CORE DISCUSSION PAPER
9823

STRATEGIC LEARNING IN GAMES WITH SYMMETRIC INFORMATION[1]

Olivier Gossner[2] and Nicolas Vieille[3]

April, 1998

Abstract

This paper 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 opponent's. We chararacterize equilibrium payoffs in terms of simple strategies in which an exploration phase is followed by a payoff acquisition phase.

KEYWORDS: public value of information, games with incomplete information, bandit problems

J.E.L. CLASSIFICATION: C72

---

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

2 CORE, 34 Voie du roman pays, Louvain-la-Neuve, B-1349. E-mail: gossner@core.ucl.ac.be

3 CEREMADE, Université Paris Dauphine, place du Maréchal de Lattre de Tassigny, FR-75016 Paris. E-mail: vieille@ceremade.dauphine.fr

This text presents research results of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The scientific responsability is assumed by the author.

1 Introduction

This paper 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 opponent's. Such a broad description includes many real life situations one may think of. We first motivate our study with a military example and an economic example.

An attacking general contemplating a strategy or a weapon that was never used before would need to try it on the field in order to know its effects. On the other side, the defender also needs to experiment to find out his best reply against the attacker's new strategy. It has been established by historians that the English victory at Crecy in 1346 was a consequence of the superiority of English longbows over the French crossbows. Such an superiority was probably not certain before the battle, has been proved at Crecy, and since then longbows began to replace crossbows in European armies.

In a world with imperfect competition, there is no reason to believe that the demand function of an opening market (seen as a function of all prices for similar goods) is known by the managers of firms. One expects managers to form beliefs on the true demand function and to try different price strategies before setting their prices optimally.

These two examples share some features on which we rely to construct our model. First, it can be assumed that the actions chosen by all agents are publicly observed. These actions are the battle plans in the first case, and the prices of goods on the market for the second example. Second, the outcomes are also publicly observed. Each general can assess the losses on the two sides, and the quantities sold can be considered as public information.

We study a model of repeated games with symmetric incomplete information in which after each turn, both the action profile played and the corresponding payoff profile are publicly announced. The two examples above can be viewed as instances of the zero-sum and of the non-zero sum cases.

The particular case of two players and zero-sum has been previously studied by Baños [4] and Megiddo [13]. They prove 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 first extend 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.

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 of the payoff functions only. Hirshleifer [11] already pointed out that public information can be socially damaging. We analyze the public value of information in a strategic model where information disclosure is a consequence of agent's actions and characterize how collective learning takes place.

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 tested during the exploration phase. 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.

It can be useful to place our work in light of the literatures of repeated games with incomplete information and multi-armed bandits, since these are also concerned with the question of learning through the repetition of a situation.

The theory of two-player repeated games with incomplete information (see [2], [8] for the general theory) usually assumes 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 [10] 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 [1] 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.

Particular 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 [12] 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 turns suffices for the whole process of information transmission. Yet, equilibria can fail to exist.

Failure of existence of equilibria can be seen as a consequence of asymmetry of information among players. Forges [7] for the zero-sum case and Neyman and Sorin [15] for the general case prove that equilibria always exists whenever information is symmetric across players. Their proofs rely on the identification of public information to a state variable in a stochastic game.

With one agent, multi-armed bandits models (see [5] for a general presentation) study the optimal allocation of time between learning and payoff optimization conditional to past information. The agent's payoff at each turn depends both on his action at that time, on some unknown state of nature, and on some extra random factor. As time goes by, the agent learns which distribution of payoffs is associated to each of his possible actions.

Building a generalization to any number of players, Bolton and Harris [6] consider a group of individuals who simultaneously face the same multi-armed bandit problem. Each agent observes other's actions and payoffs, so that one can learn from other's experiments. Bolton and Harris show that public information is a public good to which one contributes by exploring payoffs and they study the corresponding free-rider problem.

As opposed to Bolton and Harris's framework, in our model each agent's payoff depends on other's choices as well as on his. On the other hand, we assume that the same action profile yields the same payoff at each turn, and we do not discount the payoffs. These assumptions allow us to study the information acquisition separately from payoff acquisition. In a model with discounting, each stage has a strictly positive weight on the total payoff and the order in which cells are explored becomes important. Clearly, bandit problems with no discounting would be trivial since full information on the payoff distributions could be acquired at no cost. All the difficulty of bandit problems comes from the trade off between information acquisition and payoff maximization conditional to past information. Here, we show that even if players can spend time to explore payoffs at no cost, it can be socially efficient to stop this learning process before full information is obtained.

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 scenarii 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.

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 shall prove the existence of equilibria in which the true profile of payoff functions is revealed, more precisely:

Proposition 1 *Given an equilibrium payoff $x(k)$ for each game $G_\infty(k)$, there exists an equilibrium payoff of $G_\infty(p)$ in which players receive $x(k)$ when the state of nature is k.*

Proposition 1 characterizes a subset of equilibrium payoffs of $G_\infty(p)$ in which players get equilibrium payoffs of the real underlying infinitely repeated game. In such equilibria, all players completely learn their payoff functions. The following example shows that it may be better for all players not to discover the value of k, therefore leaving some uncertainty on the true underlying game.

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 will also win 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:

	W	P
W	2,-2	2,-2
P	2,-2	1,1

G_1

	W	P
W	-2,2	-2,2
P	-2,2	1,1

G_2

Players assess initial probability $p = (\frac{1}{2}, \frac{1}{2})$ 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 $G_k = ((A^i)_{i \in I}, g_k : A \to I\!R^I)$ (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 $(a_p)_p$ 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 $H_\infty = K \times A^{I\!N}$ 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 $h = (h, (a_p)_{p\geq 1}), h' = (k', (a'_p)_{p\geq 1})$. 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 $\mathcal{H}_\infty = \sigma(\cup_n \mathcal{H}_n)$; it is the coarsest σ-algebra over H_∞ which contains each $\mathcal{H}_n$.

A (behavior) strategy of player i is a sequence $\sigma^i = (f_n^i)_{n\geq 1}$, where f_n^i is a measurable map $f_n^i : (H_\infty, \mathcal{H}_n) \to \Delta(A^i)$ which describes the behavior of player i in stage n. The space of 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 $\mathcal{H}_\infty$ conditional on $k \in K$. Note that $P_{k,\sigma} = P_{\delta_k,\sigma}$ where δ_k is the Dirac mass on K and $P_{p,\sigma} = \sum_k p(k) P_{k,\sigma}$. For any $\mathcal{H}_\infty$-measurable bounded random variable X we write $E_{p,\sigma}X$ and $E_{k,\sigma}X$ 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 $\sigma \in \Sigma$, $\gamma_n(\sigma) = E_{p,\sigma}\{\frac{1}{n}\sum_{m=1}^n g_m\}$ is the expected average payoff up to stage n. Also $\gamma_n(k,\sigma) = E_{k,\sigma}\{\frac{1}{n}\sum_{m=1}^n g_m\}$ is the average payoff in state k.

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 [14] the notion of uniform equilibrium.

Definition 1 *A profile $\sigma \in \Sigma$ is an uniform equilibrium profile if the following two conditions are satisfied:*

1. *for every $k \in K$, $\gamma(k,\sigma) = \lim_{n\to\infty} \gamma_n(k,\sigma)$ 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.

The uniform equilibrium notion does not allow to associate unambiguously a payoff vector to any strategy profile. For this purpose we may use a Banach limit $\mathcal{L}$ on $\mathbb{R}^I$. We denote $G_{\mathcal{L}}(p)$ the game with strategy sets Σ^i and payoff function $\gamma_{\mathcal{L}}(\sigma) = \mathcal{L}((\gamma_n(\sigma))_n)$. A Banach equilibrium payoff of $G(p)$ is defined as an equilibrium payoff of $G_{\mathcal{L}}(p)$ for some Banach limit $\mathcal{L}$ (see [10]). Note that all uniform equilibria are also Banach equilibria since they are equilibria of $G_{\mathcal{L}}(p)$ for any Banach limit $\mathcal{L}$.

We shall denote by $E(p)$ the set of equilibrium payoffs of $G(p)$, either with the uniform equilibrium notion or with the Banach equilibrium notion. Our results will stand for both equilibrium notions and the two (similar) proofs will be provided.

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.

Given a Banach limit $\mathcal{L}$, the min max for player i in $G_{\mathcal{L}}(p)$ is defined as:

$$v^i(p) = \min_{\sigma^{-i}\in\Sigma^{-i}} \max_{\sigma^i\in\Sigma^i} \gamma^i_{\mathcal{L}}(\sigma^{-i},\sigma^i)$$

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

Although they are distinct notions, we keep the same notation for the uniform min max and the Banach min max.

If $\Gamma(p)$ 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:

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

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 [8]).

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 $E_{\mathrm{com}}(p)$. 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 $E_{\mathrm{com}}^*(p)$. 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 [3]). 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

We will rely extensively on the following characterization of min max levels which is a consequence of results due to Baños [4] and Megiddo [13] for the particular case of two players and that we extend to the N players case.

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

$$v^i(p) = \mathrm{E}_p v^i_k = \sum_k p_k v^i_k$$

Similarly, we characterize the correlated min max:

Corollary 3 *The correlated* min max *for player i in $G_\infty(p)$ exists and:*

$$w^i(p) = \mathrm{E}_p w^i_k = \sum_k p_k w^i_k$$

The preceding results are powerful tools that show that the two min max for i in $G(p)$ 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 zer-sum games.

These properties are deeply related to the observability of payoffs, and hardly to the assumption of symmetric information. In order to emphasize this point, we shall prove more than the statement of Theorem 2, and consider situations of asymmetric information. We shall prove that:

1. if player i is fully informed of k, while players $-i$ are not even informed of p, they can guarantee v^i_k in every state k;

2. if each player of the coalition $-i$ is fully informed of k, while player i is told only p, he can still defend v_k^i in every state k.

Proof of Theorem 2: We provide here only the intuition of the proof. For a detailed proof, the reader is referred to Annex A. We shall prove the claim for player i and will, for notational convenience, suppress any reference to i in the payoffs.

To guarantee v_k

We construct $\sigma^{-i} \in \Sigma^{-i}$ such that,

$$\forall \epsilon,\ \exists N_\epsilon,\ \forall \sigma^i,\ \forall n \geq N,\ \forall k \quad E_{k,\sigma}[\overline{g}_n] \leq v_k + \epsilon. \tag{1}$$

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 can not 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 $\forall \epsilon > 0$:

$$\exists N_\epsilon,\ \forall n \geq N_\epsilon, k \in K, \quad E_{k,\sigma^{-i},\sigma^i}[\overline{g}_n] \geq v_k - \epsilon. \tag{2}$$

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 $\overline{\sigma}_n^{-i}$ 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 $(1-\epsilon)\sigma^i(p_n, \overline{\sigma}_n^{-i}) + \epsilon e^i$ in stage n, where $\sigma^i(p_n, \overline{\sigma}_n^{-i})$ is a best reply of player i to the correlated distribution $\overline{\sigma}_n^{-i}$ 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 $-i$ will 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 [9] 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 $-i$ selects 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.

Proof of Corollary 3: Consider the two players game $\overline{G}(p)$ 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 $\overline{G}(p)$. Hence the result from Theorem 2.

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 Scenarii

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 $h = (k, a_1, \ldots, a_n, \ldots)$ and $n \leq |A|$, $f_n(h)$ is not in the set $\{a_1, \ldots, a_{n-1}\}$.
- t is a stopping time[4] $t : (H_\infty, \mathcal{H}_\infty) \to \{2, \ldots, |A| + 1\}$.

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 $(k, a_1, a_2, \ldots, a_{t-1})$ during the exploration phase, which can be completed to a play $e(k) = (k, a_1, a_2, \ldots, a_{t-1}, a_{t-1}, \ldots, a_{t-1}, \ldots) \in H_\infty$. This defines a map $e : K \to (H_\infty, \mathcal{H}_\infty)$. 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 undistinguishable. Therefore, $\pi_{f,t}$ represents players' partition of information on K at time t if f has been followed. It will also be useful to consider the set $A_k(e) = \{a_1, a_2, \ldots, a_{t-1}\}$ of cells explored in state k with e.

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

In state k, $\delta(k)$ is to be thought of as the distribution of player's action profiles after exploration stops, and $< \delta, g > (k) = \mathrm{E}_{\delta(k)} g_k(a)$ as the average payoff profile in the long run. We view $< \delta, g >$ 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:

[4]Recall that t being a stopping time requires the measurability condition $\{t \leq n\} \in \mathcal{H}_n$ for every n.

Definition 2 *A scenario* (f,t,δ) *is called* **admissible** *if*

$$< \delta, g > \geq \mathrm{E}_p(v|\pi_{f,t}) \qquad p \text{ almost surely}$$

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 scenarii:

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

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 will be denoted $B(p)$ instead of $A(p)$.

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 4

$$\prod_k E_k \subseteq A(p) \subseteq E(p) \subseteq \mathrm{co}(A(p)) = E^*_{pub}(p) = E_{pub}(p)$$

$$E^*_{cor}(p) = E_{com}(p) = \mathrm{co}(B(p))$$

Remark: The notation "co" stands for the convex hull.
Remark: Proposition 1 can be rewritten $\prod_k E_k \subseteq E(p)$, and is therefore a consequence of Theorem 4.
Remark: We shall prove these results both for Banach equilibria and for uniform equilibria.
Remark: In the last section we provide examples showing that each of the inclusions can be strict.
Remark: 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 4 implies:

$$E^*_{cor}(p) = E^*_{com}(p) = E_{cor}(p) = E_{com}(p) = \mathrm{co}(B(p))$$

Remark: the extension of our proofs to countable sets K is straightforward; dealing with arbitrary sets K would create measurability issues that we wish to avoid here.

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 $h_n = (\tilde{k}, a_1, \ldots, a_n) \in \mathcal{H}_n$ 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 omited):

Lemma 5 *For any mapping f from K to $\mathbb{R}$, any profile of strategies σ and $n \geq 1$, $\mathrm{E}_{\sigma,p}[f|\mathcal{H}_n] = \sum_k p_n(k) f(k)$, $P_{p,\sigma}$-a.s.*

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

Proposition 6 *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 $\{1, \ldots, |A|\}$, 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 $< \delta, g > (k) = \gamma_k$. 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. ∎

Proposition 7 *One has $A(p) \subseteq E(p)$.*

PROOF: We give here the main ideas underlying the proof. A detailed proof can be found in Annex B. Let $\gamma \in A(p)$, and (f, t, δ) an admissible scenario such that $\gamma =< \delta, g >$. 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, $E_p[< \delta, g >^i |h_n]$, 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 $E_{p,\sigma^{-i},\tau^i}[v^i(p_{n+1})|h_n]$ coincides in any case with $E_p[v^i|h_n]$. 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 $E_p[< \delta, g >^i |h_n] \geq E_p[v^i|h_n]$ 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. ∎

Proposition 8 *$E(p) \subseteq coA(p)$.*

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 $x_{\phi(n)}^k(e)$ converges, to $x^k(e)$, for each $k \in K, e \in$ Supp σ.

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 will 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 $\delta_e : K \to \Delta(A)$, such that Supp $\delta(k) \subseteq A_k(e)$, and $< \delta, g >= x(e)$.

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 $E_{p,\sigma}[\overline{g}_q | h_n]$ conditional on h_n is at least $E_p[v|h_n] - \epsilon$, 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 will at some stage fail 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 $E_{k,\sigma}[\overline{g}_q|H_n(e)]$ is close to $x^k(e)$, for each k.

The two observations yield an estimate of the kind

$$E_{p,\sigma}[x(e)|H_n(e)] \geq E_p[v|H_n(e)] - 2\epsilon.$$

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

Proposition 9 *$coB(p) = E_{cor}(p) = E_{com}(p)$.*

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

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

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

If q happens to be chosen, players follow a profile defined as in the proof of Proposition 7, 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 will not provide a detailed proof of the inclusion $E_{com}(p) \subseteq \text{co}B(p)$. We shall only briefly sketch how the proof of $E(p) \subseteq \text{co}A(p)$ can be adapted.

Let $\gamma \in E_{com}(p)$: γ is an equilibrium of $G_\infty(p)$, extended by some communication mechanism, which we denote by $G^c_\infty(p)$. 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 8, σ is viewed as a probability distribution over profiles of *pure* strategies in $G^c_\infty(p)$. The crucial point is the following: any profile of pure strategies s in $G^c_\infty(p)$ 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 $G^c_\infty(p)$ 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 8.

■

The proof of the equality $\mathrm{co}B(p) = E^*_{cor}(p) = E^*_{com}(p)$ 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 $\mathrm{co}A(p) = E^*_{pub}(p) = E_{pub}(p)$ 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 4 may be strict

Example 2: $[E(p) \neq \mathrm{co}(A(p))]$

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

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

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

Now assume $p^i_\infty > 0$ 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 $p^2_\infty \times 1 + (1 - p^2_\infty) \times 0$. 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 Dirac 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 $p = (\frac{1}{2}, \frac{1}{2})$ 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 $((\frac{3}{2}, \frac{1}{2}), (\frac{1}{2}, \frac{3}{2}))$ which is not an element of $A(p)$.

7.2 The discounted case

Example 5: Consider the following two games with probability $p = (\frac{1}{2}, \frac{1}{2})$:

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.

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.

7.3 Perfect equilibria

Player's (Bayesian) beliefs on the state of nature are well defined after any history of the game. Note however that $G(p)$ has no subgames except $G(k)$ which occurs when all information on the payoffs is revealed. The next example shows that subgame perfect equilibrium payoffs can form a strict subset of Nash equilibrium payoffs.

Example 6: Consider the following two games with probability $p = (\frac{1}{2}, \frac{1}{2})$:

	W	P
W	2,-2	2,-2
P	2,-2	1,1

$G'''(1)$

	W	P
W	-2,0	1,1
P	-2,0	1,1

$G'''(2)$

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 $((1,1),(1,1))$. Actually, the expected min max of $G'''(p)$ is $(0,-1)$ which is less than $(1,1)$ for each player. Nevertheless, the threat of playing (W, W) in $G'''(1)$ is not credible since P is a dominant strategy for player 2 in this game. The only Nash payoff of $G'''(1)$ is $(1,1)$ and the only Nash payoff of $G(2)$ is $(2,-2)$. This implies that the only subgame perfect equilibrium payoff of $G'''(p)$ is $((1,1),(2,-2))$.

References

[1] R. J. Aumann and S. Hart. Bi-convexity and bi-martingales. *Israel Journal of Mathematics*, 54:159–180, 1986.

[2] R. J. Aumann, M. B. Maschler, with the collaboration of R. E. Stearns. *Repeated games with incomplete information*. MIT Press, Cambridge, 1995.

[3] R. J. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1:67–95, 1974.

[4] A. Baños. On pseudo-games. *The Annals of Mathematical Statistics*, 39:1932–1945, 1968.

[5] D. A. Berry and B. Fristedt. *Bandit problems. Sequential allocation of experiments*. Chapman and Hall, London, 1985.

[6] P. Bolton and C. Harris. Strategic experimentation. *to appear in Econometrica*, 1997.

[7] F. Forges. Infinitely repeated games of symmetric information: symmetric case with random signals. *International Journal of Game Theory*, 11:203–213, 1982.

[8] F. Forges. Repeated games of incomplete information: non-zero sum. In R.J. Aumann and S. Hart, editors, *Handbook of Game Theory*, volume 1, chapter 6, pages 155–177. Elsevier Science Publishers, 1992.

[9] D. Fudenberg and D. K. Levine. Maintaining a reputation when strategies are imperfectly observed. *Review of Economic Studies*, 59:561–579, 1992.

[10] S. Hart. Nonzero-sum two-person repeated games with incomplete information. *Mathematics of Operations Research*, 10:117–153, 1985.

[11] J. Hirshleifer. The private and social value of information and the reward to inventive activity. *American Ecomomic Review*, 61:561–574, 1971.

[12] G. Koren. Two-person repeated games with incomplete information and observable payoffs. M.Sc. thesis, Tel-Aviv University, 1988.

[13] N. Megiddo. On repeated games with incomplete information played by non-bayesian players. *International Journal of Game Theory*, 9(3):157–167, 1980.

[14] J.-F. Mertens, S. Sorin, and S Zamir. Repeated games. CORE discussion paper 9420-9422, 1994.

[15] A. Neyman and S. Sorin. Equilibria in repeated games of incomplete information: the general symmetric case. Discussion paper 428, Laboratoire d'Économétrie de l'École Polytechnique, 1995.

[16] S. Sorin. Merging, reputation and repeated games with incomplete information. Discussion paper 467, Laboratoire d'Économétrie de l'École Polytechnique, 1997.

A Zero-sum games

In this annex, we give a detailed proof of proposition 2.

To guarantee v_k

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

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

For $j \neq i$, denote by $e^j = (\frac{1}{|A^j|}, \ldots, \frac{1}{|A^j|}) \in \Delta(A^j)$ 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 $n \in I\!N$, denote by $A^i(n)$ the set of actions $a^i \in A^i$ for which the function $g(., a^i)$ 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 $\eta = \frac{\epsilon}{I+1}$. 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 $a \in A$, $n \in I\!N$ 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 $H_n(a^i) = \cup_{a^{-i} \in A^{-i}} H_n(a^i, a^{-i}) \in \mathcal{H}_n$: it consists of those histories of length $n-1$, after which the payoff function $g(a^i, .)$ 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:

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

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 $\frac{1}{|A^{-i}|}\eta^{I-1}$. 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 10 *$\forall q, P_{k,\sigma}\{t_{q+|A|} < +\infty | t_q < +\infty\} \leq 1-\alpha$, where $\alpha = (\frac{1}{|A^{-i}|}\eta^{I-1})^{|A|}$.*

PROOF: for $n \in I\!N$, 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,

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

The result follows. ∎

Clearly, one then has

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

for every $q \in I\!N$. 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 11 $\mathrm{E}_{k,\sigma}[S] \leq |A|(1 + \frac{1}{1-\alpha})$.

PROOF:

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

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

Lemma 12 *One has $\mathrm{E}_{k,\sigma}[\overline{g}_N| \leq v_k + I\eta + \frac{1}{N}\mathrm{E}_{k,\sigma}[S]$, for every $N \in I\!N$.*

PROOF: let $n \in I\!N$. 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 $\Omega_n = \cup_{q=1}^{\infty}\{t_q = n\} \in \mathcal{H}_n$ 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

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

Therefore,

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

By summation over n, one obtains

$$\begin{aligned} \mathrm{E}_{k,\sigma}[\overline{g}_N] &\leq v_k + I\eta + \frac{1}{N}\sum_{n=1}^{N} P_{k,\sigma}\{\Omega_n\} \\ &\leq v_k + I\eta + \frac{1}{N}\mathrm{E}_{k,\sigma}[S] \end{aligned}$$

where the second inequality uses Fubini's theorem. ∎

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 15) that

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

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 $\overline{\sigma}_n^{-i}$.

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

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

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 $H_n(a^{-i}) = \cup_{a^i \in A^i} H_n(a^{-i}, a^i)$. Set $\eta = \frac{\epsilon}{6}$, 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 will be played by players $-i$. Notice that $\Omega_n \in \mathcal{H}_n$. Thus, on Ω_n, there is a probability at least $\beta = \frac{\eta\epsilon}{|A^i|}$ that a new cell is discovered at stage n.

We now state the analog of Lemma 11. We redefine $S = \sum_{n=1}^{\infty} 1_{\Omega_n}$, and we set $\sigma = (\sigma^{-i}, \sigma_\epsilon^i)$.

Lemma 13 *Set* $C = |A|(1 + \frac{1}{1-\beta^{|A|}})$. *Then* $\mathrm{E}_{k,\sigma}[S] \leq C$.

PROOF: it is straightforward to adapt the proofs of Lemmas 10 and 11. ∎

Let $n \in I\!N$. We say that the anticipation of player i in stage n is good if $\|\sigma_n^{-i}(h) - \overline{\sigma}_n^{-i}(h)\| \leq \eta$ (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 $\Theta_n = \{h, \|\sigma_n^{-i}(h) - \overline{\sigma}_n^{-i}(h)\| > \eta\} \in \mathcal{H}_n$ the corresponding set of histories. We denote by $B(h) = \{n, h \in \Theta_n\}$ the set of bad anticipations.

We shall rely on the following classical result from the literature on reputation effects. The reader is referred to [9] or [16] for a proof.

Lemma 14 (Fudenberg and Levine, 1992) *There exists* $N_0 \in I\!N$, *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 $-i$-s 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

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

Therefore,

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

Lemma 15 *One has*

$$\mathrm{E}_{k,\sigma}[\overline{g}_N] \geq v_k - (4\eta + \frac{N_0}{N} + \eta + \frac{C}{N}).$$

PROOF: set $B_N = B \cap \{1, \ldots, N\}$. By summation over n, one gets from (4)

$$\mathrm{E}_{k,\sigma}[\overline{g}_N] \geq v_k - (4\eta + \frac{1}{N}\mathrm{E}_{k,\sigma}[B_N] + \frac{1}{N}\mathrm{E}_{k,\sigma}[S]).$$

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

B non zero-sum games

PROOF OF PROPOSITION 7:

For $k \in K$, choose a sequence $a^k = (a_n^k)_n$ in $A_k(e)$ such that the empirical frequency $\frac{1}{n}\sum_{p=1}^{n} 1_{a_p^k = a}$ 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 $\{\tilde{k} = k, t \leq n\}$ (where $\tilde{k}$ is the random state of nature).

Denote by $d = \inf\{n, a_n \neq \sigma_n(k, a_1, \ldots, a_{n-1})\}$ 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, $\mathrm{E}_{k,\sigma}[\overline{g}_n] \to \gamma_k$, 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 16 $\forall n$, $\mathrm{E}_p[< \delta, g >^i | \mathcal{H}_n] \geq \mathrm{E}_p[v^i | \mathcal{H}_n]$, $P_{p,\sigma}$-*a.s.*

PROOF: notice that, $P_{p,\sigma}$-a.s., the players learn nothing on k after t. Hence, for any $f : K \to I\!R$, and $n \in I\!N$,

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

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

$$\mathrm{E}_p[< \delta, g >^i | \mathcal{H}_{\min\{n,t\}}] \geq \mathrm{E}_p[v^i|\mathcal{H}_{\min\{n,t\}}].$$

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

Lemma 17 *One has*

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

PROOF: from the study of zero-sum games, one has $v^i(p_{n+1}) = \mathrm{E}_p[v^i|\mathcal{H}_{n+1}]$, everywhere.

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

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

∎

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,

$$\gamma^i_{\mathcal{L}}(\sigma^{-i}, \tau^i) = \mathrm{E}_{p,\sigma^{-i},\tau^i}[\gamma^i_{\tilde{k}} 1_{d=+\infty} + v^i(p_{d+1}) 1_{d<+\infty}].$$

Now,

$$\begin{aligned} \mathrm{E}_{p,\sigma^{-i},\tau^i}[v^i(p_{d+1})1_{d<+\infty}] &= \mathrm{E}_{p,\sigma^{-i},\tau^i}[v^i(p_d)1_{d<+\infty}] \\ &= \mathrm{E}_{p,\sigma}[v^i(p_d)1_{d<+\infty}] \end{aligned}$$

The first equality follows from Lemma 17; 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 16, one has

$$v^i(p_{\min\{d,N\}}) \leq \mathrm{E}_p[\gamma^i_{\tilde{k}} | \mathcal{H}_{\min\{d,n\}}],\ P_{p,\sigma} - a.s.$$

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

$$\mathrm{E}_{p,\sigma}[v^i(p_{d+1})1_{d<+\infty}] \leq \mathrm{E}_{p,\sigma}[\gamma^i_k 1_{d<+\infty}],$$

hence $\gamma^i_{\mathcal{L}}(\sigma^{-i}, \tau^i) \leq \mathrm{E}_{p,\sigma}[\gamma^i_k] = \gamma^i_{\mathcal{L}}(\sigma)$.

Things are slightly more involved for uniform equilibrium. Fix some $n \in I\!N$, 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^i_k and of something which is at most the expected value of v^i, given the information at stage d. ∎

Proof of Proposition 8: let s be a profile of pure strategies. Given k, s induces a single path $(k, (a_n(k))_n$. We denote by $\{\overline{a}_1, \ldots, \overline{a}_{N_s}\}(k)$ the different action combinations which appear in this path, listed according to the order of appearance. Formally,

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

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

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

for $n < t(k)$. This condition is compatible with the informational requirements: since $s \in \Sigma$, $\overline{a}_{n+1}(k)$ depends on k only through the payoffs of the

action combinations played before, *i.e.* $\overline{a}_1(k), \ldots, \overline{a}_n(k)$. 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,

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

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 $x_n^k(e) = \mathrm{E}_{k,\sigma}[\overline{g}_n | H_n(e)]$: 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 $x = \lim_{n\to\infty} x_n$.

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

Lemma 18 *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 $H_n(e)$ will distinguish between them: the two distributions $P_{k,\sigma}\{.\,|H_n(e)\}$ and $P_{k',\sigma}\{.\,|H_n(e)\}$ coincide. Therefore $x_n^k(e) = x_n^{k'}(e)$, for every n. Taking the limit gives $x^k = x^{k'}$. ∎

Lemma 19 *$x^k(e) \in co\{g_k(a), a \in A_k(e)\}$.*

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

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 $\delta_e : K \to \Delta(A)$, such that

$$\begin{cases} < \delta_e, g >= x(e) \\ \text{Supp } \delta_e(k) \subset A_k(e), \forall k \end{cases}$$

Lemma 20 *(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 $e \in \mathcal{S}$, $P_{k,\sigma}\{\cap_n H_n(e)\} > 0$, $\forall k$. In particular, for every $\epsilon > 0$, there exists $N \in I\!N$, such that, if $q \geq n \geq N$, one has

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

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

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

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

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

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

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

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

Therefore, $x(e) \in A(p)$, for every $e \in \mathcal{S}$. Thus, Proposition 8 follows from the next lemma.

Lemma 21 $\gamma = \sum_{\mathcal{S}} \sigma(e)x(e)$.

PROOF: one has $\gamma_n(k, \sigma) = \mathrm{E}_{k,\sigma}[\overline{g}_n]$. However, one can not write $\gamma_n(k, \sigma) = \sum_{e\in\mathcal{S}} \mathrm{E}_{k,\sigma}[\overline{g}_n 1_{H_n(e)}]$: 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 $H(e) = \cap_n H_n(e)$, 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,

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

Since $x^k_n(e) = \mathrm{E}_{k,\sigma}[\overline{g}_n|H_n(e)] \to_{n\to\infty} x^k(e)$, one has $\mathrm{E}_{k,\sigma}[\overline{g}_n|H(e)] \to x^k(e)$. This yields the result. ∎