Available online at www.sciencedirect.com
SCIENCE DIRECT®
Artificial Intelligence 159 (2004) 27–47
Artificial Intelligence
www.elsevier.com/locate/artint

Efficient learning equilibrium ☆

Ronen I. Brafman [a], Moshe Tennenholtz [b,*]

[a] *Computer Science Department, Ben-Gurion University, Beer-Sheva, Israel*
[b] *Industrial Eng. & Management, Technion, Haifa, Israel 32000*

Received 5 March 2003; accepted 19 April 2004

Abstract

We introduce *efficient learning equilibrium* (ELE), a normative approach to learning in non-cooperative settings. In ELE, the learning algorithms themselves are required to be in equilibrium. In addition, the learning algorithms must arrive at a desired value after polynomial time, and a deviation from the prescribed ELE becomes irrational after polynomial time. We prove the existence of an ELE (where the desired value is the expected payoff in a Nash equilibrium) and of a Pareto-ELE (where the objective is the maximization of social surplus) in repeated games with perfect monitoring. We also show that an ELE does not always exist in the imperfect monitoring case. Finally, we discuss the extension of these results to general-sum stochastic games.

Keywords: Learning equilibrium; Ex-post equilibrium; Efficiency; Multi-agent learning; Repeated games; Stochastic games

1. Introduction

Reinforcement learning in the context of multi-agent interaction has attracted the attention of researchers in cognitive psychology, experimental economics, machine learning, artificial intelligence, and related fields for quite some time [7,18]. Much of this work uses repeated games [6,11] and stochastic games [3,17,22,26] as models of such

☆ A preliminary short version of this paper appeared at NIPS'02. This research was supported by the Israel Science Foundation under grant #91/02-1. The first author is partially supported by the Paul Ivanier Center for Robotics and Production Management.
* Corresponding author.
E-mail address: moshet@ie.technion.ac.il (M. Tennenholtz).

0004-3702/$ – see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.artint.2004.04.013

interactions. The literature on learning in games in game theory [11] is mainly concerned with the understanding of learning procedures that **if** adopted by the different agents will converge at the end to an equilibrium of the corresponding game. The game itself may be known; the idea is to show that simple dynamics lead to rational behavior, as prescribed by a Nash equilibrium. The learning algorithms themselves are not required to satisfy any rationality requirement; it is what they converge to, **if** adopted by **all** agents that should be in equilibrium.

When facing uncertainty about the game that is played, game-theorists adopt a Bayesian approach. The typical assumption in that approach is that there exists a probability distribution on the possible games, which is common-knowledge. The notion of equilibrium is extended to this context of games with incomplete information, and is treated as the appropriate solution concept. In this context, agents are assumed to be rational agents adopting the corresponding (Bayes-) Nash equilibrium, and learning is not an issue.

Our major claim is that the game-theoretic approach is not in line with the goals of multi-agent reinforcement learning research in AI and must be modified. First, the Bayesian approach used to model partial information is not in line with the common approach in theoretical computer science and computational learning for dealing with uncertainty. Second, the descriptive motivation underlying learning research in game-theory differs considerably from the normative motivation for learning research in AI, and these differences have important ramifications. We now explain these issues in more detail.

First, consider the Bayesian model of partial information. To date, most work in machine learning, and in particular, work on single-agent reinforcement learning has taken a different approach, motivated largely by work on online algorithms in computer science. Here, no distribution is assumed over the uncertain entities, and instead, our goal is to approach the behavior of an agent with complete information as closely and as quickly. Indeed, AI researchers have adopted this non-Bayesian approach in their work on learning in games, looking for algorithms that converge to an appropriate equilibrium in any game out of a class of relevant games; and we follow suit. However, researchers in multi-agent reinforcement learning did choose to adopt other assumptions made by game-theorists, despite the fact that here the differences are much more fundamental.

Work on learning in games started with descriptive motivation in mind. That is, its goal was to show that people who use simple heuristic rules for updating their behavior in a multi-agent setting (i.e., in a game) will eventually adopt behavior that corresponds to some appropriate equilibrium behavior. If that is the case, economic models based on equilibria concepts are, in some sense, justified. The assumption that all agents use the same learning rule is justified by the fact that all agents involved are people—i.e., they are all designed similarly. But in AI we are not concerned with descriptive models of human behavior—we are interested in designing artificial agents. Except in the case of cooperative systems, we have no reason to believe that agents designed by different designers will all employ the same learning algorithms. Moreover, one should view the designer's choice of learning algorithm for its agent as a fundamental decision that should follow normative criteria. Indeed, from the AI perspective, the choice of a learning algorithm is a basic action we take in a game we play against other agent designers.

There is another related point. Game-theorists adopting the descriptive stance are not too concerned with how quickly a learning rule leads to convergence—after all, we had ages

to evolve our behavior. But an agent designer wants its agent to learn quickly. He does not care about its agent's "offsprings". Thus, in AI, speed of convergence is of paramount importance.

To better align the research methodology in multiagent reinforcement learning with the AI perspective, we present in this paper a non-Bayesian normative approach to learning in games. Our approach makes no assumptions about the distribution of possible games that may be played—making it more reflective of the setting studied in machine learning and AI and in the spirit of work on online algorithms in computer science—and treats the choice of a learning algorithm itself as a game. More specifically, we adopt the framework of repeated games, and view the learning algorithm as a strategy for an agent in a repeated game. This strategy takes an action at each stage based on its previous observations, and initially has no information about the identity of the game being played.

Given the above, the following are natural requirements for the learning algorithms provided to the agents:

1. *Individual rationality*: The learning algorithms themselves should be in equilibrium. It should be irrational for each agent to deviate from its learning algorithm, as long as the other agents stick to their algorithms, *regardless* of what the actual game is.
2. *Efficiency*:
   (a) A deviation from the learning algorithm by a single agent (while the others stick to their algorithms) will become irrational (i.e., will lead to a situation where the deviator's payoff is not improved) after polynomially many stages.
   (b) If all agents stick to their prescribed learning algorithms then the expected payoff obtained by each agent within a polynomial number of steps will be (close to) the value it could have obtained in a Nash equilibrium, had the agents known the game from the outset.

A tuple of learning algorithms satisfying the above properties for a given *class* of games is said to be an *Efficient Learning Equilibrium* (ELE). Notice that the learning algorithms should satisfy the desired properties for *every* game in a given class despite the fact that the actual game played is initially unknown. Such assumptions are typical to work in machine learning. What we borrow from the game theory literature is the criterion for rational behavior in multi-agent systems. That is, we take individual rationality to be associated with the notion of equilibrium. We also take the equilibrium of the actual (initially unknown) game to be our benchmark for success; we wish to obtain a corresponding value although we initially do not know which game is played. The idea above constitutes the major conceptual contribution of this paper.

In the following section, we provide a short review of basic notions in game-theory. In the remaining sections we formalize the notion of efficient learning equilibrium, and show that it is not devoid of content, i.e., we prove the existence of an ELE for a general class of games—the class of repeated games with perfect monitoring. We also show that there are classes of games in which an ELE does not exist. Then, we generalize our results to the context of Pareto-ELE (where we wish to obtain maximal social surplus). We also discuss the extension of our results to general-sum stochastic games. Technically speaking, the results we prove rely on a novel combination of the so-called folk theorems in economics,

and a novel efficient algorithm for the punishment of deviators in games which are initially unknown.

2. Basic notions in game-theory

Game-theory provides a mathematical formulation of multi-agent interactions and multi-agent decision making. Here we review some of the basic concepts. For a good introduction to the area, see, e.g., [25].

A game is a formal description of an interaction between a set of agents. The rules of the game describe the order of moves by the agents, the available choices at each move, the information available to each agent at each point, and the final outcome for each agent (described by a pay-off function). In order to abstract away the particular order of moves, so as to be able to conveniently treat different games in a uniform manner, we employ the description of a game in *strategic form*. A game in strategic form consists of a set of players I, a set of actions A_i for each $i \in I$, and a payoff function $R_i : \times_{i \in I} A_i \to R$ for each $i \in I$. We let $\mathcal{A}$ denote the set $\times_{i \in I} A_i$ of *joint actions*. Intuitively, each $a \in A_i$ denotes a complete policy for playing the game for agent i, and describes how the agent would act in each possible situation that could arise in the course of playing the game. Such actions are often referred to as *strategies*. The resulting description is very simple, though not necessarily compact, and we adopt it in the rest of this paper.

When considering the actions an agent has to choose from in a game, we include not only the standard actions (referred to as *pure* actions), but also *mixed* actions, i.e., probability distributions over pure actions. The payoff function is extended to this extended set of actions naturally using the expectation operator.

When there are only two players, the game can be described using a (bi)-matrix whose rows correspond to the possible actions of the first agents and whose columns correspond to the possible actions of the second agent. Entry (i, j) contains a pair of values denoting the payoffs to each agent when agent 1 plays action i and agent 2 plays action j. In the rest of this paper, we concentrate on two-player games. In addition, we make the simplifying assumption that the action set of both players is identical. We denote this set by A. The extension to different sets is trivial.

In Fig. 1 we see a number of examples of two-player games. The first game is a *zero-sum* game, i.e., a game in which the sum of the payoffs of the agents is 0. This is a game of pure competition. The second game is a *common-interest* game, i.e., a game in which the agents receive identical payoffs. The third game is a well-known general-sum game, the prisoners' dilemma. In this case, the agents are not pure competitors nor do they have identical interests.

A basic concept in game-theory is that of a *Nash equilibrium*. A joint action $a \in \mathcal{A}$ is said to be a Nash equilibrium if for every agent i and every joint action a' such that a' differs from a in the action of agent i alone, it is the case that $R_i(a) \geqslant R_i(a')$. Thus, no agent has motivation to unilaterally change its behavior from a. A basic result of game theory is that every n-person game in strategic form, in which the agents' set of action is finite possesses a Nash equilibrium in mixed strategies (where each agent can select a probability distribution of its available actions) [24]. Unfortunately, in general, there can be

$$M_1 = \begin{pmatrix} 5, -5 & 3, -3 \\ -3, 3 & -2, 2 \end{pmatrix}$$
$$M_2 = \begin{pmatrix} 5, 5 & 6, 6 \\ -3, -3 & 2, 2 \end{pmatrix}$$
$$M_3 = \begin{pmatrix} 2, 2 & -10, 10 \\ 10, -10 & -5, -5 \end{pmatrix}$$

Fig. 1.

many Nash equilibria. Thus, while Nash equilibria are stable in some respect, this does not imply that any particular Nash equilibrium corresponds to the "right" or "recommended" behavior for the agents in this game. A major issue in the theory of general-sum games is what is a normatively appropriate behavior in games with diverse equilibria. There are special cases where the Nash equilibrium possesses additional properties that make it attractive. For instance, in zero-sum games, all Nash equilibria provide the same payoff to the agents. Moreover, this payoff is identical to the agent's probabilistic safety level—i.e., the maximal value it can guarantee to itself, regardless of the other agent's actions.

In order to model the process of learning in games, researchers have concentrated on settings in which agents repeatedly interact with each other—otherwise, there is no opportunity for the agent to improve its behavior.[1] In the most popular model, the agents repeatedly play a game, each time observing their reward and, possibly, the other agent's actions. In the classic work on learning in game-theory, the agents select their behavior in the next iteration of the game based on the result of previous iterations using some simple update rule. Typically, these studies had the goal of showing that some simple update rule leads the agents to eventually adopt some Nash equilibrium or that the long-term average behavior of the agents corresponds to some Nash equilibrium. Typically, the goal of such studies has been to justify the use of Nash equilibria in economic modeling. The repeated-games model has been popular with AI researchers too, although more recently, the stochastic, or Markov-game model has attracted the attention of many researchers. In the stochastic game model, the agents also engage in a series of games. However, the games the agents play can be different at each stage. In fact, the nature of each game depends probabilistically on the identity of the previous game and the agents' joint action in that game. We consider both repeated and stochastic games in this paper.

3. Efficient learning equilibrium: definition

In this section we develop a definition of efficient learning equilibrium in the context of two-player repeated games. The generalization to n-player repeated games is immediate,

[1] "Learning" in settings in which the agent has a small number of opportunities to observe the game and improve its behavior is usually modeled as a game with incomplete information. As we noted earlier, in this case the agent is assumed to have a probability distribution over the possible values of the missing information, such as the payoffs. The standard theory has been extended to handle this case.

but requires additional notation, and will not be presented here. An extension to stochastic games appears in Section 7.

Recall that in a *repeated game* (RG) the players play a given game G repeatedly. We can view a repeated game, with respect to a game G, as consisting of an infinite number of iterations, at each of which the players have to select an action in the game G. After playing each iteration, the players receive the appropriate payoffs, as dictated by that game's matrix, and move to the next iteration. For ease of exposition we normalize both players' payoffs in the game G to be non-negative reals between 0 and some positive constant $R_{\max}$. We denote this interval of possible payoffs by $P = [0, R_{\max}]$. In a *perfect monitoring* setting, the set of possible histories of length t is $(A^2 \times P^2)^t$, and the set of possible histories, H, is the union of the sets of possible histories for all $t \geqslant 0$, where $(A^2 \times P^2)^0$ is the empty history. Namely, the history at time t consists of the history of actions that have been carried out so far, and the corresponding payoffs obtained by the players. Hence, in a perfect monitoring setting, a player can observe the actions selected and the payoffs obtained in the past, but does not know the game matrix to start with. In an *imperfect monitoring* setup, all that a player can observe following the performance of its action is the payoff it obtained and the action selected by the other player. The player cannot observe the other player's payoff. An even more constrained setting is that of *strict imperfect monitoring*, where the player can observe its action and its payoff alone. The definitions of possible histories for an agent in both imperfect monitoring settings follow naturally. Given an RG, a policy for a player is a mapping from H, the set of possible histories, to the set of possible probability distributions over A. Hence, a policy determines the probability of choosing each particular action for each possible history. Notice that a learning algorithm can be viewed as an instance of a policy.

We define the *value* for player 1 of a policy profile (π, ρ), where π is a policy for player 1 and ρ is a policy for player 2, using the *expected average reward criterion* as follows: Given an RG M and a natural number T, we denote the expected T-iterations undiscounted average reward of player 1 when the players follow the policy profile (π, ρ), by $U_1(M, \pi, \rho, T)$. The definition for player 2 is similar. We define

$$U_i(M, \pi, \rho) = \lim_{T \to \infty} \inf U_i(M, \pi, \rho, T) \quad \text{for } i = 1, 2.$$

A policy profile (π, ρ) is a *learning equilibrium* if

$$\forall \pi', \rho', \quad U_1(M, \pi', \rho) \leqslant U_1(M, \pi, \rho), \quad \text{and} \quad U_2(M, \pi, \rho') \leqslant U_2(M, \pi, \rho)$$

for every game matrix M (defined over the set A of actions, and the possible payoffs).

Our first requirement is that learning algorithms will be treated as strategies. In order to be individually rational they should be the best response for one another. In addition, they should rapidly obtain a desired value. The identity of this desired value may be a parameter. We now take a natural candidate, the Nash equilibrium of the game. Another appealing alternative will be discussed later. Assume we consider games with k actions, $A = \{a_1, \ldots, a_k\}$. For every repeated game M, let $n(G) = (N_1(G), N_2(G))$ be a Nash equilibrium of the (one-shot) game G associated with M, and denote by $NV_i(n(G))$ the expected payoff obtained by agent i in that equilibrium. A policy profile (π, ρ) is an *efficient learning equilibrium* (ELE) if for every $\varepsilon > 0$, $0 < \delta < 1$, there exists some $T > 0$, polynomial in $1/\varepsilon$, $1/\delta$, and k, such that for every $t \geqslant T$ and game matrix G (and

its corresponding RG, M), $U_i(M, \pi, \rho, t) \geqslant NV_i(n(G)) - \varepsilon$ for $i = 1, 2$, for some Nash equilibrium $n(G)$, and if player 1 deviates from π to π' in iteration l, then

$$U_1(M, \pi', \rho, l + t) \leqslant U_1(M, \pi, \rho, l + t) + \varepsilon$$

with a probability of failure of at most δ. And similarly, for player 2.

Notice that a deviation is considered irrational if it does not increase the expected payoff by more than ε. This is in the spirit of ε-equilibrium in game theory, and is done in order to cover the case where the expected payoff in a Nash equilibrium equals the probabilistic maximin value.[2] In all other cases, the definition can be replaced by one that requires that a deviation will lead to a decreased value, while obtaining similar results. We have chosen the above in order to remain consistent with the game-theoretic literature on equilibrium in stochastic contexts. Notice also, that for a deviation to be considered irrational, its detrimental effect on the deviating player's average reward should manifest in the near future, not exponentially far in the future.

The above captures the insight of a normative approach to learning in non-cooperative setting. We assume that initially the game is unknown, but the agents will have learning algorithms that will rapidly lead to the values the players would have obtained in a Nash equilibrium had they known the game. Moreover, as mentioned earlier, the learning algorithms themselves should be in equilibrium.

Since learning algorithms are in fact strategies in the corresponding (repeated) game, we in fact require that the learning algorithms will be an ex-post equilibrium in a (repeated) game in informational form [16], and in particular that each strategy will be the best-response against the other agents' strategies regardless of what the payoff matrix is. This point will be discussed in more detail in Section 8.

4. Efficient learning equilibrium: existence

The definition of ELE is of lesser interest if we cannot provide interesting examples of ELE instances. In this section we prove the following constructive result:

Theorem 1. *There exists an ELE for any perfect monitoring setting.*

Below we describe a concrete algorithm with this property.

Our algorithm is based on a combination of the so-called folk theorems in economics and a novel, efficient punishment mechanism which ensures the efficiency of our approach.[3] In the folk theorems (e.g., see [12] and the extended discussion in [13]) the basic idea is that any strategy profile that leads to payoffs that are greater than or equal to the security level (probabilistic maximin) values that the agents can guarantee themselves can be obtained by directing the agents to use the prescribed strategies, and telling each agent to

[2] The probabilistic maximin value for player 1 is defined as $\max_\pi \min_\rho U_i(M, \pi, \rho, t)$ where π and ρ range over the set of policies for players 1 and 2, respectively. The definition for player 2 is similar.

[3] Efficiency here refers to the number of iterations the punishment behavior requires to attain its aim.

punish the other agent if it turns out to deviate from that behavior; the punishment remains a threat that will not be followed in equilibrium and as a result the desired strategy profile will be executed. To use this idea in our setting we need some technique for punishing without (initially) knowing the payoff matrix; moreover we need to devise an efficient punishment procedure for that setting.

Recall that we consider a repeated game M, where at each iteration G is played. In what follows, we often use the term *agent* to denote the player using the algorithm in question, and the term *adversary* to denote the other player. Both players have a set $A = \{a_1, \ldots, a_k\}$ of possible actions.

Consider the following algorithm, termed the ELE algorithm.

The ELE algorithm:

Player 1 performs action a_i one time after the other for k times for $i = 1, 2, \ldots, k$. In parallel to that player 2 performs the sequence of actions $(a_1, \ldots, a_k)$ k times.

If both players behave according to the above, a Nash equilibrium of the (now revealed) game is computed, and the players behave according to the corresponding strategies from that point on. When several Nash equilibria exist, one is selected by using a shared selection algorithm. If one of the players—whom we refer to as *the adversary*—deviates from the above, the other player—whom we refer to as *the agent*, acts as follows: The agent replaces its payoffs in G by the complements to $R_{\max}$ of the adversary payoffs. Hence, the agent will treat the game as a constant-sum game where its aim is to minimize the adversary's payoff. Notice that these payoffs might be unknown. Below we will use G and M to refer to that modified game, and describe how the agent will go about minimizing the adversary's payoff:

Initialize: Construct the following model M' of the repeated game M, where the game G is replaced by a game G' where all the entries in the game matrix are assigned the rewards $(R_{\max}, 0)$. In addition, we associate a boolean valued variable with each joint-action {*assumed*, *known*}. This variable is initialized to the value *assumed*.

Repeat:

Compute and Act: Compute the optimal probabilistic maximin of G' and execute it.

Observe and update: Following each joint action do as follows: Let a be the action the agent performed and let a' be the adversary's action. If (a, a') is performed for the first time, update the reward associated with (a, a') in G', as observed, and mark it *known*.

Claim 1. *The ELE algorithm, when adopted by both players, is indeed an ELE.*

Much of the proof of this theorem, which is non-trivial, rests on showing the agent's ability to punish the adversary quickly. The details are presented in Appendix A.

Notice that an ELE needs not be unique. Indeed, as can be easily seen from our construction, any Nash equilibrium of the one-shot game can be the basis for generating an ELE by employing an appropriate punishment phase. In some applications it is natural to assume a correlation device that will facilitate the selection of a particular ELE. This point will be discussed in Section 8.

5. Imperfect monitoring

The ELE algorithm of the previous section uses the agent's ability to view its adversary's actions and payoffs. A natural question is whether this ability is required for the existence of an ELE. In this section we show that in general, perfect monitoring is required, but there are special classes of games in which an ELE exists with imperfect monitoring.

We start with the general case:

Theorem 2. *An ELE does not always exist in the imperfect monitoring setting.*

Proof. In order to see the above consider the following games:

1. G1:
$$M_1 = \begin{pmatrix} 6, & 0 & 0, & 100 \\ 5, & -100 & 1, & 500 \end{pmatrix}.$$
2. G2:
$$M_1 = \begin{pmatrix} 6, & 9 & 0, & 1 \\ 5, & 11 & 1, & 10 \end{pmatrix}.$$

The payoffs obtained for a joint action in G1 and G2 are identical for player 1 and are different for player 2. The only equilibrium of G1 is where both players play the second action, leading to $(1, 500)$. The only equilibrium of G2 is where both players play the first action, leading to $(6, 9)$ (these are unique equilibria since they are obtained by removal of strictly dominated strategies).

Now, assume that an ELE exists, and look at the corresponding policies of the players in that equilibrium. Notice that in order to have an ELE, we must visit the entry $(6, 9)$ most of the times if the game is G2 and visit the entry $(1, 500)$ most of the times if the game is G1; otherwise, player 1 (respectively player 2) will not obtain high enough value in G2 (respectively G1), since its other payoffs in G2 (respectively G1) are lower than that.

Given the above, it is rational for player 2 to deviate and pretend that the game is always G1 and behave according to what the suggested equilibrium policy tells it to do in that case. Since the game might be actually G1, and player 1 can not tell the difference, player 2 will be able to lead to playing the second action by both players for most times also when the game is G2, increasing its payoff from 9 to 10, contradicting ELE. □

But while our approach is not Bayesian, it does not exclude the possibility that the agent knows that it is participating in a game from a particular class. Thus, there may be classes of repeated games for which an ELE exists. In particular, consider the class of *repeated common-interest games.* These are repeated games M where the underlying game G is a common-interest game, i.e., a game in which both players always receive identical payoffs. In this setting our definition of imperfect and perfect monitoring denote the same setting—if the player knows its payoff, it knows its adversary's payoff as well. Thus, we examine the case of *strict imperfect monitoring*. Recall that in this setting, the player knows only its action and its payoff.

Theorem 3. *There exists ELE for the class of common-interest games under strict imperfect monitoring.*

Proof. The idea is quite simple and, surprisingly, has not been proposed before, while other, more complex and less efficient approaches have been proposed. It does require knowledge of the number of actions available to each agent (or a polynomial bound on them). The algorithm works as follows: First, the agents go through a series of random play. They do so sufficiently many times to ensure that the probability that all joint-actions have been played is greater than $1-\delta$. During this phase, each agent maintains information about the best payoff obtained so far, and the action it used when this payoff was first obtained. Once the exploration phase is over with, the agent plays this best action repeatedly.

This is a learning equilibria because the average reward this learning strategy leads to is the maximal average reward for every agent. Thus, no agent has any motivation to deviate from it. It is an ELE because a polynomial number of steps is required to attain this average reward (see below) and any deviation will immediately reduce the average reward of the agent. We need only a polynomial number of steps to approximately obtain the maximal average reward because we need only $O(k^4 \cdot \log(k^2/\delta))$ steps of random play to ensure that all joint-actions have been played with a probability of at least $1-\delta$. This follows from the following. Given large enough k, we get that the probability that after $k^2 m$ trials the agents will not play some previously unplayed joint action can be approximated by e^{-m}. Hence, we get that after $O(k^2 \log(k^2/\delta))$ the probability we will not learn the outcome associated with a new joint action can be approximated by δ/k^2. By repeating the process k^2 times we get the desired result. □

6. Pareto-ELE

The previous sections dealt with ELE in the perfect and imperfect monitoring settings. In both cases we were interested in having a learning procedure that will enable the agents to obtain expected payoffs as the ones they would have obtained in a Nash equilibrium, had they known the game. A more ambitious objective is the following. Let $p_i(a,b)$ denote the payoff for player i in the game in question when player 1 plays a and player 2 plays b. We say that a pair of actions (a,b) is (economically) *efficient*, if $p_1(a,b) + p_2(a,b) = \max_{s\in A, t\in A} p_1(s,t) + p_2(s,t)$. That is, if the total payoff for both agents is maximized.

It is easy to see that if all the agents can do is to choose an action in G, then there is no general way to guarantee that agents will behave in an economically efficient manner. This is due to the fact that it may be the case that although (a,b) is the only economically efficient behavior, performing a (respectively b) by agent 1 (respectively 2) is irrational: $p_1(a,b)$ (respectively $p_2(a,b)$) may be lower than the probabilistic maximin value that agent 1 (respectively 2) can guarantee itself.

The classical approach in economics for dealing with economic (in)efficiency is by introducing side (monetary) payments. Formally, part of the strategy of agent i is a function $A^2 \times P^2 \to R_+$, i.e., agent i is instructed to pay a certain amount of money to the other

agent as part of its strategy.[4] If an agent's reward is p and he is paid c (where c can be positive, negative, or zero) then his utility is assumed to be $u = p + c$ (this type of utility function is termed quasi linear). The sum, over all agents, of the monetary payments is always 0, and, as a result, if the agents turn out to be using strategies that maximize $u_1 + u_2$ then they will also be economically efficient.

Now we can define the notion of a *Pareto-ELE*. A Pareto-ELE is similar to a Nash-ELE, but its aim is that the agents' behavior will be economically efficient. Therefore, the two distinctive aspects of Pareto-ELE are:

1. We require that the agents will be able to get close to an efficient outcome.
2. We allow side payments as part of the agents' behavior.

Suppose that we are considering games with k actions. For every repeated game M, let $(P_1(G), P_2(G))$ be an economically efficient joint action of the (one-shot) game associated with M, and denote by $PV_i(M)$ the payoff obtained by agent i in that joint action. A policy profile (π, ρ), which also allows side payments, is a *Pareto efficient learning equilibrium* if for every $\varepsilon > 0, 0 < \delta < 1$, we have that there exists a $T > 0$, polynomial in $1/\varepsilon$, $1/\delta$, and k, such that for every $t \geqslant T$ and game matrix G (defined over the actions in A), with corresponding RG, M, $U_1(M, \pi, \rho, t) + U_2(M, \pi, \rho, t) \geqslant PV_1(G) + PV_2(G) - \varepsilon$ for $i = 1, 2$, and if player 1 deviates from π to π' in iteration l, then $U_1(M, \pi', \rho, l + t) \leqslant U_1(M, \pi, \rho, l + t) + \varepsilon$ with a probability of failure of at most δ. And similarly for player 2.

Theorem 4. *There exists a Pareto-ELE for any perfect monitoring setting.*

Proof (*Sketch*). Consider the following algorithm that defines the policies and side payments for the agents.

Player 1 performs k iterations of the action a_i, for $i = 1, 2, \ldots, k$. In parallel to that, player 2 performs the sequence of actions $(a_1, \ldots, a_k)$ k times.

Now, that the game is known to both agents, they compute the probabilistic maximin values for agent 1 and agent 2. Denote the probabilistic maximin value of agent i by v_i and the payoff it gets in the economically efficient solution by e_i. Without loss of generality, $e_1 - v_1 > e_2 - v_2$. Choose r such that $r = e_1 + e_2 - (v_1 + v_2)$. If player 2 is paid $r/2$ by player 1 when the efficient solution is played then each player's total payoff is at least as high as his probabilistic maximin. This is easy to see by examining the two cases: $e_2 - v_2 > 0$ and $e_2 - v_2 \leqslant 0$.

From now on the agents adopt the efficient behavior with the above side-payments in all following iterations. If several economically efficient behaviors exist, some pre-determined selection algorithm is used.

In case one of players (the adversary) deviates, either in the exploration stage or the following state, the other player (the agent) will punish it as in the case of Nash-ELE.

[4] This is the definition in the perfect monitoring case. The definition in the imperfect monitoring case is similar.

It will play as if its payoffs in the game are the complements to $R_{\max}$ of the adversary payoffs.

The proof now follows the steps of the proof for the existence of ELE. □

For the case of imperfect monitoring, the same result with respect to Nash-ELE hold here.

Theorem 5. *A Pareto-ELE does not always exist in an imperfect monitoring setting.*

7. Stochastic games

Stochastic games provide a more general model of repeated multi-agent interactions. In a stochastic game the players may be in one of finitely many states, $S = \{s_1, \ldots, s_m\}$, where each state is associated with a game in strategic form. The joint action at each state not only determines the payoffs but also determine (stochastically) the identity of the next state the agents will reach. Formally, let $A = \{a_1, \ldots, a_k\}$ be the set of actions available to the agents. For each state s_i, the game associated with s_i associates a payoff $p^i_j(a, b)$ with agent j when the joint action is (a, b). In addition, for every $s_i \in S$ the probability that the next state will be s_j when the joint action is (a, b) is denoted by $P(s_i, s_j, a, b)$.

Now that we have multiple games, a policy π_i for agent i associates a (possibly mixed) action with every state and, potentially, a payment to the other agent. This policy is a function of the history of states the agent visited and the payoffs it observed. Throughout this section we assume the perfect monitoring setting, since our impossibility result for imperfect monitoring in repeated games immediately rules out the existence of an ELE in the more general context of stochastic games.

Stochastic games provide a more realistic setting, that is also more challenging technically. First, let us try and understand the issues involved. The first obstacle we face is the lack of general results on the existence of Nash equilibrium in average-reward stochastic games. Thus, we restrict our attention to the case of Pareto-ELE.

Conceptually, the required generalization is straightforward—the learning algorithm should quickly lead to an economically efficient policy for both agents, i.e., a policy that maximizes the average sum of rewards, and deviations should quickly lead to a lower reward. However, while in the case of repeated games we equated "quick" with polynomial in the size of the game and the approximation parameters ε and δ, the situation in stochastic games is more complicated. A parameter that is typically used to assess the speed of convergence of a learning algorithm in stochastic games is the ε-return mixing time [4,19]. Intuitively, the ε-return mixing time of a policy is the expected time it would take an agent that uses this policy to converge to a value that is ε close to the value of the policy. Ideally, we would like a learning algorithm to attain the optimal value in time polynomial in the ε-return mixing time of the optimal policy.

Formally, assume some fixed stochastic game M, and let (π_1, π_2) be a policy profile in M. We denote the T-step average reward of this policy profile for agent i starting at state s_0 by $U_i(s_0, \pi_1, \pi_2, T)$, and we define $U_i(s_0, \pi_1, \pi_2) = \liminf_{T\to\infty} U_i(s_0, \pi_1, \pi_2, T)$. We

let U denote $U_1 + U_2$. The *ε-return mixing time* of (π_1, π_2) is the minimal T such that for all $t > T$ and all states s, we have that $U(s, \pi_1, \pi_2, t) > U(s, \pi_1, \pi_2) - \varepsilon$. Thus, after the policy profile (π_1, π_2) is executed for T steps or longer, the agents' expected average sum of rewards will be very close to their long-term average sum of rewards. Let (π, ρ) be a policy profile that maximizes $\min_s U(s, \cdot, \cdot)$, and let T_{mix} be its ε-return mixing time.

The definition of Pareto efficient learning equilibrium in stochastic games is identical to that of repeated games, except that T must be polynomial in T_{mix} as well. Note that if the game is irreducible (i.e., for any fixed policy profile, the induced Markov chain is ergodic), $U(s, \pi, \rho)$ does not depend on s. We can show the following:

Theorem 6. *Under the following assumptions a Pareto-ELE in stochastic games exists*: (1) *The agents have perfect monitoring*; (2) T_{mix} *is known.*

Proof. The intuitive idea behind the algorithm is identical to the case of repeated games and so we elaborate only on the new issues. First, the agents run an algorithm for finding a policy profile π, ρ that maximize $U(\cdot, \cdot)$. Next, they run an algorithm for finding the best that each can accomplish on its own (i.e., assuming the other agent is trying to minimize their average payoff). From that point on they run the policy profile π, ρ, adjusted with appropriate side payments so that each agent receives more than the best it can accomplish on its own, much like in the case of repeated games. At any point, if an agent deviates, the other agent plays as if its goal is to minimize the other agent's average reward.

The learning algorithm we just described is (ε-)Pareto-optimal: The long-term average sum of rewards of this algorithm is ε-close to the optimal average sum of rewards, as desired. No agent has an incentive to deviate at any stage because the side-payment structure guarantees that it attain at least the value it could attain on its own.

To show that this algorithm is a Pareto-ELE, we also need to show that the value can be attained efficiently and punishment can be performed efficiently. We do so by resorting to recent results on efficient learning in fixed-sum stochastic games and common-interest stochastic games. First, to compute the policies π, ρ that maximizes $U(\cdot, \cdot)$ we use the algorithm described in [5]. We refer the reader there for more details. What we need to note here is that this algorithm learns the required policy profile in polynomial time. Next, to compute the values each agent can attain on its own we use R-max [4]. R-max is appropriate here because we are learning in a fixed-sum game. We this first for the fixed sum game in which the rewards are based on agent 1's rewards, and then with respect to the fixed-sum game in which the rewards are based on agent 2's rewards.

We note that given some value T' as input, R-max will learn a T'-step policy in time polynomial in T' and the other game parameters. This policy will be optimal among all policies that mix in time T'. We shall take $T' = T_{\text{mix}}$. The average reward of this policy will be used to compute the side payments structure as in the case of repeated games. In any case, the average reward of the policy profile π, ρ (suitably modified to include the side payments) will be no lower than the value that each agent can receive on its own. Thus, should an agent deviate from the above, we know that within T_{mix} steps it will attain a lower average reward, i.e., punishment can be carried out efficiently. □

Finally, note that there is a standard, though imperfect, technique for removing knowledge of T_{mix} in which we simply guess progressively higher values for T_{mix}. We refer the reader to [4] for the implications of this approach.

8. Discussion

Most previous work on learning in games fits into one of the following two paradigms:

(1) The study of learning rules that will lead to a Nash equilibrium (or other solution concept) of a game [11].
(2) The study of learning rules that will predict human behavior in non-cooperative interactions, such as the ones modeled in repeated games [7].

While the approach taken in (2) has significant merit for descriptive purposes, a normative approach to learning should go beyond recommending behavior that will eventually lead to some desired solution. The major issues one needs to face are:

(1) The learning algorithms of the agents should be individually rational.
(2) The learning algorithms should efficiently converge to the desired values if employed by the agents.
(3) A deviation from the desired learning algorithm should become irrational after a short period of time.

The concepts introduced in this paper address these issues. Both ELE and Pareto-ELE provide new basic tools for learning in non-cooperative settings. Moreover, we have been able to show constructive existence results for both ELE and Pareto-ELE in the context of repeated games with perfect monitoring. We were also able to show that if we relax the perfect monitoring assumption, the desired properties are impossible to obtain in the general case. Pareto-ELE is an appealing concept in the context of stochastic games as well, and we were able to extend our results to that context. Together, our concepts and results provide a rigorous normative approach to learning in general non-cooperative interactions.

8.1. Related work

It is useful to contrast our approach with an important line of related work that features algorithms that guarantee for the agent using them a value which is approximately equal to the value he would have attained had he known in advance how his adversary would play. Algorithms along this line appear in, e.g., [15] and in [10] (where special attention is given to the issue of efficiency). This latter result is truly in the spirit of online algorithms, where our goal is to do as much as we can online as we would have been able to do off-line. In this case, we attempt to react online to an adversary's behavior in a manner that would be similar—in terms of our average payoff—to the best we could have done had we known the adversary's behavior before hand. These results are highly valuable, but many readers may not notice a subtle, but crucial point about them: They treat the

adversary's policy as a fixed sequence of mixed strategies (probabilistic actions), and that is contrary to the spirit of game-theory. In reality, the adversary can adjust its policy in response to the agent's behavior. Imagine, for example, the following instance of the well-known Prisoner's Dilemma game:

$$M_1 = \begin{pmatrix} 2, & 2 & -10, & 10 \\ 10, & -10 & -5, & -5 \end{pmatrix}.$$

Consider the following two adversary policies: (1) If the agent initially plays row 1 (denoted *cooperate*) the adversary will always play column 1 (denoted *cooperate*, too). If the agent initially plays row 2 (denoted *defect*) the adversary will always play column 2 (denoted *defect*, too). (2) If the agent initially plays *cooperate* the adversary will always play *defect*. If the agent initially plays *defect* the adversary will always play *cooperate*. It is clear that no agent can guarantee a best-response value against such an adversary, and this approach is limited to a view of the adversary as using a (pre-determined) sequence of mixed strategies.[5] The bottom line is that, despite their practical and theoretical importance, these results, in essence, take more of a single-agent decision problem perspective, and they cannot replace concepts that are based on the notion of an equilibrium.

Another related work on normative guidelines to the design of learning algorithms is [2]. There, Bowling and Veloso suggest two criteria for learning algorithms. The first, which they call *rationality* stipulates that if the other player's policies converge to a stationary policy then the learning algorithm will converge to the best-response policy. The second, which they call *convergence* stipulates that the learner will necessarily converge to a stationary policy. Both criteria are attractive, but as with the above work, the notion of a Nash equilibrium of learning strategies is a deeper notion of rationality than that of best-response upon convergence. And convergence, though definitely desirable, ignores the issue of convergence rate. Moreover, convergence, as specified, is not well defined. Indeed, in their work, Bowling and Veloso consider the special (well-defined) case of convergence in self-play, i.e., when all agents use the same algorithm. This is the standard notion of convergence adopted by most work on learning in games uses. In fact, in the particular context of self-play investigated by Bowling and Veloso, their requirements are equivalent to the requirement that the algorithm will converge to a correlated equilibrium—a common property pursued by learning algorithms in the game-theory literature. The concept of ELE provides a more rigorous notion of *individually* rational learning strategies. Moreover, we believe that efficient (i.e., polynomial time) convergence rate should be an integral part of the definition of rationality. In many settings, what happens after an exponential number of iterations is not of great interest. This applies to the judgment of irrationality as well. An agent that makes an "irrational" choice that leads to increased reward in the near future and decreased reward only after an exponential number of steps does not seem too irrational.

[5] A strategy of always playing *defect* would be the best response for every *particular* sequence of actions taken by an adversary from one of the two classes of adversaries described above. But such an approach would completely ignore the fact that the agent's first action influences the whole future of adversary steps. It is easy to construct more complicated examples of how the agent's actions influence the adversary's choices.

When considering previous work on learning algorithms more generally, a natural question that arises is whether they are an ELE. This is not likely to be easy to answer—so far there are no polynomial-time convergence results for existing algorithm other than R-max [4]. But even if it is not possible to prove that they converge in polynomial time, it is interesting to check whether a learning agent can improve his average payoff if he knows that his adversary is using a particular learning algorithm. Such stability results of existing algorithms are highly desirable.

8.2. Perfect monitoring

Our results indicate that ELE is possible when agents have perfect monitoring, while there are various classes of games with imperfect monitoring in which ELE cannot be achieved. This limits the applicability of the ELE criteria as it cannot be used as a guideline in such cases. Yet, most previous work in AI on learning algorithms for games assumes perfect monitoring, including [14,17,21]. Work on common-interest or zero-sum games makes this assumption implicitly too, since there the perfect and imperfect monitoring settings coincide.

Perfect monitoring is not assumed by authors that attempt to devise competitive strategies, such as [10]. The differences between our setting and their setting was discussed above. Much of the earlier work on learning in games is mostly not algorithmic, and does not attempt to devise efficient algorithms. This includes work on replicator dynamics [11], which assumes the existence of (external) dynamics where agents who have obtained higher rewards "survive". In fact, one way to interpret this is as a perfect monitoring assumption whereby agents tend to mimic agents in, e.g., a symmetric game, who received higher payoffs in the past using particular actions. Notice also that in some of the game theory literature the agent is assumed to know his utility function to start with and the learning process is used to explain how actions are selected.

Despite the frequent use of the perfect monitoring assumption in the literature, the cautious reader may worry about the applicability of this assumption in practice. To see the wide applicability of that setting, consider the following scenario. Consider a set of possible games, each of which refers to different assignment to some initially unknown feature of the environment (e.g., how efficient is a particular resource). When agents select their actions while playing the game, this initially unknown feature is revealed (e.g., when agents choose their actions they learn about the speed of the appropriate resource as a result of the payoffs obtained by the different agents). Such settings fit the perfect monitoring assumption and they are typical in various economic settings. For example, work in economics refers to the setting of private values, where each agent has its own worth for a good, observed only by it, and the setting of common values, where the worth of the good is common but is revealed to the agents only after they take their actions. Many interesting intermediate cases exists (see [20] for a discussion and survey in the context of auction theory). These settings fall into the category of games in which private payoffs are revealed after the joint-action of all agents and are instances of games with perfect monitoring.

Finally, we note that the fact that ELE does not exist for certain (large) classes of games with imperfect monitoring, does not imply that an ELE does not exist for more restrictive classes, as we demonstrated in the case of common-interest games.

8.3. Multi-player games

Our formulation of ELE clearly extends to multi-player games. However, it is not apparent whether the positive results extend as well. First, observe that the technique of obtaining the probabilistic maximin value when the game is initially unknown, introduced in [4], naturally extends to the case of many players. In this paper we use this idea in order to punish a deviator. As a consequence, our result immediately applies if there is some form of communication among the players, i.e., the agents can act as one entity whose aim is to punish the deviator. In the case of perfect monitoring, deviations will be detected simultaneously by the agents and therefore they can move to the corresponding punishment mode, where all of the agents (excluding the deviator) will treat themselves as one agent with payoffs complementary to those of the deviators. Without a communication mechanism, one can obtain similar results if we consider deterministic punishments only. This restricts the class of games where we can prove that ELE exists to ones where deterministic punishment makes the payoff lower than the one obtained in the Nash equilibrium. Otherwise, the techniques used here are not applicable, and alternative approaches should be sought.

The main drawback of applying our approach in the multi-player setting is that although computations remain polynomial in the size of the explicit representation of the game, this size grows exponentially with the number of players. The discussion of ELE in the framework of succinct representation of games is beyond the scope of this paper.

8.4. ELE and ex-post equilibrium in games with incomplete information

Some of the insights behind ELE relate to work in economic mechanism design. The recent CS literature deals extensively with connections between distributed systems and the design of economic mechanisms. One central issue in that literature is the selection of a solution criterion: what can we assume about the agents' rational behavior? Of particular interest is the case where agents can communicate with one another. One must carefully consider the particular assumptions made in this setting with respect to whether or not the agents adopt strategic considerations when taking actions and passing messages (e.g., can agents strategically modify messages sent by others?; see [8,23,27] for various ways of approaching these issues). Much of the literature deals with the search for mechanisms where the agents will have dominant strategies that lead to desired behavior (e.g., maximizing social efficiency, or elicitation of agents' preferences). However, it can be shown that such mechanisms *rarely* exist. One alternative is to consider games/mechanisms where there exists an ex-post equilibrium: a strategy profile of the agents in which it is irrational to deviate from each agent's strategy, assuming the other agents stick to their strategies, and *regardless* of the state of the system. This system state may not be initially observable and might consist of various private inputs of the agents. Ex-post equilibria play a major role in central mechanisms (see [16] for general results in the context of the

famous VCG mechanisms). It turns out that the idea of equilibrium of learning algorithms can be viewed similarly. We search for strategies (termed learning algorithms) such that it will be irrational for each agent to deviate from its strategy assuming the other agent sticks to its strategy, and regardless of the state of the system (which in our setting is the initially unknown game). We believe that our results contribute to the literature on ex-post equilibria in games with incomplete information, by providing a general generic setting where they play a major role and do exist.

8.5. Correlation devices

Although our definitions are all in the spirit of equilibrium theory, in some cases (and in particular in the case of side payments in Pareto-ELE) it will be nice to motivate the source of algorithms (e.g., the designed payments scheme). One possible approach is to consider a correlation/mediator who provides the agents with algorithms to use and suggested payments to be made. This correlation device is not a designer who can enforce behaviors or payments, and it does not possess any private knowledge or aim to optimize private payoffs. Therefore, the right way to view this party is as a mediator/interested party or correlation device (as in [1,9], etc.) We find this interpretation to be convenient, although not essential. Notice that the suggested payments are just part of the algorithms, and it is up to the agents to decide whether to make them; it is the proof that the algorithms are in equilibrium that suggests that these payments will be actually executed by individually rational agents.

8.6. Cooperation and threats

One of the major issues in non zero-sum repeated games is cooperation and threat. This is manifested by the famous prisoners dilemma and similar situations. These issues are naturally handled in our framework under the umbrella of Pareto-ELE. In situations like the prisoners dilemma, one may be interested in the emergence of economically efficient outcome as an equilibrium of rational strategies. Indeed, the folk theorem in economics employed by our work is the formal justification for such behavior. For example, procedures such as Tit-For-Tat or GRIM are strategies that are in equilibrium in the repeated prisoners dilemma and yield cooperation in that setting. In a sense, Tit-For-Tat has an inherent mechanism for punishing the other player if he does not cooperate. Our results in the context of Pareto-ELE will therefore lead to, e.g., cooperation in the repeated prisoners' dilemma. Notice that in the framework of a classical prisoners dilemma, presented as a symmetric game, no side payments will be needed (since the benefits from cooperation will be identical for both agents). More generally we get that issues regarding threats and cooperation are inherited from the folk theorems, and therefore are an integral part of our approach.

Our approach, however, deals with a more general case in which the game played might be initially unknown, e.g., the game played might be the prisoners dilemma but might be also be another game. In order to therefore have credible threats one should add an efficient algorithm that "punishes" effectively and efficiently also when the game is not known. Such a procedure is provided in this paper. As a result we are able not just to

tolerate cooperation and threats, but to do so for a class of games where the identity of the game is initially unknown, as captured by the Pareto-ELE setting.

Appendix A. ELE existence proof

Claim 1. *The ELE algorithm, when adopted by both players, is indeed an ELE.*

Proof. We refer to the algorithm used by the agent when playing against the adversary in order to guarantee that the adversary will not obtain more than its maximin value as the *repeated-R-max algorithm*, as it is a version of the R-max algorithm [4]. The proof that the repeated-R-max leads to near-optimal value takes the following steps. The game matrix for G contains k^2 entries. Therefore, the number of iterations in which the agent learns some new information (i.e., has to update its model) is at most k^2. This means that the agent needs to compute its strategy at most k^2+1 times. Each computation of the probabilistic maximin strategy requires polynomial time and can be carried out using linear programming. As we shall show, except for this small number of phases in which the agent learns new information, its expected payoff is at least as high as the value of the probabilistic maximin strategy. Thus, overall, its average payoff will be high.

The actual number of iterations in which the agent learns new information may be much smaller than k^2, and it depends on how the adversary plays. The adversary can attempt to prevent the agent from learning about various aspects of the game. However, whenever it does this, the agent is guaranteed to obtain a good payoff. More precisely, in any iteration in which the agent did not learn new information, its expected payoff must be no less than the payoff of the probabilistic maximin strategy according to its model, which is at least as high as the probabilistic maximin value of the real game.

Of course, the discussion above deals with expected values. If we wish to guarantee an actual value with probability of at least $1-\delta$, we proceed as follows:

Lemma A.1. *Assume that the expected payoff of strategy s is μ. Then, the probability that our average payoff along z iterations will be less than $\mu - R_{\max}/z^{1/3}$ is bounded by* $\mathrm{e}^{-z^{1/3}/2}$.

Proof. Let X_i be the payoff in iteration i, and let $Y_i = (\mu - X_i)/R_{\max}$. Notice that $|Y_i| \leqslant 1$, and that $E(Y_i) = 0$. Hence, Chernoff bound implies that $\mathrm{Prob}(\sum_{j=1}^{z} Y_j > z^{2/3}) < \mathrm{e}^{-z^{1/3}/2}$. This implies that the average return along z iterations is at most $R_{\max}/z^{1/3}$ lower than μ with probability of at least $1-\mathrm{e}^{-z^{1/3}/2}$. □

In what follows, we choose z so that: $R_{\max}/z^{1/3} < \varepsilon/(2R_{\max}k^2)$ and $k^2\mathrm{e}^{-z^{1/3}/2} < \theta$. This holds for $z = \max(8R_{\max}^6 k^6/\varepsilon^3, 8\ln^3(k^2/\theta)) + 1$. Note that z is polynomial in k, $1/\varepsilon$, $1/\theta$, and $R_{\max}$. The precise value of θ will be determined later, and it will be polynomial in k, $1/\varepsilon$, $1/\delta$ and $R_{\max}$, as required. Assume that the adversary uses some fixed, arbitrary strategy σ, and let us use m to denote the value of the probabilistic maximin strategy for the agent. Our next step is to show the following:

Lemma A.2. *Let G' be the agent's current model of the game and let G be the real game. Let ρ be the probabilistic maximin policy with respect to G'. If the agent plays ρ for z iterations then with probability of at least $1-\theta$ the agent will either receive a payoff of at least $m - \varepsilon/R_{\max}$, or it will learn the value of a new entry.*

Proof. Let r denote the agent's actual average payoff in z iterations of ρ and σ. If $r \geqslant m - \varepsilon/R_{\max}$, we are done. Therefore, let us assume that $r < m - \varepsilon/R_{\max}$. We shall show that with probability of at least $1-\theta$ the agent will learn a new entry in the game matrix. Let $m_{G'}$ denote the probabilistic maximin value of the game G'. By definition, $m_{G'}$ is also a lower bound on the expected payoff of ρ against σ in G'. Let $\bar{m}$ be the expected value of ρ with respect to G and σ. We know that $|r - \bar{m}| \leqslant \varepsilon/(2R_{\max})$ with probability of at least $1-\theta$. Since we assume $r < m - \varepsilon/R_{\max}$, it follows that $\bar{m} < m - \varepsilon/(2R_{\max})$. However, in G' the payoffs are at least as high as the corresponding values in G. This means that $m_{G'} > m$. Define a new game $W = G' - G$, i.e., the payoff in W is the difference between the corresponding payoffs in G and G'. By definition, W is non-negative. It is strictly positive exactly in the entries that are marked unknown. The expected value of W given ρ and σ is exactly the difference between the expected value of ρ and σ on G' and on G. We know that the expected value of ρ and σ in G' is at least $m_{G'} > m$. The expected value of ρ and σ in G is $\bar{m}$. Therefore, $E(W) > \varepsilon/(2R_{\max})$. Using standard Chernoff bound analysis (e.g., as in Lemma A.1, above), it follows that after z iteration, it follows that the actual average value of W is positive with probability of at least $1-\theta$. This can happen only if an unknown entry is played. □

To finish the proof, showing that the agent can punish efficiently a deviator, we choose $Z = zk^2$ and let $T = Z + Q$, where $T\varepsilon/2 \geqslant ZR_{\max} + Q\varepsilon/(2R_{\max})$. This is satisfied if $Q > Z2R_{\max}/\varepsilon$. Since Z is polynomial in k, $1/\varepsilon$, $1/\theta$, and $R_{\max}$, we have that Q and T are polynomial in k, $1/\varepsilon$, $1/\theta$, and $R_{\max}$. Finally, we define $\theta = \delta/T$.

We claim that with probability of at least $1-\delta$, for any $t > T$, the average payoff obtained by the agent cannot be lower than the probabilistic maximin value of the game by more than ε. We know that with probability of at least $1 - T\theta = 1-\delta$, in at least Q iterations, the average payoff will be at least $m - \varepsilon/(2R_{\max})$ and in at most Z iteration the average value of the game W will be positive (i.e., learning will occur). We know that the average payoff in these Z remaining iterations is between 0 and $R_{\max}$. Therefore, by definition of T, the overall average value will be as desired. Finally, we note that the agent need not be aware of any of z, Z, $R_{\max}$ or Q, and it simply plays according to the algorithm above.

To conclude the existence proof, we notice that there are only k^2 iterations where the agents explore the unknown game matrix. If a player deviates from the exploration stage, then Lemma A.2 guarantees that this deviation will be irrational (up to an ε factor, with the desired probability of success). If the agents do follow the exploration stage then deviation from selecting the appropriate strategy (determined by a Nash equilibrium) is irrational by the definition of the Nash equilibrium. □

References

[1] R.J. Aumann, Subjectivity and correlation in randomized strategies, J. Math. Econom. 1 (1974) 67–96.
[2] M. Bowling, M. Veloso, Rational and covergent learning in stochastic games, in: Proc. IJCAI-01, Seattle, WA, 2001, pp. 1021–1026.
[3] R.I. Brafman, M. Tennenholtz, R-max—a general polynomial time algorithm for near-optimal reinforcement learning, in: Proc. IJCAI-01, Seattle, WA, 2001.
[4] R.I. Brafman, M. Tennenholtz, R-max—a general polynomial time algorithm for near-optimal reinforcement learning, J. Machine Learning Res. 3 (2002) 213–231.
[5] R.I. Brafman, M. Tennenholtz, Learning to coordinate efficiently—a model based approach, J. Artificial Intelligence Res. 19 (2003) 11–23.
[6] C. Claus, C. Boutilier, The dynamics of reinforcement learning in cooperative multi-agent systems, in: Proc. AAAI-98, Madison, WI, 1998, pp. 746–752.
[7] I. Erev, A.E. Roth, Predicting how people play games: reinforcement learning in games with unique strategy equilibrium, Amer. Economic Rev. 88 (1998) 848–881.
[8] J. Feigenbaum, S. Shenker, Distributed algorithmic mechanism design: recent results and future directions, in: Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications (DIAL-M 2002), Atlanta, GA, 2002, pp. 1–13.
[9] F. Forges, An approach to communication equilibrium, Econometrica 54 (6) (1986) 1375–1385.
[10] Y. Freund, R.E. Schapire, Adaptive game playing using multiplicative weights, Games and Economic Behavior 29 (1999) 79–103.
[11] D. Fudenberg, D. Levine, The Theory of Learning in Games, MIT Press, Cambridge, MA, 1998.
[12] D. Fudenberg, E. Maskin, The folk theorem in repeated games with discounting or with incomplete information, Econometrica 52 (1986) 533–554.
[13] D. Fudenberg, J. Tirole, Game Theory, MIT Press, Cambridge, MA, 1991.
[14] A. Greenwald, K. Hall, R. Serrano, Correlated q-learning, in: NIPS Workshop on Multi-Agent Learning, Vancouver, BC, 2002.
[15] S. Hart, A. Mas-Colell, A reinforcement procedure leading to correlated equilibrium, in: G. Debreu, W. Neuefeind, W. Trockel (Eds.), Economic Essays: A Festschrift for Werner Hildenbrand, Springer, Berlin, 2001, pp. 181–200.
[16] R. Holzman, N. Kfir-Dahav, D. Monderer, M. Tennenholtz, Bundling equilibrium in combinatorial auctions, Games and Economic Behavior 47 (2004) 104–123.
[17] J. Hu, M.P. Wellman, Multi-agent reinforcement learning: theoretical framework and an algorithms, in: Proc. 15th ICML, Madison, WI, 1998, pp. 242–250.
[18] L.P. Kaelbling, M.L. Littman, A.W. Moore, Reinforcement learning: a survey, J. Artificial Intelligence Res. 4 (1996) 237–285.
[19] M. Kearns, S. Singh, Near-optimal reinforcement learning in polynomial time, in: Proc. 15th ICML, Madison, WI, 1998, pp. 260–268.
[20] P. Klemperer, Auction theory: a guide to the literature, J. Economic Surv. 13 (3) (1999) 227–286.
[21] M. Littman, Friend or foe q-learning in general sum Markov games, in: Proc. 18th ICML, Williamstown, MA, 2001, pp. 322–328.
[22] M.L. Littman, Markov games as a framework for multi-agent reinforcement learning, in: Proc. 11th ICML, New Brunswick, NJ, 1994, pp. 157–163.
[23] D. Monderer, M. Tennenholtz, Distributed games, Games and Economic Behavior 27 (1999) 55–72.
[24] J.F. Nash, Equilibrium points in n-person games, Proc. Nat. Acad. Sci. USA 36 (1950) 48–49.
[25] G. Owen, Game Theory, second ed., Academic Press, New York, 1982.
[26] L.S. Shapley, Stochastic games, Proc. Nat. Acad. Sci. USA 39 (1953) 1095–1100.
[27] J. Shneidman, D. Parkes, Using redundancy to improve robustness of distributed mechanism implementations, in: Proc. 4th ACM Conference on Electronic Commerce (EC'03) (short version), 2003, pp. 276–277.