*, volume VIII, pages 181–200.

Hofbauer, J. and Sandholm, W. H. (2007). Evolution in games with randomly disturbed payoffs. *Journal of Economic Theory*, 132(1):47 – 69.

Isaacson, D. and Madsen, R. (1976). *Markov Chains: Theory and Applications*. John Wiley & Sons, New York.

Jaakkola, T., Jordan, M. I., and Singh, S. P. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6(6):1185–1201.

Jennings, N. R. (2001). An agent-based approach for building complex software systems. *Communications of the ACM*, 44(4):35–41.

Kandori, M., Mailath, G. J., and Rob, R. (1993). Learning, mutation, and long run equilibria in games. *Econometrica*, 61(1):29–56.

Kearns, M., Littman, M., and Singh, S. (2001). Graphical models for game theory. In *Proceedings of the 17th on Uncertainty in Artificial Intelligence (UAI–01)*, pages 253–260. Morgan Kaufmann.

Leslie, D. S. and Collins, E. J. (2005). Individual Q-learning in normal form games. *SIAM Journal on Control and Optimization*, 44:495–514.

Leslie, D. S. and Collins, E. J. (2006). Generalised weakened fictitious play. *Games and Economic Behavior*, 56:285–298.

Li, L., Littman, M. L., and Walsh, T. J. (2008). Knows what it knows: a framework for self-aware learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML '08)*, pages 568–575.

Lu, T., Pál, D., and Pál, M. (2010). Contextual multi-armed bandits. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS '10)*, pages 485–492.

Marden, J. R., Young, H. P., Arslan, G., and Shamma, J. S. (2009). Payoff–based dynamics for multi–player weakly acyclic games. *SIAM Journal on Control and Optimization*, 48:373–396.

Marschak, J. and Radner, R. (1972). *Economic Theory of Teams*. Yale University Press, New Haven and London.

Mertikopoulos, P. and Moustakas, A. L. (2010). The emergence of rational behavior in the presence of stochastic perturbations. *The Annals of Applied Probability*, 20(4):1359–1388.

Monderer, D. and Shapley, L. S. (1996). Potential games. *Games and Economic Behavior*, 14:124–143.
Papadimitriou, C. H. and Roughgarden, T. (2008). Computing correlated equilibria in multi–player games. *Journal of the ACM*, 55(3):14:1–14:29.
Rosenthal, R. W. (1973). A class of games possessing pure–strategy Nash equilibria. *International Journal of Game Theory*, 2:65–67.
Scutari, G., Barbarossa, S., and Palomar, D. P. (2006). Potential games: A framework for vector power control problems with coupled constraints. In *31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, volume 4, pages 241–244.
Singh, S. P., Jaakkola, T., Littman, M. L., and C. Szepesvári (2000). Convergence results for single–step on–policy reinforcement–learning algorithms. *Machine Learning*, 38(3):287–308.
Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning*. MIT Press, Cambridge, MA.
Touri, B. and Nedić, A. (2010). When infinite flow is sufficient for ergodicity. In *Proceedings of the 49th IEEE Conference on Decision and Control (CDC '10*, Atlanta, USA.
Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27:1134–1142.
Wardrop, J. G. (1952). Some theoretical aspects of road traffic research. *Engineers*, Part II:325–378.
Wolpert, D. H. and Tumer, K. (2002). Collective intelligence, data routing and Braess' paradox. *Journal of Artificial Intelligence Research*, 16:359–387.
Young, H. P. (1993). The evolution of conventions. *Econometrica*, 61:57–84.
Young, H. P. (2004). *Strategic Learning and its Limits*. Oxford University Press.
Young, H. P. (2009). Learning by trial and error. *Games and Economic Behavior*, 65:626–643.

Appendix A. In Section 4 we investigated a scenario where the individuals attempt to estimate their expected reward $r_i(\mathbf{a})$ for each joint action in A. However, in the *standard normal form* considered to date, the joint action space A grows exponentially with the number of agents, so this estimation problem becomes impractical. However in systems with an inherent structure, such as those with a natural spatial structure in which interaction only directly occurs between geographically close individuals, agents should only need to consider the actions of their neighbours. We now show that, if a game admits a compact form, then this representation can be exploited to improve the agents' learning rates, for two compact forms.

This first is *graphical normal form* (GNF), which can represent games in which some agents' rewards are independent of others' strategies [Kearns et al., 2001]. In this form, the nodes of a graph correspond to the set of agents, while edges connect an agent to the others with which it shares a reward dependency, called its neighbours. The neighbourhood of i is the smallest set ν_i of players such that agent i's reward is entirely determined by a_i and $\{a_j : j \in \nu_i\}$. We say an undirected reward dependency exists between i and $j(\neq i)$ if either $j \in \nu_i$ or $i \in \nu_j$.

DEFINITION A.1. *A game in GNF comprises a set of agents located on the nodes of a graph. An agent is connected to those with which it shares an undirected reward dependency, which includes its set of neighbours* $\nu_i \subseteq N$. *Its reward function,* $r_i(\mathbf{a}_{i,\nu_i})$, *is then given by an array indexed by tuples from the set* $\times_{j\in\{i,\nu_i\}}|A_j|$. *Games in* GNF with unknown noisy rewards *are defined similarly, with the difference being that when the joint action* $\mathbf{a} \in A$ *is played, agent i receives the reward*

$$R_i = r_i(a_i, \mathbf{a}_{\nu_i}) + e_i, \tag{A.1}$$

where $r_i(a_i, \mathbf{a}_{\nu_i})$ *is the true expected reward to agent i for the joint action* $(a_i, \mathbf{a}_{\nu_i})$, *and* e_i *is a random variable with zero mean and bounded variance.*

Note that in GNF, $r_i(\mathbf{a})$ depends only on a_i and $\mathbf{a}_{\nu_i}$, where $\mathbf{a}_{\nu_i}$ is the joint action of all the neighbours of i. Subsequently, we write r_i as a function of the joint actions of i and its neighbours, that is, $r_i(\mathbf{a}_{i,\nu_i})$. Also, note that games in standard normal form can be represented in GNF with a complete graph.

For games in GNF, each agent needs to learn only its rewards over it and its neighbours' joint action spaces, given by: $A_{i,\nu_i} = A_i \times_{j\in\nu_i} A_j$. For large games, this is a much more feasible task than estimating the full reward function on A. Each individual i now updates its estimates Q_i^t using the equation:

$$Q_i^{t+1}(\mathbf{a}_{i,\nu_i}) = Q_i^t(\mathbf{a}_{i,\nu_i}) + \lambda(t) I\{\mathbf{a}_{i,\nu_i}^t = \mathbf{a}_{i,\nu_i}\}\left(R_i^t - Q_i^t(\mathbf{a}_{i,\nu_i})\right) \quad \forall \mathbf{a}_{i,\nu_i} \in A_{i,\nu_i}. \tag{A.2}$$

In this case, the sequence $\{\varepsilon(t)\}_{t\to\infty}$ can be altered to take advantage of the reduced size of each agent's joint action space, while still ensuring that each Q–value is updated infinitely often.

LEMMA A.2. *In a game in GNF, let i's* neighbourhood size *be the number of neighbours of i plus 1 for i itself. Given this, let J_i be the size of the largest of the neighbourhoods of i or any j in ν_i. In a game with unknown noisy rewards, if agents select their actions using a policy in which, for all $i \in N$, $a_i \in A_i$ and $t \geq 1$,*

$$\mathbb{P}(a_i^t = a_i) \geq \varepsilon_i(t), \quad \textit{with} \quad \varepsilon_i(t) = c_\varepsilon t^{-1/J_i},$$

where $c_\varepsilon > 0$ is a positive constant, then $\forall i \in N$, $\forall \mathbf{a}_{i,\nu_i} \in A_{i,\nu_i}$:

$$\lim_{t\to\infty} |Q_i^t(\mathbf{a}_{i,\nu_i}) - r_i(\mathbf{a}_{i,\nu_i})| = 0 \quad \textit{with probability 1.} \tag{A.3}$$

Proof. If $\mathbb{P}(a_i^t = a_i) \geq c_\varepsilon t^{-1/J_i}$, then $\forall \mathbf{a}_{i,\nu_i} \in A_{i,\nu_i}$, a lower bound on $\mathbb{P}(\mathbf{a}_{i,\nu_i}^t = \mathbf{a}_{i,\nu_i})$ is:

$$\prod_{j\in\{i\}\cup\nu_i} c_\varepsilon t^{-1/J_j} \geq \left(c_\varepsilon t^{-1/(|\nu_i|+1)}\right)^{|\nu_i|+1} = (c_\varepsilon)^{|\nu_i|+1} t^{-1},$$

because $J_j \geq |\nu_i| + 1$. The result follows from observing that $\sum_{t=0}^{\infty}(c_\varepsilon)^{|\nu_i|+1}t^{-1} = \infty$ within each neighbourhood's joint action space A_{i,ν_i}. □

The second useful compact representation is *hypergraphical normal form* (HNF) [Gottlob et al., 2005; Papadimitriou and Roughgarden, 2008], which comprises hyperedges representing a set of local games that each contain several agents. An agent is typically involved in more than one local game, and its neighbours are those it is linked to via any local game.

DEFINITION A.3. *A game in HNF comprises a set of agents located on the nodes of a hypergraph. Each hyperedge represents a local game: $\Gamma = \{\gamma_1, \gamma_2, ...\}$, where $\gamma = \langle N_\gamma, \{A_i, r_{i,\gamma}\}_{i\in N_\gamma}\rangle$, defined as in SNF. Let $\Gamma_i = \{\gamma : i \in N_\gamma\}$ be the set of local games containing agent i. Player i's action set, A_i, is identical in all $\gamma \in \Gamma_i$, and it selects a single action $a_i \in A_i$ to play in all of its local games. Its neighbours in $\gamma \in \Gamma_i$ are $\nu_{i,\gamma} = N_\gamma \setminus i$, and its reward from γ, $r_{i,\gamma}(\mathbf{a}_\gamma)$ is given by an array indexed by tuples from the set $\times_{j\in N_\gamma}|A_j|$. Its full set of neighbours is given by $\nu_i = \cup_{\gamma\in\Gamma_i} N_\gamma \setminus i$, and its reward is the sum of its rewards from $\gamma \in \Gamma_i$: $r_i(a_i, \mathbf{a}_{i,\nu_i}) = \sum_{\gamma\in\Gamma_i} r_{i,\gamma}(a_i, \mathbf{a}_{\nu_{i,\gamma}})$, where $\mathbf{a}_{\nu_{i,\gamma}}$ is the joint action of i′s neighbours in γ. For games in* HNF with unknown noisy rewards*, when the joint action $a \in A$ is played, agent i receives the (independently observable) rewards*

$$R_{i,\gamma} = r_{i,\gamma}(\mathbf{a}_\gamma) + e_{i,\gamma} \quad \forall\gamma \in \Gamma_i, \tag{A.4}$$

where $r_{i,\gamma}(\mathbf{a}_\gamma)$ is the true expected reward to agent i from local game γ for the joint action $\mathbf{a}_\gamma$, and each $e_{i,\gamma}$ is a random variable with zero mean and bounded variance.

Again, since $r_i(\mathbf{a})$ now only depends on a_i and $\mathbf{a}_{\nu_i}$, we write $r_i(\mathbf{a}_{i,\nu_i})$. Any games in standard normal form can be represented in HNF with a single local game γ.

In a game in HNF, each agent can learn the payoffs for joint actions in each of its local games independently. Hence, an individual i now updates its estimate $Q_{i,\gamma}^t$ of its reward function for each γ using the equation:

$$Q_{i,\gamma}^{t+1}(\mathbf{a}_\gamma) = Q_{i,\gamma}^t(\mathbf{a}_\gamma) + \lambda(t) I\{\mathbf{a}_\gamma^t = \mathbf{a}_\gamma\}\left(R_{i,\gamma}^t - Q_{i,\gamma}^t(\mathbf{a}_\gamma)\right) \quad \forall \mathbf{a}_\gamma \in A_\gamma. \tag{A.5}$$

For games in HNF, each joint action in each local game is guaranteed to be sampled infinitely often by following the $\{\varepsilon(t)\}_{t\to\infty}$ schedule given in the following Lemma.

LEMMA A.4. *In a game in HNF, let J_i be the maximum number of participants in any single local game in Γ_i (i.e. $J_i = \max_{\gamma\in\Gamma_i}|N^\gamma|$). In a game with unknown noisy rewards, if agents select their actions using a policy in which, for all $i \in N$, $a_i \in A_i$ and $t \geq 1$,*

$$\mathbb{P}(a_i^t = a_i) \geq \varepsilon_i(t), \quad \textit{with} \quad \varepsilon_i(t) = c_\varepsilon t^{-1/J_i},$$

where $c_\varepsilon > 0$ is a positive constant, then

$$\lim_{t\to\infty} |Q_{i,\gamma}^t(\mathbf{a}_\gamma) - r_{i,\gamma}(\mathbf{a}_\gamma)| = 0 \quad \textit{with probability 1, } \forall i \in N,\ \forall\gamma \in \Gamma_i,\ \forall \mathbf{a}_\gamma \in A_\gamma. \tag{A.6}$$

The proof follows identically that of Lemma A.2 but with the appropriate definition of J_i for HNF, and by observing that $\sum_{t=0}^{\infty}(c\varepsilon)^{|\nu_i|+1}t^{-1} = \infty$ within each local game's joint action space A_γ.

We have now derived techniques for estimating an agent's reward functions that can overcome the computational problems associated with learning rewards in large games by exploiting structured interaction between the agents. When interleaved with a suitable strategy adaptation process, this will result in an algorithm that learns all rewards accurately and converges to equilibrium in games with unknown noisy rewards.

DEFINITION A.5 (Compact Q–learning variants). *Each of compact Q–learning adaptive play, better–replies with inertia, and regret matching is defined as their non–compact counterparts (i.e. Definitions 5.4, 5.15 and 5.21), but with $\varepsilon_i(t) = ct^{-1/J_i\lceil m/2\rceil}$, where J_i is the neighbourhood size as defined in GNF or HNF.*

COROLLARY A.6. *Let Γ be a game with unknown noisy rewards:*

- *If Γ has mean rewards that are weakly acyclic under best replies (WAG), then if $k \leq m/(L_\Gamma + 2)$ compact Q–learning adaptive play almost surely converges in round–by–round behaviour to a pure Nash equilibrium in Γ.*
- *If Γ has mean rewards that are weakly acyclic under better replies (WABRG), then compact Q–learning better–replies with inertia almost surely converges in round–by–round behaviour to a pure Nash equilibrium in Γ.*

Moreover, compact Q–learning adaptive play and Q–learning better–replies with inertia all converge to a pure Nash equilibrium in games with noisy unknown rewards with mean rewards that are generic and admit a potential function. Finally, if Γ has mean rewards that are generic, then compact Q–learning regret matching almost surely converges in round–by–round behaviour to a correlated equilibrium in Γ.