Optimal Efficient Learning Equilibrium: Imperfect Monitoring

Ronen I. Brafman
Computer Science Department
Stanford University
Stanford, CA 94305
brafman@cs.bgu.ac.il

Moshe Tennenholtz
Industrial Eng. & Management
Technion
Haifa, Israel 32000
moshet@ie.technion.ac.il

Abstract

Efficient Learning Equilibrium (ELE) is a natural solution concept for multi-agent encounters with incomplete information. It requires the learning algorithms themselves to be in equilibrium for any game selected from a set of (initially unknown) games. In an optimal ELE, the learning algorithms would efficiently obtain the surplus the agents would obtain in an optimal Nash equilibrium of the initially unknown game which is played. The crucial part is that in an ELE deviations from the learning algorithms would become non-beneficial after polynomial time, although the game played is initially unknown. While appealing conceptually, the main challenge for establishing learning algorithms based on this concept is to isolate general classes of games where an ELE exists. Unfortunately, it has been shown that while an ELE exists for the setting in which each agent can observe all other agents' actions and payoffs, an ELE does not exist in general when the other agents' payoffs cannot be observed. In this paper we provide the first positive results on this problem, constructively proving the existence of an optimal ELE for the class of symmetric games where an agent can not observe other agents' payoffs.

1. Introduction

Reinforcement learning in the context of multi-agent interactions has attracted the attention of researchers in cognitive psychology, experimental economics, machine learning, artificial intelligence, and related fields for quite some time (Kaelbling, Littman, & Moore 1996; Erev & Roth 1998; Fudenberg & Levine 1998). Much of this work uses repeated games (e.g. (Claus & Boutilier 1997; Kalai & Lehrer 1993; Conitzer & Sandholm 2003)) and stochastic games (e.g. (Littman 1994; Hu & Wellman 1998; Brafman & Tennenholtz 2002; Bowling & Veloso 2001; Greenwald, Hall, & Serrano 2002)) as models of such interactions. The literature on learning in games in game theory (Fudenberg & Levine 1998) 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 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. We find this perspective highly controversial. Indeed, the basic idea in game theory is that agents would adopt only strategies which are individually rational. This is the reason why the notion of equilibrium has been introduced and became the dominant notion in game theory and economics. It is only natural that similar requirements will be required from the learning algorithms.

In order to address the above issue, Brafman and Tennenholtz (Brafman & Tennenholtz 2004) introduced the notion of *Efficient Learning Equilibrium* [ELE]. In this paper we deal with an improved version of ELE, where the agents' surplus as a result of the learning process is required to be as high as the surplus of an optimal Nash equilibrium of the initially unknown game. ELE is a property of a set of learning algorithms with respect to a class of games. An optimal ELE should satisfy the following properties:

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 social surplus obtained by the agents within a polynomial number of steps will be at least (close to) the social surplus they could obtain, had the agents known the game from the outset and adopted an optimal (surplus maximizing) Nash equilibrium of it.

A tuple of learning algorithms satisfying the above properties for a given *class* of games is said to be an *Optimal Efficient Learning Equilibrium* (OELE) for that class. The definition above slightly deviates from the original definition in (Brafman & Tennenholtz 2004), since we require the outcome to yield the surplus of an optimal Nash equilibrium, while the original definition referred to the requirement that each agent would obtain expected payoff close to what he could obtain in *some* Nash equilibrium of the game.

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. This kind of requirement is typical of work in machine learning, where we require the learning algorithms to yield satisfactory results for *every* model taken from a set of models (without any Bayesian assumptions about the probability distribution over models). What the above definition borrows 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 surplus of an optimal Nash equilibrium of the (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.

In this paper we adopt the classical repeated game model. In such setting, a classical and intuitive requirement is that after each iteration an agent is able to observe the payoff it obtained and the actions selected by the other agents. Following (Brafman & Tennenholtz 2004), we refer to this as *imperfect monitoring*. In the *perfect monitoring* setting, the agent is also able to observe previous payoffs of other agents. Although perfect monitoring may seem like an exceedingly strong requirement, it is, in fact, either explicit or implicit in most previous work in multi-agent learning in AI (see e.g. (Hu & Wellman 1998)). In (Brafman & Tennenholtz 2004) the authors show the existence of ELE under perfect monitoring for any class of games, and its inexistence, in general, given imperfect monitoring. These results are based on the R-max algorithm for reinforcement learning in hostile environments (Brafman & Tennenholtz 2002). In Section 3 we show that the same results hold for OELE. However, this leaves us with a major challenge for the theory of multi-agent learning, which is the major problem tackled in this paper :

- Can one identify a general class of games where OELE exists under imperfect monitoring?

In this paper we address this question. We show the existence of OELE for the class of symmetric games – a very common and general class of games. Our proof is constructive, and provide us with appropriate efficient algorithms satisfying the OELE requirements. Indeed, our results imply the existence and the construction of an efficient protocol, that will lead to socially optimal behavior in situations which are initially unknown, when the agents follow the protocol; moreover, this protocol is stable against rational deviations by the participants. Notice that in many interesting situations, such as in the famous congestion settings studied in the CS/networks literature, the setting is known to be agent-symmetric, but it is initially unknown (e.g. the speed of service providers etc. is initially unknown). Although such symmetric settings are most common both in theoretical studies as well as in applications, dealing with the existence of an OELE in such settings is highly challenging, since in symmetric games an agent's ability to observe its own payoff (in addition to the selected joint action) does not directly teach it about other agents' payoffs (as in zero-sum games, common-interest games, and games with perfect monitoring, where ELE has been shown to exist in previous work).

Figure 1:

$$M_1 = \begin{pmatrix} 5,\ -5 & 3,\ -3 \\ -3,\ \ 3 & -2,\ \ 2 \end{pmatrix}$$

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

$$M_3 = \begin{pmatrix} 2,\ \ 2 & -10,\ 10 \\ 10,\ -10 & -5,\ -5 \end{pmatrix}$$

In the following section we provide a short review of basic notions in game-theory. Then, in Section 3, we formally define the notion of optimal efficient learning equilibrium and adapt previous results obtained on ELE to the context of OELE. In Section 4 we prove the main result of this paper: the (constructive) existence of an OELE under imperfect monitoring for a general class of games – the class of (repeated) symmetric games. The related algorithm is briefly illustrated in Section 5. For ease of exposition, we concentrate on two player games. The extension to n-player games is discussed in Section 6.

2. Game-Theory: some background

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., (Fudenberg & Tirole 1991).

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 $R_i : \times_{i \in I} A_i \rightarrow R$ for each $i \in I$. We let $\mathcal{A}$ denote the set $\times_{i \in I} A_i$ of *joint actions*. Agents' actions are also 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 there are only two players, the game can be described using a (bi)-matrix whose rows correspond to the possible actions of the first agent 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, unless stated otherwise, 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 Figure 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.

When considering the actions an agent can choose from, we allow agent i to choose among the set of probability distributions $A_i^m = \Delta(A_i)$ over his actions. The corresponding set of mixed action (strategy) profiles of the agents is denoted by $\bar{\mathcal{A}} = \times_{i \in I} A_i^m$. The payoff of an agent given such a profile is naturally defined using the expectation operator. We will therefore use $R_i(a)$ to refer to expected payoff of agent i when $a \in \bar{\mathcal{A}}$ is played.

A basic concept in game-theory is that of a *Nash equilibrium*. A joint action $a \in \bar{\mathcal{A}}$ is said to be a Nash equilibrium if for every agent i and every action profile a' such that a' differs from a in the action of agent i alone, it is the case that $R_i(a) \geq R_i(a')$. Thus, no agent has motivation to unilaterally change its behavior from a. A basic result of game theory is that every n-player game in strategic form, in which the agents' set of actions is finite possesses a Nash equilibrium in mixed strategies (where each agent can select a probability distribution of its available actions). Unfortunately, in general, there can be many Nash equilibria. An interesting type of Nash equilibria are the *optimal* Nash equilibria. A Nash equilibrium of a game is termed optimal if there is no other Nash equilibrium of the game in which the agents' surplus (i.e. the sum of agents' payoffs) is higher than in the prescribed equilibrium.

Other concepts, to be used later in the paper are the probabilistic maximin strategy and the safety-level value of a game G. A probabilistic maximin strategy for player i, is a mixed strategy $s \in argmax_{s' \in A_i^m} min_{s_{-i} \in \times_{j \neq i} A_j} R_i(s, s_{-j})$, and its value is the safety-level value.

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. The repeated-games model has been popular within both AI and game theory. In this paper we will therefore study learning in repeated games.

3. Optimal Efficient Learning Equilibrium

In this section we present a formal definition of optimal efficient learning equilibrium in the context of two-player repeated games. The generalization to n-player repeated games is relatively straightforward, but is omitted due to lack of space. We briefly discuss it in Section 6.

In a *repeated game* (RG) the players play a given game G repeatedly. We can view a repeated game M, 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 $P = [0, R_{max}]$. Let $S_{max}(G)$ be the maximal sum of agents' payoffs (a.k.a. the social surplus) obtained in some equilibrium of the game G. In the *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 $t \geq 0$, 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, given perfect monitoring, 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 the *imperfect monitoring* setup, all that a player can observe following the performance of its action is the payoff it obtained and the actions selected by the players. The player cannot observe the other player's payoff. More formally, in the imperfect monitoring setting, the set of possible histories of length t is $(A^2 \times P)^t$, and the set of possible histories, H, is the union of the sets of possible histories for all $t \geq 0$, where $(A^2 \times P)^0$ is the empty history. An even more constrained setting is that of *strict imperfect monitoring*, where the player can observe its action and its payoff alone. Given an RG, M, 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.

Assume we consider games with k actions, $A = \{a_1, \ldots, a_k\}$. For every repeated game M, selected from a class of repeated games $\mathcal{M}$, where M consists of repeatedly playing a game G defined on A, let $n(G) = (N_1(G), N_2(G))$ be an optimal Nash equilibrium of the (one-shot) game G, and denote by $NV_i(n(G))$ the expected payoff obtained by agent i in that equilibrium. Hence, $S_{max}(G) = NV_1(n(G)) + NV_2(n(G))$. A policy profile (π, ρ) is an *optimal efficient learning equilibrium*(OELE) with respect to the class $\mathcal{M}$, if for every $\epsilon > 0, 0 < \delta < 1$, there exists some $T > 0$, polynomial in $\frac{1}{\epsilon}$, $\frac{1}{\delta}$, and k, such that for every $t \geq T$ and every RG, $M \in \mathcal{M}$ (associated with a one-shot game G), $U_1(M, \pi, \rho, t) + U_2(M, \pi, \rho, t) \geq S_{max}(G) - \epsilon$, and if player 1 deviates from π to π' in iteration l, then $U_1(M, \pi', \rho, l + t) \leq U_1(M, \pi, \rho, l + t) + \epsilon$ 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. 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.

Our requirement therefore is that learning algorithms will be treated as strategies. In order to be individually rational they should be the best response for one another. The strong requirement made in OELE is that deviations will not be beneficial regardless of the actual game, where the identity of this game is initially unknown (and is taken from a set of possible games). In addition, the agents should rapidly obtain a desired value, and the loss of gain when deviating should also be materialized efficiently. 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 value the players would have obtained in an optimal Nash equilibrium had they known the game. Moreover, and most importantly, as mentioned earlier, the learning algorithms themselves should be in equilibrium. We remark that 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 (Holzman *et al.* 2004).

The definition of OELE is of lesser interest if we cannot provide interesting and general settings where OELE exists. By adapting the results of (Brafman & Tennenholtz 2004) to the context of OELE we can show that:

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

In particular, there is an algorithm that leads to an OELE for the any class of games with perfect monitoring. Thus, agents that use this algorithm can attain the average reward of an optimal Nash equilibrium of the actual game without prior knowledge about the game played, and deviation from the algorithm will not be beneficial.

However,

Theorem 2 *Under imperfect monitoring, an OELE does not always exist.*

This leaves us with a major challenge for the theory of multi-agent learning. Our aim is to identify a general setting where OELE exists under imperfect monitoring. Needless to say that a constructive proof of existence for such general setting, will provide us with a most powerful multi-agent learning technique. This is the topic of the following sections.

4. OELE for Symmetric Games with Imperfect Monitoring

A game G is symmetric if for every actions $a, b \in A$, the payoff of agent 1 for (a, b) equals the payoff of agent 2 for (b, a), i.e. $R_1(a, b) = R_2(b, a)$. In fact, the best known games from the game theory literature are symmetric.

Our aim is to show the existence of an OELE for symmetric games. We will make use of the following Lemma (proof omitted, due to lack of space).

Lemma 1 *Let G be a symmetric 2-player game where each agent can choose actions from among $A = \{1, 2, \ldots, k\}$, and agent i's payoff function is R_i ($i = 1, 2$). Let $s \in argmax_{s' \in A^2} U_1(s') + U_2(s')$, and let $r = U_1(s) + U_2(s)$; i.e., s is surplus maximizing and leads to social surplus of r. Let $v(B)$ be the safety level value that can be guaranteed by a player when both players can choose only among actions in $B \subseteq A$, and let $v = \max_{\{B: B \subseteq A\}} v(B)$. Then $v \leq \frac{s}{2}$.*

We now present our main theorem:

Theorem 3 *Let $A = \{a_1, \ldots, a_k\}$ be a set of possible actions, and consider the set of symmetric games with respect to this set of actions. Then, there exists an OELE for this set of games under imperfect monitoring.*

Proof(sketch):

Consider the following algorithm, termed the Sym-OELE algorithm.

The Sym-OELE algorithm:

Player 1 performs action a_i one time after the other for k times, for $i = 1, 2, ..., k$. 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 socially optimal (not necessary individually rational) strategy profile $s = (s_1, s_2)$ is selected, i.e. $s \in argmax_{s' \in A^2}(R_1(s') + R_2(s'))$; agent 1 then plays s_1 in odd iterations and plays s_2 in even iterations, while agent 2 plays s_2 in odd iterations and s_1 in even iterations. 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: W.l.o.g let the agent be player 1. 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 game where its aim is to minimize the adversary's payoff. Notice that these payoffs might be unknown. The corresponding punishing procedure will be described below. We will use the following general notation: given a game G_1 we will refer to the modified (constant sum) game as G'_1. A punishing strategy in the original game (minimizing the adversary's payoff) will be a probabilistic maximin of the modified game.

Initialize: The agent selects actions randomly until it knows the payoffs for all joint actions in the set $S_a = \{(x, a), x \in A\}$ for some $a \in A$.

We say that a column which corresponds to action b of the adversary is *known*, if the agent has observed her payoffs for strategy profiles (y, b) for all $y \in A$. Denote by C the set of actions that correspond to known columns at each point in time, and let G' denote the restriction of G only to actions (of both agent and adversary) that correspond to known columns, i.e. G' is a squared matrix game containing all entries of the form (a, b) such $a, b \in C$. Since G is symmetric, *all* the payoffs in G', of both of the players, are known to the agent. Let G'' denote the modified version of G', i.e., where the agent's payoffs are the R_{max} complements of the adversary's payoffs in G'.

Repeat: Compute and Act: Compute the optimal probabilistic maximin of G'' and execute it with probability $1 - \alpha$, where $\alpha = \frac{1}{Qk}$; Q will be determined later and will be polynomial in the problem parameters. With probability α uniformly select a random action 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, then we keep record of the reward associated with (a, a'). We revise G' and G'' appropriately when a new column becomes known. That is, if following the execution of (a, a') the column a' becomes known, then we should add a' to C and modify G' accordingly.

Claim 1 *The SYM-OELE algorithm, when adopted by the players, is indeed an OELE.*

Our proof will be a result of the following analysis with regard to the above algorithm.

Given the current values of C and G', then after Q^2k^4 iterations in which actions corresponding to still unknown columns are played, at least one of these actions should have been played by the adversary, at least Q^2k^3 times. (This is just the pigeonhole principle.)

The probability that if an action a, associated with unknown column, is played Q^2k^3 times by the adversary, the corresponding column will be unknown is bounded by $k(1-\frac{\alpha}{k})^{Q^2k^3}$. (This is the probability we will miss an entry (b, a) for some b, multiplied by the number of possible b's.)

Take $\alpha = \frac{1}{Qk}$, then $k(1-\frac{\alpha}{k})^{Q^2k^3} = k(1-\frac{1}{Qk^2})^{Qk^2Qk} < ke^{-Qk}$ (since $(1-\frac{1}{n})^n < \frac{1}{e}$). Hence, the probability that an unknown column will not become known after it is played Q^2k^3 times is bounded by ke^{-Qk}; the probability that no unknown column will become known after actions associated with unknown columns are played Q^2k^4 times is also bounded by ke^{-Qk}.

Choose Q such that $ke^{-Qk} < \frac{\delta}{3k}$. Notice that Q can be chosen to be bounded by some polynomial in the problem parameters.

The above implies that after $T' = Q^2k^6$ iterations, either the number of times where actions corresponding to unknown columns are selected by the adversary is less than Q^2k^5, or all the game G becomes known after T' iterations of that kind with probability greater than or equals to $1-\frac{\delta}{3}$. This is due to the fact in any Q^2k^4 iterations where actions associated with unknown columns are played, a new column will become known with probability of failure of at most $\frac{\delta}{3k}$ (as we have shown above); by applying this argument k times (i.e. for k sets of Q^2k^4 iterations like that) we get that the probability not all columns will become known is bounded by $k\frac{\delta}{3k} = \frac{\delta}{3}$.

Notice that whenever a column which corresponds to action a is known, the game G' is extended to include all and only the actions that correspond to the known columns. Since the game is symmetric, whenever the agent's payoffs for all entries of the form (a, b) for all $b \in C \subseteq A$ and for all $a \in A$ are known then the payoffs for **both** players are known to the agent in the game G' where the actions are only those in C.

Hence, after $T = QkT'$ iterations the expected payoff of an adversary, which can guarantee itself at most v (when playing against a punishing strategy in some game G' as above) is bounded by $\frac{T'R_{max}+(Qk-1)T'((1-\alpha)v+\alpha R_{max})}{T}$. This is due to our observation earlier: T' bounds the number of times in which the adversary can play an unknown column, and v is the best value that he can get playing a known column. The calculation takes also into account that with probability α, when the agent explores, the adversary might get the R_{max} payoff. Simplifying, we get that the adversary can guarantee itself no more than $v + \frac{2R_{max}}{Qk}$. This implies that the average payoff of the adversary would be smaller than $v + \epsilon$ when $Q > \frac{2R_{max}k}{\epsilon}$.

It is left to show that $v \leq \frac{s}{2}$ where $\frac{s}{2}$ is what the adversary would have obtained if we would have followed the prescribed algorithm. This however follows from Lemma 1. ∎

Although there are many missing details, the reader can verify that the Sym-OELE is efficient, and that indeed determines an OELE. In fact, the most complex operation in it is the computation of probabilistic maximin, which can be carried out using linear programming. Moreover, notice that the algorithm leads to optimal surplus, and not only to the surplus of an optimal Nash equilibrium. In no place there is a need to compute a Nash equilibrium.

5. The Sym-OELE algorithm: an example

To illustrate the algorithm, we now consider a small example of using the following 3x3 game:

(5,5)	(4,0)	(3,8)
(0,4)	(-2,-2)	(3,2)
(8,3)	(2,3)	(3,3)

If the adversary does not deviate from the algorithm, after 9 iterations, the game will become known to the agents, and (1,3) and (3,1) will be played interchangeably; here we use (i, j) to denote the fact player 1 plays action number i and player 2 plays action number j.

Suppose that the adversary deviates immediately. In that case, the first agent will select actions uniformly. With high probability, after a number of steps, she will know her payoffs for one of the columns. Assume that she knows her payoffs for column 1. In that case, $C = \{1\}$ and G'' is the single action game:

(3,5)

The agent now plays action 1 almost always, occasionally playing randomly. Suppose that the adversary always plays column 1. In that case, the adversary's payoff will be slightly less than 5, which is lower than the value he would have obtained by following the algorithm (which is 5.5). If the adversary plays other columns as well, at some point, the agent would learn another column. Suppose the agent learned column 2, as well. Now $C = \{1, 2\}$ and G'' is the game:

(3,5)	(8,0)
(4,4)	(10,-2)

Now, the agent will play action 2 most of the time. This means that the adversary's average payoff will be at most a little over 4. Finally, if the adversary plays column 3, occasionally, the agent will learn its value, too, etc.

6. n-player games

When extending to n-player games, for any fixed n, we assume that there are private channels among any pair of agents. Alternatively, one can assume that there exists a mediator who can allow agents to correlate their strategies, by sending them private signals, as in the classical definition of correlated equilibrium (Aumann 1974). Although details are omitted due to lack of space, we will mention that our setting and results are easily extended to that setting of correlated equilibrium, and it allows us to extend the discussion to the case of n-players. The above extended setting (where we either refer to private communication channels, or to a correlation device) implies that any set of agents can behave as a single master-agent, whose actions are action profiles of the set of agents, when attempting to punish a deviator. Given this, the extension to n-player games is quite immediate. The agents will be instructed first to learn the game entries, and (once the game is known) to choose a joint action which is surplus maximizing, $s = (s_1, s_2, \ldots s_n)$, and behave according to the $n!$ permutations repeatedly. This will lead each agent to an average payoff which is $\frac{\Sigma_{i=1}^{n} U_i(s)}{n}$. This will yield convergence to optimal surplus in polynomial time. In order to punish a deviator, all other players will behave as one (master-agent) whose aim is to punish an adversary.

Let $B = A^{n-1}$ be the action profiles of the above-mentioned master agent. When punishing the adversary we will say that a column corresponding to an action $a \in A$ of the adversary is known if all other agents (i.e. the master agent) know their payoffs (b, a), for every $b \in B$. Let A' be the set of actions for which the corresponding columns are known then we get that all payoffs (also of the adversary) in G', where agents can choose actions only from among A', are known. Hence, the proof now follows the ideas of the proof for the case of 2-player games: a punishing strategy will be executed with high probability w.r.t to G', and a random action profile will be selected with some small probability.

7. Conclusion

The algorithm and results presented in this paper are, to the best of our knowledge, the first ones to provide efficient multi-agent learning techniques, satisfying the natural but highly demanding property that the learning algorithms should be in equilibrium given imperfect monitoring. We see the positive results obtained in this paper as quite surprising, and extremely encouraging. They allow to show that the concept of ELE, and OELE in particular, is not only a powerful notion, but does also exist in general settings, and can be obtained using efficient and effective algorithms.

One other thing to Notice is that although the Sym-OELE algorithm has a structure which may seem related to the famous folk theorems in economics (see (Fudenberg & Tirole 1991)), it deals with issues with quite different nature. This is due to the fact we need to punish deviators under imperfect monitoring, given there is no information about entries in the game matrix.

Taking a closer look at the Sym-OELE algorithm and its analysis, it may seem that the agent needs to know the value of R_{max}, i.e., the maximal possible payoff in order to execute the algorithm. In fact, this information is not essential. The agent can base her choice of the parameter Q on the maximal observed reward so far, and the result will follow. Hence, the algorithm can be applied without any limiting assumptions.

References

Aumann, R. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1:67–96.

Bowling, M., and Veloso, M. 2001. Rational and covergent learning in stochastic games. In *Proc. 17th IJCAI*, 1021–1026.

Brafman, R. I., and Tennenholtz, M. 2002. R-max – a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research* 3:213–231.

Brafman, R. I., and Tennenholtz, M. 2004. Efficient learning equilibrium. *Artificial Intelligence* 159:27–47.

Claus, C., and Boutilier, C. 1997. The dynamics of reinforcement learning in cooperative multi-agent systems. In *Proc. Workshop on Multi-Agent Learning*, 602–608.

Conitzer, V., and Sandholm, T. 2003. Awesome: a general multiagent learning algorithm that coverges in self-play and learns best-response against stationary opponents. In *Proceedings of the 20th ICML*, 83–90.

Erev, I., and Roth, A. 1998. Predicting how people play games: Reinforcement learning in games with unique strategy equilibrium. *American Economic Review* 88:848–881.

Fudenberg, D., and Levine, D. 1998. *The theory of learning in games*. MIT Press.

Fudenberg, D., and Tirole, J. 1991. *Game Theory*. MIT Press.

Greenwald, A.; Hall, K.; and Serrano, R. 2002. Correlated q-learning. In *NIPS workshop on multi-agent learning*.

Holzman, R.; Kfir-Dahav, N.; Monderer, D.; and Tennenholtz, M. 2004. Bundling Equilibrium in Combinatorial Auctions. *Games and Economic Behavior* 47:104–123.

Hu, J., and Wellman, M. 1998. Multi-agent reinforcement learning: Theoretical framework and an algorithms. In *Proc. 15th ICML*.

Kaelbling, L. P.; Littman, M. L.; and Moore, A. W. 1996. Reinforcement learning: A survey. *Journal of AI Research* 4:237–285.

Kalai, E., and Lehrer, E. 1993. Rational Learning Leads to Nash Equilibrium. *Econometrica* 61(5):1019–1045.

Littman, M. L. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Proc. 11th ICML*, 157–163.