

Strategic learning in games with symmetric information

Olivier Gossner^{a,b,*} and Nicolas Vieille^c

^a *THEMA, Université Paris X-Nanterre, 200, Avenue de la République, 92001 Nanterre cedex, France*

^b *CORE, Université Catholique de Louvain, Belgique*

^c *École Polytechnique and HEC, Département Finance et Économie, 1, rue de la Libération, 78351 Jouy en Josas, France*

Received 10 January 2000

Abstract

This article studies situations in which agents do not initially know the effect of their decisions, but learn from experience the payoffs induced by their choices and their opponents'. We characterize equilibrium payoffs in terms of simple strategies in which an exploration phase is followed by a payoff acquisition phase.

© 2002 Elsevier Science (USA). All rights reserved.

JEL classification: C72

Keywords: Public value of information; Games with incomplete information; Bandit problems

1. Introduction

This paper analyzes situations in which agents do not initially know the effect of their decisions, but learn from experience the payoffs induced by their choices and their opponents'. Our model falls into the class of repeated games with incomplete information and signals. Our main assumption is the symmetry of information between the players. Hence, all players have the same initial information on the payoff function and receive the same additional information after every stage.

This assumption is motivated by two reasons. First, we believe it is realistic enough to apply in many economic situations (for instance, prices and quantities sold by a firm are

* Corresponding author.

E-mail addresses: olivier.gossner@u-paris10.fr (O. Gossner), vieille@hec.fr (N. Vieille).

commonly observable by others). Second, whereas equilibria may fail to exist for general repeated games with incomplete information, results due to Kohlberg and Zamir (1974) and Forges (1982) for the zero-sum case, and to Neyman and Sorin (1998) for the non-zero sum case prove their existence when information is symmetric.

We essentially characterize the set of uniform Nash equilibrium payoffs, and provide some results on perfect Bayesian equilibrium payoffs. The traditional motivation for using the notion of uniform equilibrium is that an uniform equilibrium remains an ϵ -equilibrium in many contexts of uncertainty about time-preferences and/or about the duration of the game. In addition, it highlights an essential feature of our model. In a one-player setup, the optimal level of learning/experimentation is obtained by balancing the costs and benefits of learning. In the absence of discounting, learning is *costless*. In our model, partial revelation may be an equilibrium outcome. This is linked to the public good aspect of information, and is discussed below.

In the general case, we prove that full exploration still constitutes an equilibrium. Namely, we exhibit equilibria in which players explore the payoffs induced by every action profile before they play an equilibrium of the corresponding infinitely repeated game with perfect information. Nevertheless, this family of equilibria can be Pareto dominated by equilibria with partial revelation only. Hirshleifer (1971) already pointed out that public information can be socially damaging.

More generally, we exhibit a family of equilibria in which an exploration phase is followed by a payoff acquisition phase. At each stage of the exploration phase, players choose a profile of actions which has not been played before. They can also choose to stop exploring, in which case the payoff acquisition phase starts. During this phase, which lasts forever, the only actions played are the ones which were experienced during the exploration phase (provided no player deviates). Therefore, the only information players have on the payoffs is the information obtained during the exploration phase.

Conversely, we prove that any equilibrium is payoff equivalent to a convex combination of equilibria of the preceding form. To do this, we show that we can reduce all histories on the equilibrium path in such a way that exploration only takes place during the first stages.

The particular case of zero-sum, two-player games has been studied in a strand of literature starting with Hannan (1957). It is proven that each player can guarantee the value of the true underlying game. Therefore, no player can benefit from the initial lack of information on the payoffs as long as these payoffs are announced after each turn. We need an extension of their result to any number of players. Again, we obtain that the min max level of a player is the min max level in which all information on the payoffs is revealed. This preliminary result also characterizes player's individually rational levels for the non-zero sum case.

The theory of two-player repeated games with incomplete information (see Aumann and Maschler (1995), Forges (1992) for the general theory) usually assumes that actions are observable whereas payoffs are not. With lack of information on more than one side (no player is more informed than the other) equilibria may not exist. The only general existence theorems are obtained with discounting on the payoffs (a fixed point argument applies) or with lack of information on one side only. With lack of information on one side, Hart (1985) provides a characterization of equilibrium payoffs: basically, at each stage of the repetition the informed player reveals a bit more of his information to the uninformed.

A result due to Aumann and Hart (1986) shows that this revelation process can be endless; not all equilibria are payoff-equivalent to equilibria in which revelation comes down to a finite number of stages at the beginning of the game.

Some attention has been paid to the case where each player is informed of his own payoff function. With lack of information on both sides, Koren (1988) proves that any equilibrium is payoff-equivalent to an equilibrium in which each agent is perfectly informed of the true profile of payoff functions, and shows that a finite number of stages suffices for the whole process of information transmission. Yet, equilibria can fail to exist.

We first discuss an example to introduce the main features of our model in Section 2. Section 3 presents the model. The zero-sum case is studied in Section 4. In Section 5, we introduce scenarios as a class of strategies with respect to which we characterize equilibrium payoffs in the general non-zero-sum case. Section 6 is devoted to the proof of the main theorem. Section 7 contains discussions of perfect Bayesian equilibrium, discounted games, and few miscellaneous examples.

2. Discussion and example

We are concerned with equilibria of games where players collectively learn their profile of payoff functions. Initially, players know that the game being played is one of a finite family $(G(k))_{k \in K}$, and they share a common prior p on K . We denote by $G_\infty(p)$ the infinitely repeated game in which k is drawn according to p at stage 0 and in which after each subsequent stage, the action profile played and the payoff profile yielded by k and by the action profile are publicly announced.

During the play of $G_\infty(p)$, players learn more and more about their profile of payoff functions. Eventually, they can fully learn the underlying game $G(k)$ and play in the infinite repetition $G_\infty(k)$ of $G(k)$. The Folk theorem characterizes all Nash equilibrium payoffs of $G_\infty(k)$ for each k . We characterize equilibria in terms of their corresponding levels of exploration.

In a game in which any action profile identifies the state of nature, all equilibria must be revealing. Also, zero-sum games have the property that all equilibria are payoff equivalent to full revelation of the payoff function. Nevertheless, in the general case, some equilibria of $G_\infty(p)$ can be sustained only if there is no complete learning of k , as shown by the coming example.

Example 1. Consider a situation of duopoly in which each firm can be peaceful (P) or initiate a war (W). When a war is initiated by any of the two firms, a winner is declared that also wins all subsequent wars. For instance we may imagine that one of the two firms possesses a stronger technology but the identity of the stronger is unknown until a war occurs. The true game played can be $G(1)$ or $G(2)$, where $G(i)$ happens when i is the strongest firm:

$$\begin{array}{cc}
 & \begin{array}{cc} W & P \end{array} \\
 \begin{array}{c} W \\ P \end{array} & \begin{array}{|cc|} \hline 2, -2 & 2, -2 \\ \hline 2, -2 & 1, 1 \\ \hline \end{array}
 \end{array}
 G(1), \quad
 \begin{array}{cc}
 & \begin{array}{cc} W & P \end{array} \\
 \begin{array}{c} W \\ P \end{array} & \begin{array}{|cc|} \hline -2, 2 & -2, 2 \\ \hline -2, 2 & 1, 1 \\ \hline \end{array}
 \end{array}
 G(2).$$

Players assess initial probability $p = (1/2, 1/2)$ on the game being $G(1)$ or $G(2)$.

First, note that it is an equilibrium of $G(p)$ to play (W, W) forever, thus revealing the true payoff function and playing a Nash equilibrium of the associated infinitely repeated game. In fact, the only equilibrium payoffs of $G_\infty(1)$ and $G_\infty(2)$ are $(2, -2)$ and $(-2, 2)$, respectively.

There also exist equilibria in which war is never declared. After W is played once, the payoff function is revealed and one of the two players has W as a dominant strategy. Thus after a war the winner gets 2 forever and the loser gets -2 forever. If at some stage no war has ever been declared, each player anticipates to being strongest or the weakest with equal probabilities. The expected payoff if a war is declared is 0, which is less than the payoff of 1 if peace lasts forever. Therefore it is an equilibrium that players remain peaceful forever. In this equilibrium no war is ever declared because each player fears being the loser.

3. Model

3.1. The game

The set of players is a finite set I . Each player i has a finite set of actions A^i . The finite set K of states of nature is initially endowed with probability $p \in \Delta(K)$ with full support (for any finite set S , $\Delta(S)$ is the set of probabilities over S). For each $k \in K$ is given a game in strategic form $G_k = ((A^i)_{i \in I}, g_k: A \rightarrow \mathbb{R}^J)$ (as usual $A = \prod_i A^i$, $A^{-i} = \prod_{j \neq i} A^j$ and we use similar notations whenever convenient).

The game $G_\infty(p)$ unfolds as follows.

step 0: a state $k \in K$ is drawn according to some distribution p .

step n , $n \geq 1$: The players are told the past sequence of actions profiles $(a_t)_{t < n}$ and the corresponding sequence of payoffs. They then choose independently actions a_n^i , $i \in I$.

The above description, including p , is common knowledge. Notice that all the players have the same information about k , and receive the *same* additional information. Hence, no asymmetry of information can possibly arise during the play.

We make the innocuous assumption that a state of nature contains no more than the information relative to the payoffs: for any two distinct states k_1, k_2 , the payoff functions g_{k_1} and g_{k_2} differ.

3.2. Strategies

We denote by $H_\infty = K \times A$ the set of plays. For $n \geq 1$, we define a σ -algebra \mathcal{H}_n on H_∞ which represents the information available at stage n . Let $h, h' \in H_\infty$, with $h = (k, (a_p)_{p \geq 1})$, $h' = (k', (a'_p)_{p \geq 1})$. We say that h and h' are n -equivalent if $a_p = a'_p$, and $g_k(a_p) = g_{k'}(a_p)$, for each $p < n$. It captures the intuitive idea that, prior to playing in stage n , the players are unable to distinguish the two plays h and h' . This equivalence relation partitions H_∞ into finitely many equivalence classes. We denote by \mathcal{H}_n the σ -algebra over H_∞ induced by this partition. Note that $(\mathcal{H}_n)_n$ is a filtration over H_∞ , i.e.,

$\mathcal{H}_n \subset \mathcal{H}_{n+1}$ for each n . We define $\mathcal{H}_\infty = \sigma(\bigcup_n \mathcal{H}_n)$; it is the coarsest σ -algebra over H_∞ which contains each \mathcal{H}_n .

A pure (respectively behavioral) strategy of player i is a sequence $\sigma^i = (f_n^i)_{n \geq 1}$, where f_n^i is a measurable map $f_n^i: (H_\infty, \mathcal{H}_n) \rightarrow A^i$ (respectively to $\Delta(A^i)$) which describes the behavior of player i in stage n . The space of behavioral strategies of player i is denoted by Σ^i .

Given p , any profile $\sigma \in \Sigma$ induces a probability distribution $P_{p,\sigma}$ over the set of plays $(H_\infty, \mathcal{H}_\infty)$. We write $P_{k,\sigma}$ for the distribution on \mathcal{H}_∞ conditional on $k \in K$. Note that $P_{k,\sigma} = P_{\delta_k,\sigma}$ where δ_k is the unit mass on K and $P_{p,\sigma} = \sum_k p(k)P_{k,\sigma}$. For any \mathcal{H}_∞ -measurable bounded random variable X we write $\mathbf{E}_{p,\sigma} X$ and $\mathbf{E}_{k,\sigma} X$ for the expectations of X under $P_{p,\sigma}$ and $P_{k,\sigma}$ respectively.

The action a_n^i played by i and the action profile $a_n = (a_n^i)_{i \in I}$ at stage n are random variables over $(H_\infty, \mathcal{H}_\infty)$. Then, $g_n = g_k(a_n)$ is the payoff vector in stage n if the true state of nature is k , and for $\sigma \in \Sigma$, $\gamma_n(\sigma) = \mathbf{E}_{p,\sigma} \{\frac{1}{n} \sum_{m=1}^n g_m\}$ is the expected average payoff up to stage n . Also $\gamma_n(k, \sigma) = \mathbf{E}_{k,\sigma} \{\frac{1}{n} \sum_{m=1}^n g_m\}$ is the average payoff in state k . Subscripts are then used to denote both stages and states of nature, but no confusion will possibly arise.

We denote by $G_n(p)$ the n -stage version of $G_\infty(p)$, it has strategy sets Σ^i , and payoff function γ_n .

3.3. Equilibrium notions

We recall from (Mertens et al., 1994) the notion of uniform equilibrium.

Definition 3.1. A profile $\sigma \in \Sigma$ is an uniform equilibrium profile if the following two conditions are satisfied:

- (1) for every $k \in K$, $\gamma(k, \sigma) = \lim_{n \rightarrow \infty} \gamma_n(k, \sigma)$ exists;
- (2) for each $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that, provided $n \geq N$, σ is an ε -equilibrium in $G_n(p)$.

We then say that $\gamma(\sigma) = (\gamma(k, \sigma))_{k \in K}$ is an uniform equilibrium payoff.

These are about the most stringent requirements for equilibrium: the *same* profile is an ε -equilibrium in *every* finitely repeated game provided the number of repetitions is large enough. Furthermore this implies that this profile is also an ε -equilibrium in *every* discounted game, provided the payoffs are sufficiently little discounted.

We denote by $E(p)$ the set of equilibrium payoffs of $G_\infty(p)$.

3.4. Individually rational levels

As usual for repeated games, it is essential to characterize the level at which players other than i can punish player i . The corresponding concept is that of min max.

We say that $v^i(p)$ is the (uniform) min max for player i if the following two conditions are satisfied:

(1) *Players $-i$ can guarantee $v^i(p)$* : there exists $\sigma^{-i} \in \Sigma^{-i}$ such that

$$\limsup_n \max_{\sigma^i} \gamma_n^i(\sigma^{-i}, \sigma^i) \leq v^i(p);$$

(2) *Player i can defend $v^i(p)$* : for every $\sigma^{-i} \in \Sigma^{-i}$, there exists σ^i such that

$$\liminf_n \gamma_n^i(\sigma^{-i}, \sigma^i) \geq v^i(p).$$

If $G_\infty(p)$ happens to be a game of complete information ($|K| = 1$, or p is a unit mass on some $k \in K$), the min max for player i exists and coincides with the min max of the corresponding one-shot game, defined as:

$$v_k^i = \min_{s^{-i} \in \prod_{j \neq i} \Delta(A^j)} \max_{s^i \in \Delta(A^i)} \mathbf{E}_{s^{-i}, s^i} g_k^i(a^{-i}, a^i).$$

When players $j \neq i$ can correlate their strategies, Σ^{-i} and $\prod_{j \neq i} \Delta(A^j)$ in the above definitions must be replaced by $\Delta(\Sigma^{-i})$ and $\Delta(A^{-i})$, respectively. This defines the correlated min max for player i in $G(p)$ and $G(k)$ that we denote $w^i(p)$ and w_k^i .

In general, $w^i(p) < v^i(p)$ and $w_k^i < v_k^i$, except with two players where equality holds. In Section 4 we characterize $v^i(p)$ and $w^i(p)$.

3.5. Correlated and communication equilibria

In many situations, it is natural to assume that players have the opportunity to communicate during the play of the game. In the most general framework, players can communicate between any two stages through the use of any communication mechanism that sends them back private, stochastically drawn signals (Forges, 1992).

When we assume players can communicate between any two stages using any communication mechanism, the (uniform or Banach) equilibrium payoffs induced on the infinitely repeated game are called the extensive form communication equilibria. Their set is denoted $E_{\text{com}}(p)$. We also consider some common limitations on the mechanisms used to communicate. First, if players can only communicate before the game starts, we speak of normal form communication equilibria, and the corresponding set of equilibrium payoffs is $E_{\text{com}}^*(p)$. Second, if we assume that players' signals do not depend on their messages (of equivalently if the mechanism receives no inputs), the communication mechanism is called a correlation device (Aumann, 1974). This defines the two corresponding sets of extensive form correlated equilibrium payoffs $E_{\text{cor}}(p)$ and normal form correlated equilibrium payoffs $E_{\text{cor}}^*(p)$. Furthermore, when the correlation devices are restricted to be public (every player gets the same signal), the equilibria are called public correlated equilibria (in extensive form or not) and the sets of equilibrium payoffs are denoted $E_{\text{pub}}^*(p)$ and $E_{\text{pub}}(p)$.

4. The zero-sum case

The characterization of the min max in the two-player, zero-sum case is well-known. The following result is an immediate consequence of a much more powerful result obtained

independently by Hannan (1957), and Baños (1968) among others. We refer the reader to Foster and Vohra (1999) for a discussion of this result and the relevant literature.

Theorem 4.1. *Assume $N = 2$. The min max for player i in $G_\infty(p)$ exists and:*

$$v^i(p) = \mathbf{E}_p v_k^i = \sum_k p_k v_k^i.$$

Let now the number N of players be arbitrary. By viewing players $-i$ as a single player, the following characterization of the correlated min max is a direct consequence of Theorem 4.1.

Corollary 4.1. *The correlated min max for player i in $G_\infty(p)$ exists and:*

$$w^i(p) = \mathbf{E}_p w_k^i = \sum_k p_k w_k^i.$$

To study situations where correlation mechanisms are ruled out, we need an extension of Theorem 4.1. We now state this extension.

Theorem 4.2. *The min max for player i in $G_\infty(p)$ exists and*

$$v^i(p) = \mathbf{E}_p v_k^i = \sum_k p_k v_k^i.$$

The preceding results are powerful tools that show that the two min max for i in $G_\infty(p)$ are the same as in the game in which the state of nature is *publicly* revealed.

In other words, as long as payoffs are publicly revealed, i cannot be worse off neither can he take advantage of the fact that the game has initially incomplete information on the payoffs. Of course, this holds only for zero-sum games.

The property of Theorem 4.2 is deeply related to the observability of payoffs, and hardly to the assumption of symmetric information. In order to emphasize this point, we prove more than the statement of Theorem 4.2, and consider situations of asymmetric information. We prove that:

- (1) *even if player i is fully informed of the realized state k , while players $-i$ are not even informed of p , players $-i$ can punish player i down to v_k^i , whatever be k ;*
- (2) *even if player i is told only p , while each player of the coalition $-i$ is fully informed of k , player i can still defend v_k^i in every state k .*

Proof of Theorem 4.2. We provide here only the intuition of the proof. For a detailed proof, the reader is referred to Appendix A. We prove the claim for player i and, for notational convenience, suppress any reference to i in the payoffs.

To guarantee v_k . We construct $\sigma^{-i} \in \Sigma^{-i}$ such that,

$$\forall \epsilon, \exists N_\epsilon, \forall \sigma^i, \forall n \geq N, \forall k \quad \mathbf{E}_{k,\sigma}[\bar{g}_n] \leq v_k + \epsilon. \quad (1)$$

First, we argue that it is enough to construct, for each ϵ , a profile σ_ϵ^{-i} for which (1) is satisfied. Indeed, for any sequence (ϵ_n) decreasing to 0, the profile σ^{-i} defined as: play $\sigma_{\epsilon_1}^{-i}$ for N_{ϵ_1} stages, then forget the past and play $\sigma_{\epsilon_2}^{-i}$ for N_{ϵ_2} stages, etc. would then satisfy (1) for each ϵ .

Therefore, let $\epsilon > 0$. Denote by $A^i(n)$ the set of those actions $a^i \in A^i$ which consequences are known at stage n , i.e., those a^i such that *all* action combinations (a^i, a^{-i}) , $a^{-i} \in A^{-i}$ have been played at least once prior to stage n .

We define σ^{-i} as: play $(1 - \epsilon)\sigma^{-i}(k, A^i(n)) + \epsilon e^{-i}$ in stage n , where $\sigma^{-i}(k, A^i(n))$ is an optimal strategy of players $-i$ in the (complete information) one-shot game where player i is restricted to $A^i(n)$, and e^{-i} is some distribution with full support. (At stage n , player i knows the restriction of g_k to $A^i(n)$; therefore, this restricted game may be viewed as a one-shot game with complete information.)

At every stage, every action combination of players $-i$ is played with a positive probability, bounded away from 0. Therefore, there cannot be many stages, on average, in which player i chooses an action which consequences are not yet fully known. On the other hand, whenever player i chooses an action in $A^i(n)$, his expected payoff against $\sigma^{-i}(k, A^i(n))$ does not exceed v_k .

To defend v_k . We prove that for every $\sigma^{-i} \in \Sigma^{-i}$, there exists $\sigma^i \in \Sigma^i$ such that $\forall \epsilon > 0$:

$$\exists N_\epsilon, \forall n \geq N_\epsilon, k \in K, \quad \mathbf{E}_{k, \sigma^{-i}, \sigma^i}[\bar{g}_n] \geq v_k - \epsilon. \quad (2)$$

Moreover, N_ϵ may be chosen independently of σ^{-i} .

As in the first part of the proof, we let $\epsilon > 0$ and σ^{-i} . We define a strategy σ_ϵ^i and prove that it satisfies (2).

We denote by $\bar{\sigma}_n^{-i}$ the distribution of players $-i$'s actions in stage n , conditional on the information held by player i , and by p_n the conditional distribution over K .

Define σ_ϵ^i as: play $(1 - \epsilon)\sigma^i(p_n, \bar{\sigma}_n^{-i}) + \epsilon e^i$ in stage n , where $\sigma^i(p_n, \bar{\sigma}_n^{-i})$ is a best reply of player i to the correlated distribution $\bar{\sigma}_n^{-i}$ in the game with payoff function $\sum_k p_n(k)g_k$.

To establish (2), two main arguments are used. First, it is shown as in the previous part of the proof that there are not too many stages in which there is a non-small probability that players $-i$ pick an action combination which consequences have not been fully experienced in the past. Second, we rely on a classic result in the literature on reputation effects or merging due to Fudenberg and Levine (1992) which states roughly that most of the time, the distribution of players $-i$'s actions anticipated by player i is quite close to the *true* distribution.

Bringing these two parts together yields the result. Consider any stage in which *both* the anticipation of player i is good *and* there is only a small probability that players $-i$ select an action combination which is not completely known. In that stage, the expected payoff to player i is at least v_k minus some small quantity. \square

Proof of Corollary 4.1. Consider the two players game $\bar{G}(p)$ where player I has strategy set A^{-i} player II has strategy set A^i and the payoff function to II is g^i . Observe that the correlated min max for i in $G(p)$ is equal to the min max for II in $\bar{G}(p)$. Hence the result from Theorem 4.2. \square

5. The general case

We analyze equilibria of $G_\infty(p)$ with respect to simple strategies in which all exploration takes place during the first stages of repetition.

5.1. Scenarios

A scenario is a profile of strategies under which players first explore a new action combination in each stage, then this exploration process *stops*, and play forever (a convex combination of) the cells which have been uncovered in the exploration phase. We now formalize this intuitive notion.

We first define how players explore their payoffs. An *exploration rule* is a pair $e = (f, t)$ where:

- $f = (f_n)_n$ is a profile of pure strategies such that for every play $h = (k, a_1, \dots, a_n, \dots)$ and $n \leq |A|$, $f_n(h)$ is not in the set $\{a_1, \dots, a_{n-1}\}$.
- t is a stopping time¹ $t: (H_\infty, \mathcal{H}_\infty) \rightarrow \{2, \dots, |A| + 1\}$.

f describes the order in which cells are explored, whereas $t - 1 \leq |A|$ is the last stage at which exploration takes place. The condition on t ensures that the decision whether to stop or not at stage n depends only on their information at stage n . Note that the definition of f matters only up to stage $|A|$ since $t \leq |A| + 1$.

An exploration rule e together with a state of nature k induce a history $(k, a_1, a_2, \dots, a_{t-1})$ during the exploration phase, which can be completed to a play $e(k) = (k, a_1, a_2, \dots, a_{t-1}, a_{t-1}, \dots, a_{t-1}, \dots) \in H_\infty$. This defines a map $e: K \rightarrow (H_\infty, \mathcal{H}_\infty)$. We let $\pi_{f,t} = e^{-1}(\mathcal{H}_\infty)$ be the coarsest σ -algebra on K for which this map is measurable. Two states of K are in the same atom of $\pi_{f,t}$ if and only if the histories they induce during the exploration with e are indistinguishable. Therefore, $\pi_{f,t}$ represents players' partition of information on K at time t if f has been followed. It is also useful to consider the set $A_k(e) = \{a_1, a_2, \dots, a_{t-1}\}$ of cells explored in state k with e .

A *scenario* (e, δ) is defined by an exploration rule e and by a measurable mapping $\delta: (K, \pi_{f,t}) \rightarrow \Delta(A)$ such that if k induces the history $(k, a_1, a_2, \dots, a_{t-1})$ during the exploration phase, $\text{supp}(\delta(k)) \subset \{a_1, \dots, a_{t-1}\}$.

In state k , $\delta(k)$ is to be thought of as the distribution of player's action profiles after exploration stops, and $\langle \delta, g \rangle(k) = \mathbf{E}_{\delta(k)} g_k(a)$ as the average payoff profile in the long run. We view $\langle \delta, g \rangle$ as a random variable on $(H_\infty, \mathcal{H}_\infty)$. The conditions on δ ensures that (1) $\delta(k)$ is known to the players at the end of the exploration phase and (2) after stage t , players keep playing cells already discovered.

The σ -algebra of events before t is denoted by \mathcal{H}_t . It is formally given by the set of $B \in \mathcal{H}_\infty$ such that for all n , $B \cap \{t \leq n\} \in \mathcal{H}_n$.

A scenario naturally defines strategies in $G_\infty(p)$ in which players follow f up to stage $t - 1$, then play pure actions with frequencies given by $\delta(k)$. For these strategies to form equilibria, one needs to impose some individual rationality condition. Hence we define:

¹ I.e., $\{t \leq n\} \in \mathcal{H}_n$ for every n .

Definition 5.1. A scenario (f, t, δ) is called *admissible* if

$$\langle \delta, g \rangle \geq \mathbf{E}_p(v \mid \pi_{f,t}) \quad p \text{ almost surely.}$$

In an admissible scenario each player receives at least the expectation of his min max conditional to his information after the exploration phase.

In terms of payoffs, $A(p)$ represents the subset of $\mathbb{R}^{I,K}$ induced by admissible scenarios:

$$A(p) = \{(\langle \delta, g \rangle(k))_k \in \mathbb{R}^{I,K} \text{ for some admissible scenario } (f, t, \delta)\}.$$

When the min max level v^i is replaced by the correlated min max level w^i in the definition of an admissible scenario, the corresponding set of induced payoffs is denoted by $B(p)$.

5.2. Statement of the results

Our main result is the following characterization of equilibrium payoffs of $G_\infty(p)$ in terms of $A(p)$:

Theorem 5.1. $\prod_k E_k \subseteq A(p) \subseteq E(p) \subseteq \text{co}(A(p)) = E_{\text{pub}}^*(p) = E_{\text{pub}}(p),$

$$E_{\text{cor}}^*(p) = E_{\text{com}}(p) = \text{co}(B(p)).$$

The notation “co” stands for the convex hull. In the last section we provide examples showing that each of the inclusions can be strict. Going from normal form to extensive form and from correlation devices to communication mechanisms, one increases the set of communication possibilities which are open to the players and the corresponding set of equilibrium payoffs. Therefore, Theorem 5.1 implies:

$$E_{\text{cor}}^*(p) = E_{\text{com}}^*(p) = E_{\text{cor}}(p) = E_{\text{com}}(p) = \text{co}(B(p)).$$

We present some remarks dealing with possible extensions of our results. In some situations of economic interest, it is more natural to assume that only *own* payoffs are observable. In that case, $E(p)$ may be empty, as shown by Koren (1988).

Nevertheless, the monitoring assumption may be weakened to allow for *symmetric* information functions. We may assume that in any stage, the players receive a public signal which includes the action profile. With the exception of the first inclusion, the result in Theorem 5.1 still holds, *modulo* an obvious adaptation of $\pi_{f,t}$ in the definition of an admissible scenario. The first inclusion needs not hold: it relies on the possibility of identifying the true game being played by exploration. If the public signal is always uninformative, $G(p)$ is equivalent to the average game, with payoff function $\sum_k p_k g(k, \cdot)$.

Finally, all the results would still hold for Banach equilibrium payoffs, i.e., if a Banach limit was used to define average payoffs (Hart, 1985).

6. Proofs

First, notice that after any stage, player's beliefs on k depend on the observed history and not on the strategies followed. More precisely, the probability of the true state of nature being k conditional to $h_n = (\tilde{k}, a_1, \dots, a_n) \in \mathcal{H}_n$ is:

$$p(k | \mathcal{H}_n)(h_n) = \begin{cases} \frac{p(k)}{p(\{k', \forall p < n \ g_{k'}(a_p) = g_{\tilde{k}}(a_p)\})} & \text{if } \forall p < n, \ g_k(a_p) = g_{\tilde{k}}(a_p), \\ 0 & \text{otherwise.} \end{cases}$$

We denote p_n this conditional probability, and view it as a random variable on $(H_\infty, \mathcal{H}_\infty)$.

This implies the following lemma that we shall use extensively (the proof is straightforward and omitted):

Lemma 6.1. *For any mapping f from K to \mathbb{R} , any profile of strategies σ and $n \geq 1$, $E_{\sigma,p}[f | \mathcal{H}_n] = \sum_k p_n(k) f(k)$, $P_{p,\sigma}$ -a.s.*

We can now prove the first inclusion of the main theorem:

Proposition 6.1. *One has $\prod_k E_k \subseteq A(p)$.*

Proof. Let $\gamma = (\gamma_k)_k \in \prod_k E_k$. Choose an enumeration of the possible action combinations, i.e., a bijective map from A to $\{1, \dots, |A|\}$, and define a profile $f \in \Sigma$ as: play in stage n the action profile labeled n , whatever be the information available.

Set $t = |A| + 1$, and $e = (f, t)$. For $k \in K$, choose $\delta(k) \in \Delta(A)$, such that $\langle \delta, g \rangle(k) = \gamma_k$. Under f , all the action combinations have been tested by stage $|A|$. Hence π_e is the discrete σ -algebra over K . Therefore, δ is π_e -measurable. On the other hand, $\gamma_k \in E_k$ implies $\gamma_k \geq v_k$. Thus, (e, δ) is an admissible scenario. \square

Proposition 6.2. *One has $A(p) \subseteq E(p)$.*

Proof. We give here the main ideas underlying the proof. A detailed proof can be found in Appendix B. Let $\gamma \in A(p)$, and (f, t, δ) an admissible scenario such that $\gamma = \langle \delta, g \rangle$. An equilibrium profile with payoff γ is described as follows.

On the equilibrium path, the play is divided into a learning phase and a payoff accumulation phase. In the learning phase, the players follow f , therefore discover the payoffs induced by some action combinations. This phase is ended at time t . From then on, the players play a specific sequence of elements of A , among those which have been *discovered* (i.e., *played*) prior to t . It is chosen so that the asymptotic frequency along this sequence of each $a \in A$ converges to $\delta(a)$. Of course, it has to depend on the realized state of nature. However, since δ is π_e -measurable, the sequences followed in the different states can be chosen in a π_e -measurable way: playing the correct sequence can be done using only the information available at t .

Any deviation from this equilibrium path is punished forever: if player i deviates, the coalition $-i$ switches to an optimal strategy in the corresponding zero-sum game (with symmetric incomplete information).

The fact that this constitutes indeed an equilibrium profile with payoff γ is derived from the following arguments.

In order to evaluate the impact of deviating after a given history $h_n \in \mathcal{H}_n$, player i has to compare his continuation payoff, i.e., the payoff he would get by not deviating, $\mathbf{E}_p[\langle \delta, g \rangle^i | h_n]$, to the level at which he would be punished, would he deviate at that stage. This punishment level is equal to $v^i(p_{n+1})$, where p_{n+1} is the posterior distribution over K , after the deviation has taken place. At h_n , the value of $v^i(p_{n+1})$ may be unknown, since it might be the case that a new action combination is tried at that stage (and it may depend upon the specific deviation from the equilibrium path). A crucial step is to show that the expected level of punishment $\mathbf{E}_{p, \sigma^{-i}, \tau^i}[v^i(p_{n+1}) | h_n]$ coincides in any case with $\mathbf{E}_p[v^i | h_n]$. This is easily deduced from a martingale argument and from the fact that $v(p) = \sum_k p_k v_k, \forall p$ (cf. the study of the zero-sum case).

Finally, the fact that $\mathbf{E}_p[\langle \delta, g \rangle^i | h_n] \geq \mathbf{E}_p[v^i | h_n]$ follows from the admissibility of the scenario (f, t, δ) . Therefore, the continuation payoff of player i always exceeds the payoff he would get in case of a deviation. \square

Proposition 6.3. $E(p) \subseteq \text{co } A(p)$.

Proof. Let $\gamma \in E(p)$, and σ be an uniform equilibrium profile associated to γ . The decomposition of γ as a convex combination of elements of $A(p)$ is obtained by interpreting σ as a *mixed* strategy, i.e., as a probability distribution over pure strategies, rather than as *behavioral* strategies.

Any profile of pure strategies induces a family of plays, one for each state of nature. On each of these plays, *experimentation* may occur at various stages, but must eventually end. For each play, delete *all* the stages prior to the last experimentation stage in which no experimentation takes place. One thereby obtains a new family of plays in which all the learning is done right at the beginning of the play. Therefore, we have associated an exploration rule to any profile of pure strategies. σ may thus be viewed as a probability distribution over the *finite* set of exploration rules.

We now construct payoffs. Let e be an exploration rule in the support of σ . For $n \geq 1$, it makes sense to compute the average payoff $x_n(e)$ up to stage n , conditional on the fact that the observed history is compatible with e (i.e., is consistent with the hypothesis that the profile of pure strategies selected by σ induces e).

There is no reason why the various sequences $(x_n^k(e))_{k \in K, e \in \text{supp } \sigma}$ should converge. However, since the number of states and exploration rules is finite, we may choose a subsequence $\phi(n)$ such that $x_n^k(e)$ converges, say to $x^k(e)$, for each $k \in K, e \in \text{supp } \sigma$.

If two states k and k' are not distinguished by e (that is, belong to the same atom of π_e), then no history consistent with e can distinguish between them. Thus, $x^k(e) = x^{k'}(e)$. On the other hand, if the true state happens to be k , then, on any history consistent with e , all the action combinations which are played belong to $A_k(e)$. Therefore, one can construct a π_e -measurable function $\delta_e : K \rightarrow \Delta(A)$, such that $\text{supp } \delta_e(k) \subseteq A_k(e)$, and $\langle \delta_e, g \rangle = x(e)$.

It is straightforward to check that $\gamma = \sum_e \sigma(e)x(e)$. To conclude the proof, it remains to be proved that, for each e in the support of σ , the scenario (e, δ_e) is admissible. This property is derived from the following two observations.

On the one hand, let h_n be an history of length n (atom of \mathcal{H}_n) with positive probability under σ . Then, for $\epsilon > 0$, the expected average payoff $\mathbf{E}_{p,\sigma}[\bar{g}_q \mid h_n]$ conditional on h_n is at least $\mathbf{E}_p[v \mid h_n] - \epsilon$, provided q is large enough. Indeed, if this were not true, say for player i , player i would find it profitable to deviate from stage n , if h_n occurred. This is ruled out since σ is an equilibrium profile.

On the other hand, provided n is large enough, the probability that the play fails at some stage to be consistent with e , given that it is consistent up to stage n , is close to 0 (otherwise, e would not be in the support of σ). Therefore, denoting by $H_n(e)$ the set of histories consistent with e up to n , the expected payoff $\mathbf{E}_{k,\sigma}[\bar{g}_q \mid H_n(e)]$ is close to $x^k(e)$, for each k .

The two observations yield an estimate of the kind

$$\mathbf{E}_{p,\sigma}[x(e) \mid H_n(e)] \geq \mathbf{E}_p[v \mid H_n(e)] - 2\epsilon.$$

The result follows by taking the limit n to infinity, using the fact that ϵ was arbitrary. \square

Proposition 6.4. $\text{co } B(p) = E_{\text{cor}}(p) = E_{\text{com}}(p)$.

Proof. We first prove that $\text{co } B(p) \subseteq E_{\text{cor}}(p)$. Let $\gamma \in \text{co } B(p)$. Write γ as a convex combination of payoffs in $B(p)$:

$$\gamma = \sum_{q=1}^Q \alpha_q \gamma_q, \quad \text{where } \alpha_q \geq 0, \gamma_q \in A(p) \text{ for each } q, \text{ and } \sum_{q=1}^Q \alpha_q = 1.$$

Extend $G_\infty(p)$ by the following public correlation mechanism which takes place in stage 0: $q \in \{1, \dots, Q\}$ is chosen according to the distribution $\alpha = (\alpha_1, \dots, \alpha_Q)$, and publicly announced.

If q happens to be chosen, players follow a profile defined as in the proof of Proposition 6.2, with the following modification. At each stage, a correlation device is available, which is used if some player, say player i , deviated from the equilibrium path: it enables players $-i$ to correlate their actions, in order to achieve the correlated min max level.

We do not provide a detailed proof of the inclusion $E_{\text{com}}(p) \subseteq \text{co } B(p)$. We only briefly explain how the proof of $E(p) \subseteq \text{co } A(p)$ can be adapted.

Let $\gamma \in E_{\text{com}}(p)$: γ is an equilibrium of $G_\infty(p)$, extended by some communication mechanism, which we denote by $G_\infty^c(p)$. Add one fictitious player which *controls* the communication mechanisms (whose strategy is to choose the outputs as a function of the inputs he gets). Let σ be a corresponding equilibrium profile (of course, the strategy of the fictitious player coincides with the description of the communication mechanisms). As in the proof of Proposition 6.3, σ is viewed as a probability distribution over profiles of *pure* strategies in $G_\infty^c(p)$. The crucial point is the following: any profile of pure strategies s in $G_\infty^c(p)$ can be *identified* to a profile of pure strategies \tilde{s} in $G_\infty(p)$: intuitively, every round of communication is useless since its result is known in advance (actually, is common knowledge). Slightly more formally, given any history \tilde{h}_n of length n in $G_\infty(p)$, each player is able to compute the vector of inputs which have been sent, according to s , in the previous stages, therefore also the outputs since the fictitious player is also using a pure

strategy. Thus, there is exactly one history h_n of length n in $G_\infty^c(p)$ which is consistent with h_n and s . Hence, it is meaningful to define \tilde{s} as: play after \tilde{h}_n what s would play after h_n . The rest of the proof is similar to the proof of Proposition 6.3. \square

The proof of the equality $\text{co} B(p) = E_{\text{cor}}^*(p) = E_{\text{com}}^*(p)$ is obtained along the same lines as the previous proposition, by setting all the correlation or communication devices used along the play *before* the beginning of the play.

The proofs of $\text{co} A(p) = E_{\text{pub}}^*(p) = E_{\text{pub}}(p)$ are similar. The use of correlated devices with public signals makes it impossible to a coalition of players to correlate themselves in a *private* way. Therefore, $B(p)$ is here to be replaced by $A(p)$. (If we did replace public *correlation* devices by public *communication* devices, private correlation would again be possible; we do not wish to elaborate on this point).

7. Comments

7.1. All inclusions of Theorem 5.1 may be strict

Example 2 ($E(p) \neq \text{co}(A(p))$). Consider the example of duopoly previously studied, and let (σ^1, σ^2) be a Nash equilibrium of $G(1/2, 1/2)$. Let $p^i(t)$ denote the probability that player i plays P at stage t if (P, P) has always been played before. If

$$p_\infty^i = \lim_{T \rightarrow \infty} \prod_{1 \leq t \leq T} p^i(t) = 0 \quad \text{for } i = 1 \text{ or } i = 2,$$

then war occurs with probability 1. The induced equilibrium payoff is $(2, -2)$ if $k = 1$ and $(-2, 2)$ if $k = 2$.

Now assume $p_\infty^i > 0$ for $i = 1, 2$. Player 1's incentives are to minimize the probability with which a war is declared, since after war is declared his expected payoff is 0 whereas if war is never declared his expected payoff is 1. Therefore it is a best reply for player 1 to play P until W has been played by 2, and his best reply in $G(k)$ after. This way, 1's expected payoff is $p_\infty^2 \times 1 + (1 - p_\infty^2) \times 0$. Therefore 1 never declares war before 2 does. Similarly 2 does not play W until 1 does. Thus, both players always play P , and the induced equilibrium payoff is $(1, 1)$ in both states.

Hence we have shown that

$$E(p) = \{(2, -2), (-2, 2)\} \cup \{(1, 1), (1, 1)\}$$

which is not a convex set.

Example 3 ($\prod E_k \neq A(p)$). In the previous duopoly game, one has $\prod E_k = ((2, -2), (-2, 2))$ since when k is known, there is only one equilibrium payoff. We define an exploration rule e by: examine cell (P, P) then stop. This exploration process is completed to a scenario with the distribution on cells which is a unit mass at (P, P) . This scenario is admissible since it yields to each player a payoff of 1 which is greater than the expected min max of 0. Yet it yields a payoff which is not element of $\prod E_k$.

Example 4 ($A(p) \neq E(p)$). Consider the following version $G'(p)$ of $G(p)$ in which strategy P has been duplicated. The initial probability is $p = (1/2, 1/2)$ on payoff matrices:

	W	P_1	P_2
W	2, -2	2, -2	2, -2
P_1	2, -2	1, 1	1, 1
P_2	2, -2	1, 1	1, 1

$G'(1)$

	W	P_1	P_2
W	-2, 2	-2, 2	-2, 2
P_1	-2, 2	1, 1	1, 1
P_2	-2, 2	1, 1	1, 1

$G'(2)$

The same arguments as before show that $A(p) = \{(2, -2), (-2, 2)\} \cup \{(0, 0), (0, 0)\}$. Now, we define strategies in $G'(p)$ in which both players:

- Stage 1. Play $(\frac{1}{2}P_1, \frac{1}{2}P_2)$.
- Stage $n \geq 2$. Play P_1 if (P_1, P_1) or (P_2, P_2) was played in stage 1. Otherwise play W .
- If some player played W instead of P_1 at any stage $n \geq 2$, play W from stage $n + 1$ on.

No player has incentives to deviate from (W, W) since it is a Nash equilibrium. As before, (P_1, P_1) is an equilibrium path if a deviation to W leads to an infinite repetition of (W, W) . Stage 1 is a jointly controlled lottery used to randomize between the two basic equilibria: Peace or War. Hence these strategies form a Nash equilibrium; it yields an equilibrium payoff of $((3/2, 1/2), (1/2, 3/2))$ which is not an element of $A(p)$.

7.2. The discounted case

We first deal with the zero-sum case. Define $v_\lambda^i(p)$ to be the min max value for player i of the λ -discounted game with incomplete information in which the initial distribution over states is p . Since the uniform min max $v(p)$ exists, $\lim_{\lambda \rightarrow 1} v_\lambda(p)$ exists, and is equal to $v(p)$. In particular, an application of Theorem 4.2 shows that $v_\lambda(p)$ is close to $\sum_k p_k v_\lambda(k)$, provided λ is close enough to 1.

Example 5. Consider the following two games, one of which is selected according to $p = (1/2, 1/2)$:

T	1, 0
B_1	1, 1
B_2	0, 0

$G''(1)$,

T	1, 0
B_1	0, 0
B_2	1, 1

$G''(2)$.

The action T always gives a payoff of 1 to player 1, so that player 1 can guarantee 1 in any (discounted or not) repetition of the game. Note also that $(1, 1)$ is an equilibrium payoff of both $G''(1)$ and $G''(2)$.

If payoffs are not discounted, player 1 can explore during the first stage, and play the action that leads to $(1, 1)$ at each consecutive stage. The payoff vector associated to this equilibrium is $((1, 1), (1, 1))$, which is consistent with the fact that $\prod_k E_k \subseteq E(p)$.

If payoffs are discounted, the only way for player 1 to get a payoff of 1 is to play T at each stage. Therefore, payoffs are not explored at an equilibrium. This shows that the inclusion $\prod_k E_k \subseteq E(p)$ does not hold if payoffs are discounted.

This last example, which is a maximization problem for a single agent, shows that the set of equilibrium payoffs $E_\lambda(p)$ where payoffs are discounted with discount factor λ may not converge to $E(p)$. This is in fact a classical phenomenon in the literature of repeated games with incomplete information.

In this example, there is no strictly individually rational payoff in E_k . It is well known that, in such circumstances, the Folk theorem may fail to hold, even for repeated games with complete information (Fudenberg and Maskin, 1986). One is therefore led to ask what happens in non-degenerate situations. We report some related results. Let $(e_k)_k$ be a feasible payoff vector ($e_k \in \text{co}\{g(k, A)\}$ for every k), such that $e_k^i > v_k^i$, for every player i . It is not difficult to adapt the proof of Theorem 5.1 to show that (e_k) is an equilibrium payoff in the λ -discounted game, provided λ is close enough to 1.

More generally, the inclusion $A(p) \subseteq E(p)$ can be adapted as follows. Let $e \in A(p)$. It is associated to an admissible scenario (f, t, δ) (see Definition 5.1). Assume that $e > \mathbf{E}_p(v \mid \pi_{f,t})$, p -a.s. As above, it is not difficult to show that e is an equilibrium payoff of the λ -discounted game, provided λ is close enough to 1.

7.3. Perfect equilibria

For any history h_n of length n let $p(h_n)$ be the conditional probability on K after h_n . We say that the strategy profile σ is a perfect (Bayesian) equilibrium of $G(p)$ if the continuation strategies $(\sigma_{h_n}^i)_i$ after every h_n form an equilibrium of $G(p(h_n))$. We denote by $E'(p)$ the set of perfect equilibrium payoffs profiles of $G(p)$.

Clearly $E'(p) \subseteq E(p)$. Next is an example where the inclusion is strict.

Example 6. Consider the following two games with probability $p = (1/2, 1/2)$:

$$\begin{array}{cc|cc} & W & P & & & \\ W & 2, -2 & 2, -2 & G'''(1), & W & -2, 0 & 1, 1 & G'''(2). \\ P & 2, -2 & 1, 1 & & P & -2, 0 & 1, 1 & \\ \hline & & & & & & & \end{array}$$

The strategies:

- play (P, P) if W has never been played before;
- play (W, W) otherwise.

Constitute a Nash equilibrium of $G'''(p)$ inducing payoff $((1, 1), (1, 1))$. The min max of $G'''(p)$ is $(0, -1/2)$ which is less than $(1, 1)$ for each player. Nevertheless, the threat of playing (W, W) in $G'''(2)$ is not credible since P is a dominant strategy for player 2 in this game. The only Nash payoff of $G'''(1)$ is $(2, -2)$ and the only Nash payoff of $G'''(2)$ is $(1, 1)$. Therefore, every perfect Bayesian equilibrium yields a payoff of at least $3/2$ to player 1. This implies that the probability that (P, P) is played forever is 0 in every perfect Bayesian equilibrium: the true state is uncovered, a.s. Hence the only subgame perfect equilibrium payoff of $G'''(p)$ is $((2, -2), (1, 1))$.

Here are some remarks on the structure of $E'(p)$.

- (i) Note that the perfect Folk theorem asserts that $E'_k = E_k$ for every k .

(ii) One can easily prove that $\prod_k E_k \subseteq E'(p)$ using the following fully revealing strategies (x_k denotes a fixed element of E_k).

EXPLORE. Play sequentially each combination of actions in A , thus revealing k .

PAYOFFS. Once k is revealed, play a subgame equilibrium of $G(k)$ implementing x_k .

PUNISHMENTS. If player i deviates from EXPLORE at stage n , play the punishing strategies defined in Section 4 for n stages, then start back EXPLORE.

Clearly, no player has incentives to deviate from PAYOFFS. By deviating in EXPLORE, player i is in the long run punished to his min max level in PUNISHMENT, which cannot be more than what he would obtain in PAYOFFS (recall that $x_k^i \geq v_k^i$). Note also that no deviation from PUNISHMENT can be profitable since each punishment is of finite length.

(iii) Let us define $z_k^i = \min\{c^i, x \in E_k^i\}$. This is the worst payoff for player i in a perfect Bayesian equilibrium payoff of G_k . Note that we may have $z_k^i > v_k^i$. Then, for $p \in \Delta(K)$ let $z_k^i(p) = \mathbf{E}_k z_k^i$. Say that a scenario is p -admissible when one replaces v^i by z^i in the definition of admissibility. We let $A'(p)$ represent the set of payoff profiles induced by p -admissible scenarios. One has $A'(p) \subseteq E'(p)$. The proof is similar to the one of $A(p) \subseteq E(p)$, except that one replaces the punishments of i by an equilibrium of the kind defined in (ii) in which i receives z_k^i , in state k .

(iv) Finally, do we have $E'(p) \subseteq \text{co}(A'(p))$? The answer is no, as shown by the next example.

Example 7. There are three players, player 3 has only one possible action, and the game is one of the following two with probability $p = (1/2, 1/2)$.

	W	P		W	P		
W	2, -2, 4	2, -2, 4	G''''(1),	W	-2, 2, 4	-2, 2, 4	G''''(2).
P	2, -2, 4	1, 1, 0		P	-2, 2, 4	1, 1, 0	

One has $z_k^3 = 4$ for each k , and thus $z_k^3(p) = 4$. However, $(1, 1, 0) \in E'(p)$.

Acknowledgments

The authors are grateful to Martin Cripps, Ivar Ekeland, Françoise Forges, Jean-François Mertens, and Rakesh Vohra for comments and stimulating discussions.

Appendix A. Zero-sum games

In this appendix, we give a detailed proof of Proposition 4.2. We assume w.l.o.g. in what follows that $\max_{a,i} |g^i(a)| \leq 1$.

To guarantee v_k . Let $\epsilon > 0$. We define below a profile σ_ϵ^{-i} and prove that it satisfies

$$\exists N, \forall \sigma^i, n \geq N, k \in K, \mathbf{E}_{k, \sigma_\epsilon^{-i}, \sigma^i} [\bar{g}_n] \leq v_k + \epsilon. \tag{3}$$

For $j \neq i$, denote by $e^j = (1/|A^j|, \dots, 1/|A^j|) \in \Delta(A^j)$ the uniformly mixed strategy of player j .

For each subset \tilde{A}^i of A^i and $k \in K$, choose an optimal profile $\sigma^{-i}(k, \tilde{A}^i)$ of players $-i$ in the (one-shot, complete information) game with payoff function g_k where player i is restricted to \tilde{A}^i . We may obviously assume that the two profiles $\sigma^{-i}(k, \tilde{A}^i)$ and $\sigma^{-i}(k', \tilde{A}^i)$ coincide if the restrictions of g_k and $g_{k'}$ to $\tilde{A}^i \times A^{-i}$ coincide.

For $n \in \mathbb{N}$, denote by $A^i(n)$ the set of actions $a^i \in A^i$ for which the function $g(a^i, \cdot)$ is known at the beginning of stage n . Notice that this is a set-valued process adapted to (\mathcal{H}_n) .

For $j \neq i$, define σ_ϵ^j as: play according to e^j if $A^i(n) = \emptyset$, and $(1 - \epsilon)\sigma^j(k, A^i(n)) + \eta e^j$ otherwise, where $\eta = \epsilon/(I + 1)$. Set $\sigma_\epsilon^{-i} = (\sigma_\epsilon^j)_{j \neq i}$.

Let σ^i be a pure strategy of player i and set $\sigma = (\sigma^i, \sigma_\epsilon^{-i})$ for notational convenience. For $a \in A$, $n \in \mathbb{N}$ denote by

$$H_n(a) = \{h \in H_\infty, \forall p < n, a_p \neq a\}$$

the set of plays on which a has not been played prior to stage n . Notice that $H_n(a) \in \mathcal{H}_n$. For $a^i \in A^i$ set $H_n(a^i) = \bigcup_{a^{-i} \in A^{-i}} H_n(a^i, a^{-i}) \in \mathcal{H}_n$: it consists of those histories of length $n - 1$, after which the payoff function $g(a^i, \cdot)$ is not yet fully known.

We denote by (t_p) the successive stages in which player i chooses an action which consequences are not fully known:

$$t_1 = 1, \quad t_{p+1}(h) = \inf\{n > t_p(h), h \in H_n(\sigma^i(h))\}, \quad p \geq 1.$$

Notice that (t_p) is a non-decreasing sequence of stopping times (possibly infinite) for the filtration (\mathcal{H}_n) .

In each of the stages t_p , the probability that a new cell is discovered is at least $(1/|A^{-i}|)\eta^{I-1}$. This implies that the sequence $(P_{k,\sigma}\{t_p < +\infty\})_p$ decreases exponentially fast to 0. This is the content of the next lemma.

Lemma A.1. $\forall q, P_{k,\sigma}\{t_{q+|A|} < +\infty \mid t_q < +\infty\} \leq 1 - \alpha$, where $\alpha = ((1/|A^{-i}|)\eta^{I-1})^{|A|}$.

Proof. For $n \in \mathbb{N}$, we denote by $N_n(h) = |\{a \in A, h \in H_n(a)\}|$ the number of action combinations which are unknown prior to stage n (i.e., which have not been previously played). Notice that $0 \leq N_n \leq |A|$, $\forall n$, and $N_{n+1} \leq N_n$. Also, N_n may only decrease in the stages t_p and $N_{t_p} > 0$ on $\{t_p < +\infty\}$. Moreover,

$$P_{k,\sigma}\{N_{t_p+1} = N_{t_p} - 1 \mid t_p < +\infty\} \geq \frac{1}{|A^{-i}|} \eta^{I-1}.$$

The result follows. \square

Clearly, one then has

$$P_{k,\sigma}\{t_{q|A|} < +\infty\} \leq (1 - \alpha)^{q-1},$$

for every $q \in \mathbb{N}$. Denote by $S = \max\{p, t_p < +\infty\}$ the number of stages in which player i plays an unknown action. We now prove that S is bounded in expectation.

Lemma A.2. $\mathbf{E}_{k,\sigma}[S] \leq |A|(1 + 1/(1 - \alpha))$.

Proof.

$$\begin{aligned} \mathbf{E}_{k,\sigma}[S] &= \sum_{q=1}^{\infty} P_{k,\sigma}\{S \geq q\} = \sum_{q=1}^{\infty} P_{k,\sigma}\{t_q < +\infty\} \leq |A| \left(1 + \sum_{q=1}^{\infty} P_{k,\sigma}\{t_{q|A|} < +\infty\} \right) \\ &\leq |A| \left(1 + \frac{1}{1 - \alpha} \right). \quad \square \end{aligned}$$

We are now in a position to prove that σ_ϵ^{-i} almost guarantees v_k in state k , for long games. Property (3) follows from the next result.

Lemma A.3. One has $\mathbf{E}_{k,\sigma}[\bar{g}_N] \leq v_k + I\eta + \frac{1}{N}\mathbf{E}_{k,\sigma}[S]$, for every $N \in \mathbb{N}$.

Proof. Let $n \in \mathbb{N}$. With probability at least $(1 - \eta)^{I-1} \geq 1 - I\eta$, players $-i$ follow in stage n the profile $\sigma^{-i}(A^i(n))$. In that case, if player i selects an action a^i within $A^i(n)$, the expected payoff to player i in stage n is at most v_k .

Denote by $\Omega_n = \bigcup_{q=1}^{\infty} \{t_q = n\} \in \mathcal{H}_n$ the set of those plays on which player i chooses an action outside $A^i(n)$ in stage n .

By the previous paragraph, one has

$$\mathbf{E}_{k,\sigma} [g_n \mathbb{1}_{\Omega_n^c}] \leq ((1 - I\eta)v_k + I\eta) P_{k,\sigma} \{\Omega_n^c\}.$$

Therefore,

$$\mathbf{E}_{k,\sigma} [g_n] \leq v_k + I\eta + P_{k,\sigma} \{\Omega_n\}.$$

By summation over n , one obtains

$$\mathbf{E}_{k,\sigma} [\bar{g}_N] \leq v_k + I\eta + \frac{1}{N} \sum_{n=1}^N P_{k,\sigma} \{\Omega_n\} \leq v_k + I\eta + \frac{1}{N} \mathbf{E}_{k,\sigma} [S]$$

where the second inequality uses Fubini's theorem. \square

To defend v_k . Let $\sigma^{-i} \in \Sigma^{-i}$, and $\epsilon > 0$. We construct $\sigma_\epsilon^i \in \Sigma^i$ and prove (see Lemma A.6) that

$$\forall k, \quad \mathbf{E}_{k,\sigma^{-i},\sigma_\epsilon^i} [\bar{g}_n] \geq v_k - \epsilon,$$

provided n is large enough.

Denote by (p_n) the process of posterior beliefs held by player i , knowing that players $-i$ use σ^{-i} .

Notice that the distribution of players $-i$'s actions in stage n , conditional on the information available to player i , is a correlated distribution, denoted by $\bar{\sigma}_n^{-i}$.

The strategy σ_ϵ^i is defined as: play according to $(1 - \epsilon)\sigma^i(p_n, \bar{\sigma}_n^{-i}) + \epsilon e^i$ in stage n , where $\sigma^i(p_n, \bar{\sigma}_n^{-i})$ is a best reply of player i to the correlated distribution $\bar{\sigma}_n^{-i}$ in the game with payoff function $\sum_k p_n(k)g_k$.

We prove that, whatever be the true state of nature k , playing σ_ϵ^i against σ^{-i} ensures that player i 's average payoffs eventually exceeds $v_k - \epsilon$.

As above $H_n(a) = \{h, \forall p < n, a_p \neq a\}$ is the set of histories up to stage n for which the content of cell a has not been discovered. We set $H_n(a^{-i}) = \bigcup_{a^i \in A^i} H_n(a^{-i}, a^i)$. Set $\eta = \epsilon/6$, and define

$$\Omega_n = \{h, \exists a^{-i} \in A^{-i}, h \in H_n(a^{-i}) \text{ and } \sigma_n^{-i}(h)[a^{-i}] \geq \eta\}.$$

$h \in \Omega_n$ is at stage n , there is a non-negligible probability that an *unknown* action is played by players $-i$. Notice that $\Omega_n \in \mathcal{H}_n$. Thus, on Ω_n , there is a probability at least $\beta = \eta\epsilon/|A^i|$ that a new cell is discovered at stage n .

We now state the analog of Lemma A.2. We redefine $S = \sum_{n=1}^{\infty} \mathbb{1}_{\Omega_n}$, and we set $\sigma = (\sigma^{-i}, \sigma_\epsilon^i)$.

Lemma A.4. Set $C = |A|(1 + 1/(1 - \beta^{|A|}))$. Then $\mathbf{E}_{k,\sigma} [S] \leq C$.

Proof. It is straightforward to adapt the proofs of Lemmas A.1 and A.2. \square

Let $n \in \mathbb{N}$. We say that the anticipation of player i in stage n is good if $\|\sigma_n^{-i}(h) - \bar{\sigma}_n^{-i}(h)\| \leq \eta$ (the real distribution on players $-i$'s move in stage n is quite close to the anticipated distribution). We otherwise say that the anticipation is bad. We denote by $\Theta_n = \{h, \|\sigma_n^{-i}(h) - \bar{\sigma}_n^{-i}(h)\| > \eta\} \in \mathcal{H}_n$ the corresponding set of histories. We denote by $B(h) = \{n, h \in \Theta_n\}$ the set of bad anticipations.

We rely on the following classical result from the literature on reputation effects. The reader is referred to (Fudenberg and Levine, 1992) or (Sorin, 1999) for a proof.

Lemma A.5 (Fudenberg and Levine, 1992). There exists $N_0 \in \mathbb{N}$, such that $P_{k,\sigma} \{|B| \geq N_0\} < \eta$.

We now compute an estimate on the average payoff in any stage $n \geq 1$. Let h_n be an history up to stage n included in $(\Omega_n \cup \Theta_n)^c$. After h_n , the anticipated distribution of players $-i$ actions is good, which implies that

$\sigma_n^i(h_n)$ is an 2η -best reply to the actual distribution $\sigma_{k,n}^{-i}(h_n)$. Moreover, the probability of an unknown action combination by players $-i$ is at most η . Therefore, any best reply of player i to $\sigma_{k,n}^{-i}(h_n)$ yields an expected payoff of at least $v_k - \eta$.

In conclusion, one has

$$\mathbf{E}_{k,\sigma}[g_n \mathbb{1}_{(\Omega_n \cup \Theta_n)^c}] \geq (v_k - 4\eta) P_{k,\sigma}\{(\Omega_n \cup \Theta_n)^c\}.$$

Therefore,

$$\mathbf{E}_{k,\sigma}[g_n] \geq v_k - 4\eta - (P_{k,\sigma}(\Omega_n) + P_{k,\sigma}(\Theta_n)). \quad (4)$$

Lemma A.6. *One has*

$$\mathbf{E}_{k,\sigma}[\bar{g}_n] \geq v_k - \left(4\eta + \frac{N_0}{N} + \eta + \frac{C}{N}\right).$$

Proof. Set $B_N = B \cap \{1, \dots, N\}$. By summation over n , one gets from (4)

$$\mathbf{E}_{k,\sigma}[\bar{g}_n] \geq v_k - \left(4\eta + \frac{1}{N} \mathbf{E}_{k,\sigma}[B_N] + \frac{1}{N} \mathbf{E}_{k,\sigma}[S]\right).$$

Now, $B_N \leq N$, and $P_{k,\sigma}\{B_N \geq N_0\} < \eta$. The result follows. \square

Appendix B. Non-zero-sum games

Proof of Proposition 6.2. For $k \in K$, choose a sequence $a^k = (a_n^k)_n$ in $A_k(e)$ such that the empirical frequency $\frac{1}{n} \sum_{p=1}^n \mathbb{1}_{a_p^k=a}$ of each $a \in A$ in the sequence converges to $\delta(k)[a]$. Moreover, we choose the sequences a^k so that the map $k \mapsto a^k$ is π_e -measurable. This is feasible, since δ is π_e -measurable.

We define a profile σ of pure strategies as follows. It coincides with f until t (learning phase). In other words, $\sigma_n^i = f_n^i$ on $\{t > n\}$. From t on, in state k , σ implements $(a_n^k)_n$ (payoff phase): $\sigma_n = (a_n^k)$ on $\{\bar{k} = k, t \leq n\}$ (where k is the random state of nature).

Denote by $d = \inf\{n, a_n \neq \sigma_n(k, a_1, \dots, a_{n-1})\}$ the first stage in which a player deviates from the main path. Notice that $d + 1$ is a stopping time for (\mathcal{H}_n) . If i is the deviating player, players $-i$ switch to *punishment path* i : they compute the posterior distribution p_{d+1} over K , given the information available at stage $d + 1$, and play optimal strategies in the corresponding game of incomplete information, where player i faces players $-i$.

Under σ , the main path is followed up to the end of the game. Given k , the players explore until t , and then follow the sequence a^k . Therefore, $\mathbf{E}_{k,\sigma}[\bar{g}_n] \rightarrow \gamma_k$, for each $k \in K$.

We now prove that no deviation of player i can improve upon σ^i . Let τ^i be a pure strategy of player i .

Our first statement compares conditional continuation payoffs to expected levels of individual rationality under σ .

Lemma B.1. $\forall n, \mathbf{E}_p[(\delta, g)^i | \mathcal{H}_n] \geq \mathbf{E}_p[v^i | \mathcal{H}_n], P_{p,\sigma}$ -a.s.

Proof. Notice that, $P_{p,\sigma}$ -a.s., the players learn nothing on k after t . Hence, for any $f: K \rightarrow \mathbb{N}$, and $n \in \mathbb{N}$,

$$\mathbf{E}_p[f | \mathcal{H}_n] = \mathbf{E}_p[f | \mathcal{H}_{\min\{n,t\}}], P_{p,\sigma}$$
-a.s. (5)

By assumption, $\mathbf{E}_p[(\delta, g)^i | \mathcal{H}_t] \geq \mathbf{E}_p[v^i | \mathcal{H}_t], P_{p,\sigma}$ -a.s. Conditioning with respect to $\mathcal{H}_{\min\{n,t\}}$ yields

$$\mathbf{E}_p[(\delta, g)^i | \mathcal{H}_{\min\{n,t\}}] \geq \mathbf{E}_p[v^i | \mathcal{H}_{\min\{n,t\}}].$$

The claim follows then from (5), used both for $(\delta, g)^i$ and v^i . \square

Lemma B.2. *One has*

$$\forall n \geq 1, \mathbf{E}_{p,\sigma^{-i},\tau^i}[v^i(p_{n+1}) | \mathcal{H}_n] = \mathbf{E}_p[v^i | \mathcal{H}_n].$$

Proof. From the study of zero-sum games, one has $v^i(p_{n+1}) = \mathbf{E}_p[v^i \mid \mathcal{H}_{n+1}]$, everywhere.

On the other hand, notice that $(\mathbf{E}_p[v^i \mid \mathcal{H}_n])_n$ is a $(H_\infty, (\mathcal{H}_n)_n, P_{p, \sigma^{-i}, \tau^i})$ -martingale. Therefore,

$$\mathbf{E}_{p, \sigma^{-i}, \tau^i}[v^i(p_{n+1}) \mid \mathcal{H}_n] = \mathbf{E}_{p, \sigma^{-i}, \tau^i}[\mathbf{E}_p[v^i \mid \mathcal{H}_{n+1}] \mid \mathcal{H}_n] = \mathbf{E}_p[v^i \mid \mathcal{H}_n]. \quad \square$$

It is easy now to derive the claim for Banach equilibria. Let \mathcal{L} be a Banach limit. Consider the paths induced by the two profiles σ and (σ^{-i}, τ^i) when the state of nature is k . If these two paths coincide, the payoffs induced by σ and (σ^{-i}, τ^i) are both equal to γ_k . If not, they differ in stage d and, from stage $d+1$ on, player i is punished. Therefore,

$$\gamma_{\mathcal{L}}^i(\sigma^{-i}, \tau^i) = \mathbf{E}_{p, \sigma^{-i}, \tau^i}[\gamma_k^i \mathbb{1}_{d=+\infty} + v^i(p_{d+1}) \mathbb{1}_{d<+\infty}].$$

Now,

$$\mathbf{E}_{p, \sigma^{-i}, \tau^i}[v^i(p_{d+1}) \mathbb{1}_{d<+\infty}] = \mathbf{E}_{p, \sigma^{-i}, \tau^i}[v^i(p_d) \mathbb{1}_{d<+\infty}] = \mathbf{E}_{p, \sigma}[v^i(p_d) \mathbb{1}_{d<+\infty}].$$

The first equality follows from Lemma B.2; the second from the fact that the paths induced by (σ^{-i}, τ^i) and σ coincide until d : $P_{p, \sigma} = P_{p, \sigma^{-i}, \tau^i}$ on $(H_\infty, \mathcal{H}_{\min\{d, n\}})$, for each n . From Lemma B.1, one has

$$v^i(p_{\min\{d, n\}}) \leq \mathbf{E}_p[\gamma_k^i \mid \mathcal{H}_{\min\{d, n\}}], \quad P_{p, \sigma}\text{-a.s.}$$

for each n . By taking expectations, and letting $n \rightarrow \infty$, one obtains

$$\mathbf{E}_{p, \sigma}[v^i(p_{d+1}) \mathbb{1}_{d<+\infty}] \geq \mathbf{E}_{p, \sigma}[\gamma_k^i(p_d) \mathbb{1}_{d<+\infty}],$$

hence $\gamma_{\mathcal{L}}^i(\sigma^{-i}, \tau^i) \leq \mathbf{E}_{p, \sigma}[\gamma_k^i] = \gamma_{\mathcal{L}}^i(\sigma)$.

Things are slightly more involved for uniform equilibrium. Fix some $n \in \mathbb{N}$, large compared to the time needed for a punishment to be effective, and to the time needed for average payoffs under σ to be close to γ . We only give the general idea of the computation. Details are standard and left to the reader.

Given k , either (σ^{-i}, τ^i) induces the same path up to n as σ , in which case the average payoff up to n , given k , are the same for the profiles. Or the two paths differ in stage d . The average payoff up to n is a convex combination of the average payoffs up to d and from $d+1$ up to n . The former coincides, (with the exception of stage d), with the average payoff up to d induced by σ . The latter corresponds to payoffs in the i -punishment phase.

If d is small compared to n , the weight of the first part is negligible, and the average payoff up to n is at most the expectation of v^i (up to some ε), given the information available at stage d . If d is close to n , the weight of the second part is negligible, and the average payoff up to n is close to γ_k . Otherwise, the average payoff to player i up to n is close to a convex combination of γ_k^i and of something which is at most the expected value of v^i , given the information at stage d . \square

Proof of Proposition 6.3. Let s be a profile of pure strategies. Given k , s induces a single path $(k, (a_n(k))_n)$. We denote by $\{\bar{a}_1, \dots, \bar{a}_{N_s}\}(k)$ the different action combinations which appear in this path, listed according to the order of appearance. Formally,

$$t_1(k) = 1, \quad \bar{a}_1(k) = a_1, \\ t_{p+1}(k) = \inf\{n > t_p, a_n \notin \{\bar{a}_1(k), \dots, \bar{a}_p(k)\}\}, \quad \bar{a}_{p+1}(k) = a_{t_{p+1}}(k), \quad \text{for } p \geq 1.$$

Choose a profile $f_s = (f_{s, n})_{n \geq 1}$ of pure strategies such that

$$f_{s, 1}(k) = \bar{a}_1(k) \quad \text{and} \quad f_{s, n+1}(k, \bar{a}_1(k), \dots, \bar{a}_n(k)) = \bar{a}_{n+1}(k), \quad (6)$$

for $n < t(k)$. This condition is compatible with the informational requirements: since $s \in \Sigma$, $\bar{a}_{n+1}(k)$ depends on k only through the payoffs of the action combinations played before, i.e., $\bar{a}_1(k), \dots, \bar{a}_n(k)$. Notice also that there are many exploration processes compatible with (6).

We say that $e_s = (f_s, N_s)$ is the exploration rule induced by s . Since σ may be viewed as a probability distribution over the profiles of pure strategies, it may also be viewed as a probability distribution over the set of exploration rules. We then denote by \mathcal{S} its support.

For $e \in \mathcal{S}$, we denote by $C(e) = \{s, e_s = e\}$ the set of profiles of pure strategies which induce e . For $n \geq 1$, and $s \in C(e)$, the set $\{h = (k, (a_p)_{p \geq 1}, s_p(h) = a_p, \forall p < n) \in \mathcal{H}_n$ is the event: *at stage n , the past play is consistent with the hypothesis that players are using s* . Therefore,

$$H_n(e) = \bigcup_{s \in C(e)} \{h, s_p(h) = a_p, \forall p < n\}$$

is the set of plays h compatible with e up to n . Notice that $H_n(e) \in \mathcal{H}_n$. For $k \in K$, define $x_n^k(e) = \mathbf{E}_{k,\sigma}[\bar{g}_n | H_n(e)]$: it is the average payoff up to n in state k , conditional upon the information being coherent with e . Set $x_n = (x_n^k(e))_{k \in K, e \in \mathcal{S}}$.

Since K and \mathcal{S} are finite, we may choose a convergent subsequence of (x_n) . For notational convenience, we still denote by (x_n) this subsequence, and we set $x = \lim_{n \rightarrow \infty} x_n$.

In the next three lemmas, $e \in \mathcal{S}$ is fixed.

Lemma B.3. *The map $k \mapsto x^k(e)$ is π_e -measurable.*

Proof. For any two states k, k' , the behaviors of the players in these states are identical until one of them is ruled out by the observations. Therefore, if k, k' belong to the same atom of π_e , no history in $H_n(e)$ distinguishes between them: the two distributions $P_{k,\sigma}\{\cdot | H_n(e)\}$ and $P_{k',\sigma}\{\cdot | H_n(e)\}$ coincide. Therefore $x_n^k(e) = x_n^{k'}(e)$ for every n . Taking the limit gives $x^k = x^{k'}$. \square

Lemma B.4. $x^k(e) \in \text{co}\{g_k(a), a \in A_k(e)\}$.

Proof. In state k , on $H_n(e)$, the only action combinations which can possibly appear are the elements of $A_k(e)$. Thus, $\mathbf{E}_{k,\sigma}[g_p | H_n(e)] \in \text{co}\{g_k(a), a \in A_k(e)\}$, for each $p \leq n$. This implies $x_n^k(e) \in \text{co}\{g_k(a), a \in A_k(e)\}$. \square

If k and k' belong to the same atom of π_e , $x^k(e) = x^{k'}(e)$, $A_k(e) = A_{k'}(e)$, and $g_k(a) = g_{k'}(a)$, for every $a \in A_k(e)$. Therefore, one can construct a π_e -measurable map $\delta_e : K \rightarrow \Delta(A)$, such that

$$\begin{cases} \langle \delta_e, g \rangle = x(e), \\ \text{Supp } \delta_e(k) \subset A_k(e), \quad \forall k. \end{cases}$$

Lemma B.5. (e, δ_e) is an admissible scenario.

Proof. By construction, it is a scenario. We prove that it is admissible. Notice that $(H_n(e))_n$ is a decreasing sequence of subsets of H_∞ . Since $e \in \mathcal{S}$, $P_{k,\sigma}\{\bigcap_n H_n(e)\} > 0, \forall k$. In particular, for every $\epsilon > 0$, there exists $N \in \mathbb{N}$, such that, if $q \geq n \geq N$, one has

$$\forall k, \quad P_{k,\sigma}\{H_n(e) \setminus H_q(e)\} < \epsilon. \quad (7)$$

It is straightforward to derive from (7) that, if X is an \mathcal{H}_∞ -measurable random variable with values in $[-1, 1]$,

$$|\mathbf{E}_{p,\sigma}[X | H_n(e)] - \mathbf{E}_{p,\sigma}[X | H_q(e)]| < 2\epsilon. \quad (8)$$

On the other hand, since σ is a uniform equilibrium profile, one has, for q large enough (depending on ϵ),

$$\mathbf{E}_{p,\sigma}\left[\frac{1}{q-n+1} \sum_{l=n}^q g_l \mid H_n(e)\right] \geq \mathbf{E}_{p,\sigma}[v | H_n(e)] - \epsilon. \quad (9)$$

From (8) and (9), one deduces that, for q large enough,

$$\mathbf{E}_{p,\sigma}[\bar{g}_q | H_q(e)] \geq \mathbf{E}_{p,\sigma}[v | H_q(e)] - 3\epsilon,$$

i.e., $\bar{x}_q(e) \geq \mathbf{E}_{p,\sigma}[v | H_q(e)] - 2\epsilon$. The result follows by taking the limit $q \rightarrow \infty$, using the fact that ϵ is arbitrary. \square

Therefore, $x(e) \in A(p)$, for every $e \in \mathcal{S}$. Thus, Proposition 6.3 follows from the next lemma.

Lemma B.6. $\gamma = \sum_S \sigma(e)x(e)$.

Proof. One has $\gamma_n(k, \sigma) = \mathbf{E}_{k, \sigma}[\bar{g}_n]$. However, one cannot write $\gamma_n(k, \sigma) = \sum_{e \in \mathcal{S}} \mathbf{E}_{k, \sigma}[\bar{g}_n \mathbb{1}_{H_n(e)}]$: the sets $(H_n(e))_{e \in \mathcal{S}}$ may overlap, hence do not constitute a partition of H_∞ ; a given atom of \mathcal{H}_n may be consistent with several exploration rules in \mathcal{S} .

Yet, set $H(e) = \bigcap_n H_n(e)$, for $e \in \mathcal{S}$. $(H(e))_{e \in \mathcal{S}}$ is a (finite) partition of H_∞ . Moreover, $\sigma(e) = P_{k, \sigma}(H(e))$, for each $k \in K$. Therefore,

$$\gamma_n(k, \sigma) = \sum_{e \in \mathcal{S}} \mathbf{E}_{k, \sigma}[\bar{g}_n \mathbb{1}_{H(e)}] = \sum_{e \in \mathcal{S}} \sigma(e) \mathbf{E}_{k, \sigma}[\bar{g}_n | H(e)].$$

Since $x_n^k(e) = \mathbf{E}_{k, \sigma}[\bar{g}_n | H_n(e)] \xrightarrow{n \rightarrow \infty} x^k(e)$, one has $\mathbf{E}_{k, \sigma}[\bar{g}_n | H(e)] \rightarrow x^k(e)$. This yields the result. \square

References

- Aumann, R.J., Hart, S., 1986. Bi-convexity and bi-martingales. *Israel J. Math.* 54, 159–180.
- Aumann, R.J., Maschler, M.B., 1995. With the collaboration of Stearns R.E., *Repeated Games with Incomplete Information*. MIT, Cambridge.
- Aumann, R.J., 1974. Subjectivity and correlation in randomized strategies. *J. Math. Econ.* 1, 67–95.
- Baños, A., 1968. On pseudo-games. *Ann. Math. Statist.* 39, 1932–1945.
- Forges, F., 1982. Infinitely repeated games of symmetric information: symmetric case with random signals. *Int. J. Game Theory* 11, 203–213.
- Forges, F., 1992. Repeated games of incomplete information: non-zero sum. In: Aumann, R.J., Hart, S. (Eds.). In: *Handbook of Game Theory*, Vol. 1. Elsevier, pp. 155–177, Chapter 6.
- Foster, D., Vohra, R., 1999. Regret in the on-line decision problem. *Games Econ. Behav.* 29, 7–35.
- Fudenberg, D., Levine, D.K., 1992. Maintaining a reputation when strategies are imperfectly observed. *Rev. Econ. Studies* 59, 561–579.
- Fudenberg, D., Maskin, E., 1986. The Folk theorem in repeated games with discounting and with incomplete information. *Econometrica* 54, 533–554.
- Hannan, J., 1957. Approximation to Bayes risk in repeated plays. In: Dresher, M., Tucker, A.W., Wolfe, P. (Eds.), *Contributions to the Theory of Games*, Vol. 3. Princeton Univ. Press, pp. 97–139.
- Hart, S., 1985. Nonzero-sum two-person repeated games with incomplete information. *Math. Oper. Res.* 10, 117–153.
- Hirshleifer, J., 1971. The private and social value of information and the reward to inventive activity. *Amer. Econ. Rev.* 61, 561–574.
- Kohlberg, E., Zamir, S., 1974. Repeated games of incomplete information: the symmetric case. *Ann. Statist.* 2, 1010–1041.
- Koren, G., 1988. Two-person repeated games with incomplete information and observable payoffs. MSc thesis. Tel-Aviv University.
- Mertens, J.-F., Sorin, S., Zamir, S., 1994. Repeated games. CORE discussion paper 9420-9422.
- Neyman, A., Sorin, S., 1998. Equilibria in repeated games of incomplete information: the general symmetric case. *Int. J. Game Theory* 27, 201–210.
- Sorin, S., 1999. Merging, reputation and repeated games with incomplete information. *Games Econ. Behav.* 29, 274–308.