Stochastic Games with Imperfect Monitoring

Dinah Rosenberg[*], Eilon Solan[†] and Nicolas Vieille[‡] [§]

July 23, 2003

Abstract

This chapter provides an introductory exposition of stochastic games with imperfect monitoring. These are stochastic games in which the players imperfectly observe the play. We discuss at length a few basic issues, and describe selected contributions.

Our objective in this chapter is to provide an introductory exposition of some recent work on zero-sum stochastic games with imperfect monitoring. We will try to avoid many of the technical subtleties inherent to this type of work, by discussing at length some fundamental issues, before we proceed to introduce the basic insights of the known results. This introduction briefly recalls historical developments of the theory, discussed more extensively later, and describes the organization of the chapter.

Stochastic games are played in stages. At every stage $n \in \mathbf{N}$ the players are to play one matrix game, taken from a finite set of possible games, called *states*. The matrix game played at stage n depends on the actions that were played at stage $n-1$ and on the previous state. In the present chapter, we limit ourselves to zero-sum games, *i.e.,* to the case where each component matrix game is a (two-player) zero-sum game. *Imperfect monitoring* refers to a situation where past moves of a player are imperfectly observed by his/her opponent, as opposed to perfect monitoring. Most work on stochastic games assumes perfect monitoring.

[*]Laboratoire d'Analyse Géométrie et Applications, Institut Galilée, Université Paris Nord, avenue Jean-Baptiste Clément, 93430 Villetaneuse, France. e-mail: dinah@zeus.math.univ-paris13.fr

[†]MEDS Department, Kellogg School of Management, Northwestern University, *and* the School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel. e-mail: eilons@post.tau.ac.il, e-solan@kellogg.northwestern.edu

[‡]Département Finance et Economie, HEC, 1, rue de la Libération, 78351 Jouy-en-Josas, France. e-mail: vieille@hec.fr

[§]We acknowledge the financial support of the Arc-en-Ciel/Keshet program for 2001/2002. The research of the second author was supported by the Israel Science Foundation (grant No. 69/01-1).

Stochastic games were introduced in a seminal paper of Shapley (1953). Shapley introduced discounted games in which each player i uses a discounted evaluation, that is, he wishes to maximize the discounted sum $\lambda \sum_{n=1}^{\infty}(1-\lambda)^{n-1} r_n^i$, where $\lambda \in (0,1)$ is the common discount factor, and r_n^i is the payoff to player i at stage n. He proved that any λ-discounted zero-sum stochastic game with perfect monitoring has a value v_λ. In addition, he proved that each player has an optimal strategy that is stationary: it depends only on the current state, and not on past history. Blackwell (1962) analyzed one-player stochastic games, better known as Markov decision process, or stochastic dynamic programming problems. For such games, Blackwell proved that there is a stationary strategy that is optimal for every discount factor λ sufficiently close to zero. This robustness, or uniformity, result was extended by Mertens and Neyman (1981) to the class of zero-sum stochastic games with perfect monitoring. Specifically, Mertens and Neyman proved that, given any $\varepsilon > 0$, each player has a strategy that is ε-optimal in the λ-discounted game, for *every* $\lambda < \lambda(\varepsilon)$. Thus, a single strategy is approximately optimal, whatever be the discount factor being used, provided it is sufficiently small. However, by contrast to Shapley's and Blackwell's results, in general this strategy cannot be taken to be stationary.

The consequences of imperfect monitoring have been widely explored within the framework of repeated games, see, e.g., Radner (1981), Rubinstein and Yaari (1983) and Lehrer (1989, 1990, 1992a, 1992b). Most of the interest focused on trying to provide a characterization of the set of equilibrium payoffs.

Stochastic games with imperfect monitoring were first analyzed in a series of papers by Coulomb (1992, 1999, 2001). In these papers, Coulomb analyzes absorbing stochastic games. These are stochastic games in which the state changes at most once along the play. Coulomb provides an example where the value does not exist, and proves that in this class of games the max-min and min-max values always exist (see below for definitions). This existence result was recently extended to all zero-sum stochastic games with imperfect monitoring independently by Coulomb (2003) and Rosenberg et *al.* (2003). Flesch et *al.* (2000) showed that a slight amount of imperfect monitoring in non-zero-sum games can prevent the existence of equilibrium payoffs.

This chapter is organized as follows. In Section 1, we discuss a number of classical examples, in order to highlight a few fundamental issues related to imperfect monitoring. In particular, we will argue that the issue of imperfect monitoring is irrelevant both for zero-sum repeated games and for λ-discounted stochastic games. Alternatively, it is relevant only for zero-sum stochastic games, in connection with the uniformity property mentioned above. In addition, we illustrate on two examples the consequences of imperfect monitoring. The discussion in this section is mainly kept at a heuristic level. In Section 2 we will be more specific in introducing a formal

model and in stating existence results. Section 3 contains a discussion of the proofs. It first summarizes the main insights of the proof of Mertens and Neyman (1981). It then explains how those insights are used in the analysis of games with imperfect monitoring. We conclude by discussing related work.

1 Basic observations and examples

1.1 Repeated Matching Pennies

We start with one of the simplest games, *Matching Pennies*. A version of the strategic form of this game is given by the table

	L	R
T	0	1
B	1	0

in which player 1 and player 2 are respectively the row and column players, and whose entries contain the payoff paid by player 2 to player 1.

The value of the one-shot Matching Pennies is $1/2$. Each player has a unique optimal strategy, which is mixed and assigns probability $1/2$ to both actions.

Suppose that the game is repeated over time. Consider the strategy σ of player 1 that tosses a fair coin at each stage, independent of previous tosses, and that plays T or B depending on the outcome. Let τ be any strategy of player 2. Such a strategy specifies, for each $n \in \mathbf{N}$, the mixed move (that is to say a probability distribution over the set $\{L, R\}$) to be used at stage n, as a function of all the information available at stage n. For each given n, the (conditional) distribution of player 1's move at stage n, given the information known to player 2, assigns probability $1/2$ to each action. Therefore, the (conditional) expected payoff at stage n under (σ, τ), given player 2's information, is equal to $1/2$. By averaging over all possible information sets of player 2 at stage n, this implies that the expected payoff under (σ, τ) at each stage n is $1/2$.

As a consequence, the strategy σ guarantees that the (expected) payoff to player 1 in the repeated game is exactly $1/2$, whatever be the weights assigned to the different stages. A similar analysis holds true for player 2.

In other words, the value of the infinitely repeated Matching Pennies is the same as that of the one-shot Matching Pennies. Moreover, the strategy in the repeated game that repeats an optimal strategy of the one-shot game is optimal. Plainly, these conclusions are not specific to the Matching Pennies

game, and hold more generally for every repeated two-player zero-sum game. Note that the preceding observation holds, whatever be the information about past play that is available to the players. Therefore, the nature of monitoring – perfect or imperfect – is irrelevant for the analysis of repeated zero-sum games.

To conclude this example, it may be helpful to realize why the conclusions are dramatically different for non-zero-sum repeated games. Let G be a given strategic form game, that is repeated over time. Generalizing upon the above observation, the strategy profile that consists of repeating over time a given equilibrium profile x of the game G is a Nash equilibrium of the repeated game, whose payoff coincides with the payoff induced by x in G. By contrast to zero-sum games, this need not be the unique equilibrium payoff of the repeated game. When it comes to a characterization of the equilibrium payoffs in the repeated game, the nature of monitoring is crucial. Typical proofs of the so-called *Folk Theorem* (see, e.g., Sorin, 1990. or Aumann and Shapley, 1994), proceed along the following lines: given a payoff vector, a play path is identified that induces this payoff. A strategy profile is next designed, under which the players are required to follow the play path, and to "punish" reciprocally in case of deviations from this path. Clearly, whether or not this profile is an equilibrium depends on the extent to which deviations are observed and deviators identified. A complete characterization of equilibrium payoffs is not yet available. A solution has been provided by Lehrer (1989,1990,1992a,1992b) for various notions of undiscounted equilibrium and/or under specific assumptions on the monitoring structure. Only partial results have been established for discounted games, see, *e.g.*, the January 2002 special issue of the Journal of Economic Theory, and the references therein.

1.2 The Big Match

We here recall well-known results on the Big Match game, an example of an absorbing stochastic game introduced in Gillette (1957) and later analyzed by Blackwell and Ferguson (1968). Although the formal model of stochastic games has yet to be introduced, this example will clarify why the issue of monitoring is irrelevant for the analysis of the discounted games, but not if one seeks to establish optimality properties which are uniform w.r.t. the discount factor.

The Big Match is described by the table

	L	R
T	0^*	1^*
B	1	0

Both players have two actions. As long as player 1 plays the Bottom row,

his payoff is given by the above table. As soon as he plays the Top row, say at stage θ, the payoff to player 1 at this stage and *all* subsequent stages is 0 or 1, depending on whether player 2 played the Left or the Right column at stage θ. Equivalently, at stage θ, the play moves to one of two possible trivial (absorbing) states, in which the payoff is constant.

As proven in Shapley (1953), the value v_λ of the λ-discounted game satisfies a dynamic programming principle. Indeed, v_λ is uniquely characterized as the value of the following one-shot zero-sum game $\Gamma(v_\lambda)$:

	L	R
T	0	1
B	$\lambda + (1-\lambda)v_\lambda$	$(1-\lambda)v_\lambda$

Moreover, let x_λ denote an optimal mixed move of player 1 in the game $\Gamma(v_\lambda)$. Then the stationary strategy that plays x_λ at every stage is an optimal strategy in the λ-discounted stochastic game.[1]

Again, this property is not specific to the Big Match. More generally, the existence of stationary optimal strategies in λ-discounted games ensures that the λ-discounted value v_λ is independent of the type of monitoring, as long as both players are always informed of the current state. Hence the issue of monitoring is irrelevant for the analysis of λ-discounted games.

We now discuss the existence of strategies that are ε-optimal in all λ-discounted games, provided λ is close enough to zero. We shall discuss only the problem faced by player 1, since the stationary strategy of player 2 that assigns probability 1/2 to both actions is optimal for each $\lambda > 0$. Assume first that the assumption of perfect monitoring holds, that is, at each stage n, player 1 knows the complete sequence of actions selected by player 2 up to stage n.

Blackwell and Ferguson (1968) devised a parametric family $(\sigma_N)_{N\in\mathbf{N}}$ of strategies. For $n \in \mathbf{N}$, define e_n to be the excess number of stages up to n in which player 2 selected the Left column: $e_n = l_n - r_n$, where l_n and r_n are respectively the number of stages up to stage n in which player 2 played the Left and the Right columns. The strategy σ_N plays the Top row with probability $\frac{1}{(N+e_n)^2}$ (assuming player 1 played the Bottom row in all previous stages). Then there is a constant λ_N such that for every strategy τ, the λ-discounted payoff induced by σ_N and τ is at least $\frac{N-2}{2N}$, provided $\lambda < \lambda_N$; for a proof of this result see Blackwell and Ferguson (1968) and Coulomb (1996).

[1] One can check that for the Big Match, $v_\lambda = 1/2$, player 1 has a unique optimal strategy that assigns probability $\lambda/(1+\lambda)$ to the Top row, and player 2 has a unique optimal strategy that assigns probability 1/2 to each column.

The intuition behind this strategy is as follows. Suppose that $e_n > 0$, so that so far player 2 played the Left column more often than the Right column. In this case player 1 does not want the game to terminate: as long as the frequency of Left is higher, his payoff by playing the Bottom row is more than 1/2. Therefore, player 1 decreases the probability of playing the Top row. If, on the other hand, $e_n < 0$, player 1 would like the game to terminate, so he increases the probability of playing the Top row. The effect of any given stage on the probability of playing the Top row is small, so that any strategic manipulation of future behavior of player 1 by player 2 comes at the cost of being absorbed to a bad payoff while the manipulation takes place. However, this effect is large enough, so that if $\liminf_{n\to\infty}(e_n) < 0$, player 1 will eventually play the Top row.

We postpone the discussion of the strategies devised by Mertens and Neyman (1981) to Section 3.1. In a nut-shell, according to Mertens and Neyman, at every stage n player 1 plays a stationary λ_n-discounted strategy x_{λ_n}, where λ_n is defined recursively as a function of λ_{n-1} and of the choice of player 2 in stage $n-1$.

To contrast with the full monitoring case, we now assume that player 1 receives no information about past moves of player 2. Since the game stops at the first stage θ in which player 1 chooses the Top row, a strategy of player 1 reduces here to a sequence $\mathbf{x} = (x_n)_{n\in\mathbf{N}}$, with the interpretation that x_n is the probability assigned to the Top row at stage n, assuming $\theta \geq n$.

Let such a strategy $\mathbf{x}$ be given. We shall now check that, given $\varepsilon > 0$, there is a reply τ of player 2 such that the expected λ-discounted payoff under $\mathbf{x}$ and τ is at most ε, provided λ is close enough to zero. Thus, by playing properly, player 2 can lower the expected payoff of player 1 as close to 0 as he wishes. This will prove that player 1 cannot guarantee a positive payoff in all discounted games with sufficiently low discount factor.

Let ε be given. As $\mathbf{x}$ is given, there is N sufficiently large such that the probability that the game reaches stage N and player 1 plays the Top row at least once after that stage is at most $\varepsilon/2$, that is, $\mathbf{P}(N \leq \theta < +\infty) \leq \varepsilon/2$. Let τ be the pure strategy that plays the Left column up to stage N, and the Right column afterwards. If player 1 plays the Top row at some stage $n \leq N$, the terminal payoff is 0. Since the probability that player 1 plays the Top row after stage N is at most $\varepsilon/2$, and since after stage N player 2 plays the Right column, the expected λ-discounted payoff from stage N on, when restricted to the event that play reaches stage N, is at most $\varepsilon/2$. Therefore, if λ is sufficiently small so that the contribution of the first N stages to the λ-discounted payoff is at most $\varepsilon/2$, we deduce that the λ-discounted payoff under $(\mathbf{x}, \tau)$ is at most ε.

1.3 A modified Big Match

We here discuss a striking example due to Coulomb (1992), which is a modification of the Big Match. The game is given by the matrix

	L	M	R
T	0^*	1^*	γ
B	1	0	γ

where $\gamma \geq \frac{1}{2}$ is arbitrary. If either the action combination (T, L) or the action combination (T, M) is played, the play moves to an absorbing state with constant payoff. Note that the current state can change at most once along any play, as in the Big Match. This game differs from the Big Match by the adjunction of the column R.

Assuming perfect monitoring, this extra column makes no difference, since $\gamma \geq 1/2$. Indeed, let σ be any strategy in the Big Match, and define a strategy $\widetilde{\sigma}$ in the present game as follows. Given a history $\widetilde{h}$, $\widetilde{\sigma}$ plays at $\widetilde{h}$ the mixed move that would be played by σ at h, where h is obtained from $\widetilde{h}$ by deleting all stages in which player 2 played R. It can be checked that $\widetilde{\sigma}$ guarantees $1/2-\varepsilon$ in the λ-discounted game, as soon as σ guarantees $1/2-\varepsilon$ in the λ-discounted Big Match.

We shall now assume that player l is only imperfectly informed of player 2's past choices. Specifically, we shall assume that, whenever player 1 plays B, he is "told" "L" if player 2 played L, and "M or R" otherwise. The information received by player 1 upon playing T is irrelevant for the present analysis, as well as the signals for player 2.

We now check that in this game player 2 can do much better than in the Big Match. Specifically, given any strategy σ of player 1 and any $\varepsilon > 0$, we shall exhibit a strategy τ of player 2 such that the expected λ-discounted payoff under (σ, τ) is at most ε, for every λ close enough to zero. This result is striking because the signalling structure and the payoffs are such that this game is a Big Match with perfect monitoring with an additional column which payoffs can be as high as we want. Nevertheless, the highest quantity that player 1 can guarantee in any discounted game with small enough discount factor decreases from 1/2 which is the value of the Big Match with perfect monitoring to 0.

Define θ to be the first stage in which either (T, L) or (T, M) is played, so that θ is the stage at which the game effectively terminates.

Let y be the stationary strategy of player 2 that plays L and R with probabilities $\varepsilon/2$ and $1-\varepsilon/2$ respectively, and let y' be the stationary strategy of player 2 that plays L and M with probabilities $\varepsilon/2$ and $1-\varepsilon/2$ respectively.

If the probability that under (σ, y) the game terminates in finite time is 1, then the probability that under (σ, y) the game terminates by (T, L) is 1,

so that for every λ sufficiently small, the λ-discounted payoff under (σ, y) is at most ε, as desired.

So assume that this probability is strictly less than 1. Therefore, there is N such that the probability that under (σ, y) stage N is reached and player 1 plays the Top row after stage N is at most $\varepsilon/2$. Let τ be the strategy that coincides with y up to stage N, and with y' afterwards. Since the signal to player 1 is "M or R", whether the action pair is (B, M) or (B, R), as long as player 1 follows σ and does not play the Top row, he *cannot* tell whether he plays against y or against τ. Since the probability that player 1 plays the Top row after stage N is at most $\varepsilon/2$, the probability that player 1 can distinguish between y and between τ is at most $\varepsilon/2$. This means that the probability that under (σ, τ) player 1 plays the Top row after stage N is at most $\varepsilon/2$, while the probability that player 2 plays L at any given stage after stage N is $1-\varepsilon/2$, so that the expected λ-discounted payoff, restricted to the event that the game is not terminated before stage N, is at most ε. If λ is sufficiently small so that the contribution of the first N stages to the discounted payoff is at most ε, we deduce that the expected λ-discounted payoff under (σ, τ) is at most 2ε.

The two phases in the definition of τ have a natural interpretation. In the first phase, player 2 exhausts the probability that the play will end up in an absorbing state. In the second phase, player 2 switches to a mixed move that yields a low stage payoff. The fact that the mixed moves used by player 2 in the two phases cannot be distinguished by player 1 (as long as he plays B) guarantees that the probability that the play moves to an absorbing state in the second phase is very low. This two-part definition of a reply of player 2 to a given strategy of player 1 turns out to be a powerful tool, see Section 3.4.

2 The model and the results

2.1 Stochastic games with imperfect monitoring

We proceed to the model of stochastic games with imperfect monitoring. Given a finite set K, $\Delta(K)$ will denote the set of probability distributions over K. An element $k \in K$ will be identified with the element of $\Delta(K)$ that assigns probability one to k.

A *two-person zero-sum stochastic game with imperfect monitoring* is described by: (i) a set S of states, (ii) action sets A and B for the two players, (iii) a daily reward function $r : S \times A \times B \to \mathbf{R}$, (iv) signal sets M^1 and M^2 and (v) a transition function $\psi : S \times A \times B \to \Delta(M^1 \times M^2 \times S)$. The sets S, A, B, M^1 and M^2 are assumed to be finite.

The game is played in stages. An initial state s_1 is given and known to both players. At each stage $n \in \mathbf{N}$, (a) the players independently choose actions $a_n \in A$ and $b_n \in B$; (b) a triple (m_n^1, m_n^2, s_{n+1}) is drawn according

to $\psi(s_n, a_n, b_n)$; (c) players 1 and 2 are *only* told m_n^1 and m_n^2 respectively and (d) the game proceeds to stage $n+1$.

We denote by ψ^1 the marginal of ψ on M^1. It stands for the distribution of player 1's signal, as a function of the current state and the current action choices. Unlike in the previous examples, note that we allow for the case where the signal depends stochastically on (s, a, b). We will always assume that each player always knows the current state, and has perfect recall (i.e., remembers his past choices and past information).

This implies that ψ^1 is such that $(s, a) = (s', a')$ as soon as $\psi^1(s, a, b)[m^1] > 0$ and $\psi^1(s', a', b')[m^1] > 0$, for some $m^1 \in M^1$.

We define accordingly ψ^2 as the marginal of ψ over M^2. We also denote by q the marginal of ψ on S. Thus, $q(s, a, b)[s']$ is the probability of moving from s to s', if the players play a and b.

In a sense, ψ^1 provides all the information player 1 has on player 2's current move. However, since the signals to the two players can be correlated, the pair (ψ^1, ψ^2) does not fully describe the information available to player 1 on player 2's signal. Therefore, our model is more general than a model in which, given (s, a, b), the next state and the signals are chosen independently.

A behavior strategy of player 1 is a sequence $\sigma = (\sigma_n)_{n\geq 1}$ of functions $\sigma_n : H_n^1 \to \Delta(A)$, where $H_n^1 = S \times (M^1)^{n-1}$ is the set of "private" histories of player 1 at stage n. A stationary strategy depends only on the current stage. Hence, a stationary strategy of player 1 is described by a vector $(x^s)_{s\in S}$ in $(\Delta(A))^S$, with the interpretation that x^s is the mixed move used whenever the current state is $s \in S$. Strategies τ of player 2 are defined analogously, with obvious changes. We let $H_\infty = (S \times A \times B \times M^1 \times M^2)^{\mathbf{N}}$ denote the set of *plays*. For $i = 1, 2$, $\mathcal{H}_n^i$ denotes the cylinder algebra over H_∞ induced by H_n^i, and we let $\mathcal{H}_\infty = \sigma(\mathcal{H}_n^1, \mathcal{H}_n^2, n \in \mathbf{N})$ denote the σ-algebra generated by all cylinder sets. A given strategy pair (σ, τ), together with an initial state $s \in S$, induces a probability distribution $\mathbf{P}_{s,\sigma,\tau}$ over $(H_\infty, \mathcal{H}_\infty)$. Expectations w.r.t. $\mathbf{P}_{s,\sigma,\tau}$ are denoted $\mathbf{E}_{s,\sigma,\tau}$. Note that the initial state is a parameter, and not a data of the game.

Given $\lambda \in (0, 1)$, the λ-discounted payoff induced by a strategy pair (σ, τ) starting from state $s \in S$, is given by

$$\gamma_\lambda(s, \sigma, \tau) = \mathbf{E}_{s,\sigma,\tau}\left[\lambda \sum_{n=1}^{+\infty} (1-\lambda)^{n-1} r(s_n, a_n, b_n)\right].$$

The seminal result of Shapley (1953) asserts that the λ-discounted game has a value v_λ that does not depend on ψ. That is, for each $s \in S$, the zero-sum game with payoff function $\gamma_\lambda(s, \cdot, \cdot)$ has a value. We now introduce the definitions of the uniform properties we will be dealing with.

Definition 1 *Let $\phi \in \mathbf{R}^S$. Player 1 can* guarantee *ϕ if for every $\varepsilon > 0$ there exists a strategy σ and $\lambda_0 \in (0,1)$ such that:*

$$\forall s \in S, \forall \tau, \forall \lambda \in (0, \lambda_0),\ \gamma_\lambda(s, \sigma, \tau) \geq \phi(s) - \varepsilon.$$

We then say that the strategy σ guarantees $\phi - \varepsilon$.

Definition 2 *Let $\phi \in \mathbf{R}^S$. Player 2 can* defend *ϕ if for every $\varepsilon > 0$ and every strategy σ of player 1 there exists a strategy τ of player 2 and $\lambda_0 \in (0,1)$, such that:*

$$\forall s \in S, \forall \lambda \in (0, \lambda_0), \gamma_\lambda(s, \sigma, \tau) \leq \phi(s) + \varepsilon.$$

We then say that the strategy τ defends $\phi + \varepsilon$ against σ.

The definitions of a vector guaranteed by player 2, and defended by player 1, are similar, with the roles of the two players exchanged.

Definition 3 *A vector $v \in \mathbf{R}^S$ is*

- *the (uniform)* value *of Γ if both players can guarantee v;*
- *the (uniform) max-min if player 1 can guarantee v and player 2 can defend v;*
- *the (uniform) min-max if player 1 can defend v and player 2 can guarantee v.*

Assume that player 1 cannot guarantee ϕ. Then, for every strategy σ and every $\lambda_0 \in (0,1)$, there is a strategy τ such that $\gamma_\lambda(s, \sigma, \tau) < \phi(s) - \varepsilon$, for *some* $\lambda \in (0, \lambda_0)$ and some $s \in S$. Plainly, it does not follow that player 2 can defend ϕ. Therefore the existence of the max-min is not at all a trivial matter.

Note that the value coincides with $\lim_{\lambda \to 0} v_\lambda$, as soon as it exists. Also, if both the max-min and the min-max exist, one has max-min $\leq$ min-max.

2.2 The results

We first quote two known results. A stochastic game has perfect monitoring if the signal received by a player always reveals the current state and the action choices. Formally, given $(s, a, b) \neq (s', a', b')$, the supports of the probability distributions $\psi^i(s, a, b)$ and $\psi^i(s', a', b')$ are disjoint, for $i = 1, 2$.

Theorem 4 (Mertens and Neyman, 1981) *Every two-player zero-sum stochastic game with perfect monitoring has a value.*

Actually, the proof of Theorem 4 is valid as soon as the signals the players receive at each stage n contain the new state s_{n+1} and the payoff r_n at stage n.

Let Γ be a stochastic game. A state $s \in S$ is *absorbing* if $q(s,a,b)[s] = 1$, for every $(a,b) \in A \times B$. The game Γ is *absorbing* if all states but one are absorbing.

Theorem 5 (Coulomb, 1992, 1999, 2001) *Every two-player zero-sum absorbing stochastic game has a max-min and a min-max. The max-min (resp. the min-max) depends on ψ only through ψ^1 (resp. through ψ^2).*

In the rest of this chapter, we will report on the following theorem, obtained independently by Coulomb (2003) and Rosenberg et *al.* (2003). Our goal is to identify the main ideas of the proof, and to strip the exposition from details. The interested reader should consult Coulomb (2003) or Rosenberg et *al.* (2003)

Theorem 6 (Coulomb, 2003, Rosenberg, Solan and Vieille, 2003) *Every two-player zero-sum stochastic game with imperfect monitoring has a max-min and a min-max. The max-min (resp. the min-max) depends on ψ only through ψ^1 (resp. through ψ^2).*

The proof of Theorem 6 is quite related to the proof of Theorem 4. It is independent of the proof of Theorem 5.

W.l.o.g. we focus on the existence of the max-min value, and assume that payoffs are non-negative and bounded by 1.

3 Existence of the max-min – Highlights

3.1 On Mertens and Neyman's (1981) proof

The proof of Theorem 6 builds upon the proof of Theorem 4. We therefore start by recalling the main insights of Mertens and Neyman (1981, hereafter MN). We will next single out the main computational step in their proof, and discuss the additional issues that arise in games with imperfect monitoring.

MN offer a wide class of ε-optimal strategies σ for player 1. All share the following structure. The play is divided into blocks of random finite length L_k. On each block k, the strategy requires to play an optimal strategy in the λ_k-discounted game. Both L_k and λ_k depend on an auxiliary parameter, z_k: $L_k = L(z_k)$ and $\lambda = \lambda(z_k)$. In a sense, $z_k \in \mathbf{R}$ is a statistic that summarizes all the relevant aspects of the play, up to the beginning of block k.

To be specific, given two functions $L : [0, +\infty) \to \mathbf{N}$ and $\lambda : [0, +\infty) \to \mathbf{R}$, and $M \in \mathbf{R}$, the sequences $(z_k), (L_k), (\lambda_k)$ are defined recursively by

$$\begin{aligned} &z_1 \geq Z, B_1 = 1,\ \lambda_k = \lambda(z_k), L_k = L(z_k), \\ &B_{k+1} = B_k + L_k, z_{k+1}\ = \max\left\{ Z, z_k + \frac{\varepsilon}{2} + \sum_{B_k \leq n < B_{k+1}} (r_n - v(s_{B_{k+1}})) \right\}, \end{aligned} \tag{1}$$

where $v = \lim_{\lambda \to 0} v_\lambda$ and $r_n = r(s_n, a_n, b_n)$ is the payoff in stage n. In a first approximation, the new value z_{k+1} of the statistic is obtained by adding to the previous value z_k the excess of payoffs received over the values of the states visited along the block.

MN provide sufficient conditions on the functions $\lambda(\cdot)$ and $L(\cdot)$ under which the above strategy σ is ε-optimal, for M large. These conditions are in particular satisfied for each of the two following simple functions.

Case 1: $\lambda(z) = z^{-\beta}$ and $L(z) = \lceil \lambda(z)^{-\alpha} \rceil$,[2] where $\alpha \in (0, 1)$ satisfies $\|v_\lambda - v\|_\infty < \lambda^{1-\alpha}$ for every λ sufficiently close to 0, and $\beta > 1$ satisfies $\alpha\beta < 1$;

Case 2: $\lambda(z) = 1/(z \ln^2 z)$ and $L(z) = 1$.

Given an appropriate choice for $\lambda(\cdot)$ and $L(\cdot)$, MN prove that $\gamma_\lambda(s, \sigma, \tau) \geq v(s) - \varepsilon$ for each τ, and that $\mathbf{E}_{s,\sigma,\tau}\left[\sum_{k=1}^{+\infty} \lambda_k L_k\right] < +\infty$.

The proof relies on the semi-algebraicity of the map $\lambda \mapsto v_\lambda$, due to Bewley and Kohlberg (1976), and on inequality (2) below, which holds for every τ, since during block k σ follows an optimal strategy in the λ_k-discounted game:

$$\mathbf{E}_{s,\sigma,\tau}\left[\lambda_k \sum_{n=B_k}^{B_{k+1}-1} (1-\lambda_k)^{n-B_k} r_n + (1-\lambda_k)^{L_k} v_{\lambda_k}(s_{B_{k+1}}) | \mathcal{H}_{B_k} \right] \geq v_{\lambda_k}(s_{B_k}). \tag{2}$$

We conclude this section by a list of standing issues, that need to be addressed in order to adapt MN's proof to games with imperfect monitoring. This list is not exhaustive.

- In games with imperfect monitoring, the max-min need not be equal to the limit of the λ-discounted values. The above proof asserts that player 1 can guarantee $\lim_{\lambda \to 0} v_\lambda$. Therefore, we will have to define *auxiliary* discounted games. The proof when imperfect monitoring

[2] For every $c \in \mathbf{R}$, $\lceil c \rceil$ is the minimal integer greater than or equal to c.

is present will assert that the max-min is equal to the limit of the solutions to these auxiliary discounted games. The definition of these auxiliary games will take into account the structure ψ of signals.

- The solution w_λ of these auxiliary games will have to be semi-algebraic as a function of λ.
- In (1), the updating formula for z_k involves $\sum_{B_k \leq n < B_{k+1}} r_n$, the payoffs received in the previous block. Since this quantity is not available to player 1, we will have to estimate it using only the information that is available to player 1. In effect, we will use a measure of the worst payoff that is consistent with the distribution of the signals received in the elapsed block.
- Finally, the ε-optimality computation will have to be adapted.

As it so turns out, the last issue is easy. Specifically, replace in (1) the term $\sum_{n=B_k}^{B_{k+1}-1} r_n$ by an $\mathcal{H}^1_{B_k}$-measurable variable $L_k \widehat{r}_k$, and let $\lambda \mapsto w_\lambda$ be a $\mathbf{R}^S$-valued semi-algebraic function, with $w := \lim_{\lambda \to 0} w_\lambda$. Let $\lambda(\cdot)$, $L(\cdot)$ satisfy MN's sufficient conditions, and M be large enough. A close inspection of MN's proof reveals that the following Proposition holds

Theorem 7 *There exists $\lambda_0 \in (0,1)$ such that the following holds. Let (σ, τ) be a strategy pair such that*

$$\mathbf{E}_{s,\sigma,\tau}\left[\lambda_k L_k \widehat{r}_k + (1 - \lambda_k L_k) w_{\lambda_k}(s_{B_{k+1}}) | \mathcal{H}^1_{B_k}\right] \geq w_{\lambda_k}(s_{B_k}) - \frac{\varepsilon}{12}\lambda_k L_k, \quad (3)$$

$\mathbf{P}_{s,\sigma,\tau}$-a.s. for every k. Then for each $\lambda \in (0, \lambda_0)$,

$$\mathbf{E}_{s,\sigma,\tau}\left[\lambda \sum_{n=1}^{+\infty} (1-\lambda)^{n-1} \widehat{R}_n\right] \geq w(s) - \varepsilon, \text{ where } \widehat{R}_n = \widehat{r}_k \text{ for } B_k \leq n < B_{k+1}. \quad (4)$$

Moreover,

$$\mathbf{E}_{s,\sigma,\tau}\left[\sum_{k=1}^{+\infty} \lambda_k L_k\right] < +\infty. \quad (5)$$

3.2 Auxiliary discounted games

We let a stochastic game $\Gamma = (S, A, B, M^1, M^2, \psi, r)$ be given. We here define an auxiliary family of stochastic games. The stage payoff of these games incorporates the structure of signals.

A preliminary comment is in order. Consider first a *repeated* game with imperfect monitoring. Assume that player 1 and player 2 consider using

mixed moves $x \in \Delta(A)$ and $y \in \Delta(B)$ in some given stage. If player 2 replaces y by another mixed move $y' \in \Delta(B)$, this replacement can possibly have an effect on the future behavior of player 1 only if it alters the distribution of signals to player 1 at that stage. In other words, if y' is *indistinguishable* from y, in the sense that the distributions $\psi^1(x,y)$ and $\psi^1(x,y')$ of signals to player 1 coincide, then switching from y to y' while player 1 is using x has no incidence whatever on player 1's future behavior.[3] This suggests that a proper modified payoff function for player 1 is $\widetilde{r}(x,y) = \inf r(x,y')$, where the infimum is taken over all $y' \in \Delta(B)$ that are indistinguishable from y given x. That is, $\widetilde{r}(x,y)$ is the worst payoff to player 1, given that player 1's signals are consistent with y.

This equivalence relation and the corresponding modified payoff function have played an important role in the analysis of games with imperfect monitoring, see Aumann and Maschler (1995), Lehrer (1989, 1990, 1992a, 1992b) and Coulomb (1999, 2001).

However, this relation is not well suited for general stochastic games with imperfect monitoring. Indeed, a mixed move y' can be practically indistinguishable from y if the probability that player 1 can distinguish between y and y' is quite small compared to the discount factor. We therefore amend it as follows. Given a discount factor $\lambda \in (0,1)$, a state $s \in S$, a mixed move $x \in \Delta(A)$, and an additional parameter $\varepsilon \in (0,1)$, we say that $y \in \Delta(B)$ and $z \in \Delta(B)$ are indistinguishable, written $y \sim_{\lambda,\varepsilon,s,x} z$ if

$$\psi^1(s,a,y) = \psi^1(s,a,z) \text{ for every } a \text{ such that } x[a] \geq \lambda/\varepsilon.$$

Accordingly, we set

$$\widetilde{r}^{\varepsilon}_{\lambda}(s,x,y) = \min_{z \sim_{\lambda,\varepsilon,s,x} y} r(s,x,z). \tag{6}$$

As above, it can be thought of as the worst payoff consistent with a given distribution of signals to player 1. The specific role of the parameter ε will be clarified later.

We briefly mention some basic properties of $\widetilde{r}^{\varepsilon}_{\lambda}$. Note first that $\widetilde{r}^{\varepsilon}_{\lambda} \leq r$. Since $\sim_{\lambda,\varepsilon,s,x}$ is an equivalence relation, one has $\widetilde{r}^{\varepsilon}_{\lambda}(s,x,y) = \widetilde{r}^{\varepsilon}_{\lambda}(s,x,z)$ whenever $z \sim_{\lambda,\varepsilon,s,x} y$. In addition, it can be checked that, for fixed λ, ε and s, the function $\widetilde{r}^{\varepsilon}_{\lambda}(s,.,.)$ is continuous with respect to y and upper semi-continuous in the pair (x,y). Finally, the map $(\varepsilon,\lambda,x,y) \mapsto \widetilde{r}^{\varepsilon}_{\lambda}(s,x,y)$ is semi-algebraic.

We now proceed to introducing a vector $v^{\varepsilon}_{\lambda}$, that will play the role of the "value" of the auxiliary discounted game. Specifically, define $v^{\varepsilon}_{\lambda} \in \mathbf{R}^S$ as

[3] This is equivalent to the requirement that $\psi^1(a,y) = \psi^1(a,y')$ for every action $a \in A$ that is played with positive probability under x.

the unique solution to the fixed-point equation

$$v_\lambda^\varepsilon(s) := \max_{x\in\Delta(A)} \min_{y\in\Delta(B)} \left\{\lambda \widetilde{r}_\lambda^\varepsilon(s,x,y) + (1-\lambda)\mathbf{E}_{q(s,x,y)}[v_\lambda^\varepsilon(\cdot)]\right\},\ w \in \mathbf{R}^S, \tag{7}$$

where $\mathbf{E}_{q(\cdot|s,x,y)}$ is the expectation w.r.t. $q(s,x,y)$.[4] It follows from this fixed-point property that the map $(\lambda,\varepsilon) \mapsto v_\lambda^\varepsilon(s)$ is semi-algebraic.

One can relate v_λ^ε to the sup inf of some non-standard λ-discounted game. Indeed, define the (ε,λ)-game to be a λ-discounted game, in which the stage payoff is $\widetilde{r}_\lambda^\varepsilon$. The (ε,λ)-game differs from standard stochastic games in several respects. At each stage, the players choose *mixed moves* in $\Delta(A)$ and $\Delta(B)$ (and not actions in A and B). In addition, the stage payoff function depends on the discount factor being used. It can be checked that v_λ^ε coincides with the sup-inf of the (ε,λ)-game, when players are restricted to pure strategies.

We conclude this section by offering a candidate for the max-min. Since the map $\lambda \mapsto v_\lambda^\varepsilon(s)$ is semi-algebraic for fixed ε, the limit $\lim_{\lambda\to 0} v_\lambda^\varepsilon(s)$ exists for every $\varepsilon > 0$. In addition, the auxiliary reward $\widetilde{r}_\lambda^\varepsilon$ is non-decreasing w.r.t. ε, hence so is $\lim_{\lambda\to 0} v_\lambda^\varepsilon(s)$. As a consequence, the limit $v := \lim_{\varepsilon\to 0}\lim_{\lambda\to 0} v_\lambda^\varepsilon$ exists. It turns out that v is the max-min of the game Γ, as we explain in the following two sections.

3.3 Guaranteeing v

We here explain why player 1 can guarantee v. We shall rely on the tools introduced in Section 3.1, and we first introduce the function w_λ that will be used. Using the theory of semi-algebraic sets, there is a semi-algebraic function $\lambda \in (0,1) \mapsto \varepsilon(\lambda) \in (0,1)$ such that $\lambda \le \varepsilon(\lambda)^2$ for each λ, and $\lim_{\lambda\to 0} v_\lambda^{\varepsilon(\lambda)} = v$. We set $w_\lambda := v_\lambda^{\varepsilon(\lambda)}$. Besides, there is a semi-algebraic map $\lambda \in (0,1) \mapsto x_\lambda = (x_\lambda^s)_{s\in S} \in \Delta(A)^S$, such that, for each $s \in S$, x_λ^s achieves the maximum in the definition of $v_\lambda^{\varepsilon(\lambda)}$, see (7). By semi-algebraicity again, the set $\overline{A}(s) = \{a \in A : x_\lambda^s[a] \ge \lambda/\varepsilon(\lambda)\}$ is, for λ close enough to zero, independent of λ.

We now define the estimate $\widehat{r}_k$ that is used by player 1 at the end of block k to update the statistic z_k. At the end of block k, player 1 collects the signals he received during the block. For each state $s \in S$, player 1 computes a mixed move $\widehat{y}^s \in \Delta(B)$ that is "most likely" given the signals he received in state s. Specifically, for each state $s \in S$, and each action $a \in \overline{A}(s)$, player 1 computes the empirical distribution $\rho_{s,a}$ of signals that he received in those stages in which he played a while at state s (if there was no such stage, the definition of $\rho_{s,a}$ is irrelevant). The mixed move $\widehat{y}^s$ is chosen to minimize over $y \in \Delta(B)$ the maximal discrepancy $\max_{a\in\overline{A}(s)} \left\|\rho_{s,a} - \psi^1(s,a,y)\right\|_\infty$.

[4] The justification of why the max and min in (7) are achieved is omitted.

Finally, player 1 sets

$$\widehat{r}_k = \frac{1}{L_k} \sum_{s \in S} N_s \widetilde{r}^{\varepsilon(\lambda_k)}_{\lambda_k}(s, x^s_{\lambda_k}, \widehat{y}^s),$$

where N_s is the number of times the state s was visited during block k. In a sense, $\widehat{r}_k$ is the worst (average) payoff in block k, given that player 2 played a stationary strategy that is consistent with the signals to player 1.

The strategy of player 1 is defined as in Section 3.1, taking **Case 1** specifications for $\lambda(\cdot)$ and $L(\cdot)$. To be precise, we first choose $d > 0$ such that $\varepsilon(\lambda) \leq \lambda^d$ for λ close enough to zero. We next choose $\alpha \in (1-d, 1)$, $\beta \in (1, 1/\alpha)$ and we set $\lambda(z) = z^{-\beta}$, $L(z) = \lceil \lambda(z)^{-\alpha} \rceil$.

We turn to the intuition of the proof. The crucial part is to show that the inequality (3) is satisfied, provided M is large enough. To this end, we introduce, for each $s \in S$, the average mixed move $\overline{y}^s$ used by player 2 in state s.[5] This average mixed move $\overline{y}$ can be related to the strategy $\widehat{y}$ that is reconstructed by player 1 at the end of the block. Indeed, fix a state $s \in S$, and some action $a \in \overline{A}(s)$. By the definition of $\overline{A}(s)$, at any visit to the state s, the action a is played with probability at least $\lambda/\varepsilon(\lambda) \geq \lambda^{1-d}$, which is much larger than $1/L_k$, provided M is large enough. Thus, provided the number of visits to s exceeds a small fraction of L_k, the action a will typically be played many times. Since the action choices of the two players are independent (conditional on past play), it is quite likely that the empirical distribution of signals $\rho_{s,a}$ will be very close to $\psi^1(s, a, \overline{y}^s)$. As a result, provided the state s is visited more than a negligible fraction of L_k stages, the reconstructed strategy $\widehat{y}$ will be such that $\left\| \psi^1(s, a, \overline{y}^s) - \psi^1(s, a, \widehat{y}^s) \right\|_\infty$ is close to zero. By continuity, this will imply that $\widetilde{r}^{\varepsilon(\lambda_k)}_{\lambda_k}(s, x^s_{\lambda_k}, \overline{y}^s)$ is close to $\widetilde{r}^{\varepsilon(\lambda_k)}_{\lambda_k}(s, x^s_{\lambda_k}, \widehat{y}^s)$.[6] On the other hand, states that are visited less than a negligible fraction of L_k stages hardly contribute to $\widehat{r}_k$. Therefore, the expectation $\mathbf{E}_{s,\sigma,\tau}\left[L_k \widehat{r}_k | \mathcal{H}^1_{B_k} \right]$ of $\widehat{r}_k$ given past history is close to $\mathbf{E}_{s,\sigma,\tau}\left[\sum_{s \in S} N_s \widetilde{r}^{\varepsilon(\lambda_k)}_{\lambda_k}(s, x^s_{\lambda_k}, \overline{y}^s) | \mathcal{H}^1_{B_k} \right]$.

Using the optimality of x_{λ_k}, it can be checked – although this is not a trivial observation – that the difference

$$\mathbf{E}_{s,\sigma,\tau}\left[\lambda_k \sum_{s \in S} N_s \widetilde{r}^{\varepsilon(\lambda_k)}_{\lambda_k}(s, x^s_{\lambda_k}, \overline{y}^s) + (1 - \lambda_k L_k) v^{\varepsilon(\lambda_k)}_{\lambda_k}(s_{B_{k+1}}) | \mathcal{H}^1_{B_k} \right] - v^{\varepsilon(\lambda_k)}_{\lambda_k}(s_{B_k})$$

is minorized by an amount of the order $\varepsilon \lambda_k L_k$. As a consequence, (3) holds.

[5] It is given by $\overline{y}^s = \frac{1}{N_s} \sum_{n : s_n = s} y_n$, where the summation runs over all stages of block k, and where $y_n = \tau(h^2_n)$ is the mixed move used by player 2 in stage n.

[6] The formal proof involves many technical complications.

3.4 Defending v

We here deal with the other side of the analysis. We will prove that player 2 can defend $v^\varepsilon := \lim_{\lambda\to 0} v^\varepsilon_\lambda$ for each $\varepsilon > 0$. Let $\varepsilon > 0$, and a strategy σ of player 1 be given. Generalizing upon the example of Section 1.3, we shall define a reply τ in two steps.

First, we use the tools of Section 3.1 to construct a strategy τ_1 that yields a low discounted payoff against σ, when measured in terms of $\widetilde{r}^\varepsilon_\lambda$. It is convenient here to use the specifications of **Case 1** for the functions $\lambda(\cdot)$ and $L(\cdot)$: $L(z) = 1$ and $\lambda(z) = 1/(z\ln^2 z)$. In effect, player 2 updates his summary z_k at every stage. We define simultaneously and inductively the strategy τ_1 and the estimate $\widehat{r}_k$.

Consider a given stage $n \in \mathbf{N}$, and assume that τ_1 has already been defined for the first $n-1$ stages, together with $\widehat{r}_1, \cdots, \widehat{r}_{n-1}$. Consequently, player 2 has in mind a fictitious discount factor $\lambda_n = \lambda(z_n)$, as determined by (1). At stage n, player 2 computes the conditional distribution of player 1's action choice in stage n, given the past sequence of states. To be specific, we set $\xi_n[\cdot] = \mathbf{P}_{s,\sigma,\tau_1}(a_n = \cdot \mid s_1, \ldots, s_n)$. Note that this distribution involves only the restriction of τ_1 to the first $n-1$ stages, and the observation of past states, so that player 2 is indeed in a position to compute ξ_n. The strategy τ_1 recommends playing a mixed move $y_n \in \Delta(B)$, that satisfies

$$\lambda_n \widetilde{r}^\varepsilon_{\lambda_n}(s_n, \xi_n, y_n) + (1-\lambda_n)\mathbf{E}_{q(s_n,\xi_n,y_n)}\left[v^\varepsilon_{\lambda_n}\right] \leq v^\varepsilon_{\lambda_n}(s_n). \tag{8}$$

We set $\widehat{r}_n = \widetilde{r}^\varepsilon_{\lambda_n}(s_n, \xi_n, y_n)$. This completes the definition of τ_1.

By the choice of y_n and the definition of $\widehat{r}_n$, (3) trivially holds (with the inequality reversed), which implies that

$$\mathbf{E}_{s,\sigma,\tau_1}\left[\lambda\sum_{n=1}^{+\infty}(1-\lambda)^{n-1}\widetilde{r}^\varepsilon_{\lambda_n}(s_n, \xi_n, y_n)\right] \leq v^\varepsilon(s) + \varepsilon, \tag{9}$$

provided λ is close enough to zero.

The inequality (9) says that, if the payoff at every given stage were defined as the *worst* payoff $\widetilde{r}^\varepsilon_{\lambda_n}$, consistent with the actual choice of player 2, then the discounted payoff would be low. Since $\widetilde{r}^\varepsilon_{\lambda_n} \leq r$, it however fails to imply $\gamma_\lambda(s,\sigma,\tau_1) \leq v(s) + 2\varepsilon$. We now address this issue.

Given a stage n, we let $z_n \in \Delta(B)$ be a mixed move such that $z_n \sim_{\lambda_n,\varepsilon,s_n,\xi_n} y_n$ and $r(s_n, \xi_n, z_n) = \widetilde{r}^\varepsilon_{\lambda_n}(s_n, \xi_n, y_n)$. Hence, z_n achieves the minimal payoff against ξ_n, among all the mixed moves that are indistinguishable from y_n. By the definition of the equivalence relation $\sim_{\lambda_n,\varepsilon,s_n,\xi_n}$, the probability (given the sequence of states) that at stage n, player 1 plays an action that might possibly distinguish y_n from z_n is at most $|A|\,\lambda_n/\varepsilon$. We next make use of the fact that

$$\mathbf{E}_{s,\sigma,\tau_1}\left[\sum_{n=1}^{+\infty}\lambda_n\right] < +\infty$$

(see Theorem 7) to choose $N \in \mathbf{N}$ such that $\mathbf{E}_{s,\sigma,\tau}\left[\sum_{n=N}^{+\infty} \lambda_n\right] < \frac{\varepsilon^2}{|A|}$. Finally, we let τ be the strategy that coincides with τ_1 up to stage N, and plays z_n rather than y_n in each subsequent stage $n \geq N$.

By the choice of N, the probability that player 1 will ever, from stage N on, play an action that might possibly distinguish τ from τ_1 is at most $\frac{|A|}{\varepsilon}\mathbf{E}_{s,\sigma,\tau}\left[\sum_{n=N}^{+\infty} \lambda_n\right] \leq \varepsilon$. This implies that the probability distributions $\mathbf{P}_{s,\sigma,\tau}$ and $\mathbf{P}_{s,\sigma,\tau_1}$ induced over the sequences of states differ by at most ε. Therefore,

$$\begin{aligned} \mathbf{E}_{s,\sigma,\tau}\left[r_n\right] = \mathbf{E}_{s,\sigma,\tau}\left[r(s_n,\xi_n,z_n)\right] &= \mathbf{E}_{s,\sigma,\tau}\left[\widetilde{r}^{\varepsilon}_{\lambda_n}(s_n,\xi_n,y_n)\right] \\ &\leq \mathbf{E}_{s,\sigma,\tau_1}\left[\widetilde{r}^{\varepsilon}_{\lambda_n}(s_n,\xi_n,y_n)\right] + \varepsilon. \end{aligned} \tag{10}$$

The first equality simply states that the payoff at stage n is the payoff function evaluated at the current state, with current mixed actions. The second equality follows by the choice of z_n. The inequality follows from the previous claim.

Together with (9), (10) implies that $\gamma_\lambda(s,\sigma,\tau) \leq v^\varepsilon(s) + 2\varepsilon$, provided ε is small enough.

Note that the strategy τ uses only the sequence of states, and not any additional signal that player 2 may receive. It is important to observe that z_n need not satisfy (8) since $q(s_n,\xi_n,y_n) \neq q(s_n,\xi_n,z_n)$. Hence, the two-part definition of τ cannot be avoided.

4 Concluding comments

The results discussed in the present chapter raise additional questions. We here mention just a few.

Within the framework of this survey, it would be useful to characterize the games that have a value. More precisely, given S, A, B, M^1 and M^2, it is interesting to know for which signalling structures ψ the game has a value, for every payoff function r.

The examples of Flesch et *al.* (2000) suggest that the analysis of non-zero-sum stochastic games with imperfect monitoring will need additional insights. This field is yet unexplored.

Finally, challenging problems arise as soon as one drops the assumption that the current state is observed. The one-player case has been investigated in Rosenberg et *al.* (2002). They prove that the value exists, in the sense that the player can guarantee $\lim_{\lambda\to 0} v_\lambda$. However, they leave unanswered basic questions on the nature of ε-optimal strategies.

In the two-player case, the model is related to stochastic games with incomplete information (see Sorin, 1984, 1985, 2002, Sorin and Zamir, 1985, Rosenberg and Vieille, 2000, Rosenberg et *al.*, 2002). Most work in this area focused on the case where the state is a pair (k,ω), and (i) the k-component

is fixed at the outset of the game and known to one player only, while (ii) the ω-component can change from stage to stage, but is observed by both players. A recent exception is the paper by Renault (2003), in which the state s follows a Markov chain, that is, the evolution of s is unaffected by action choices, and is observed only by one player. In this framework, Renault proves the existence of the value. This literature assumes that actions are observed.

References

[1] Aumann R.J. and Maschler M., with the collaboration of R.E. Stearns (1995), Repeated Games with Incomplete Information, MIT Press

[2] Aumann R.J. and Shapley L.S. (1994) Long-Term Competition—a Game-Theoretic Analysis. Essays in game theory (Stony Brook, NY, 1992), 1–15, Springer, New York

[3] Blackwell D. (1965) Discounted Dynamic Programming, *Ann Math. Stat.*, **36**, 266-235

[4] Blackwell D. and Ferguson T. (1968) The Big Match, *Ann. Math. Stat.,* **39**, 159-163

[5] Bewley T. and Kohlberg E. (1976) The Asymptotic Theory of Stochastic Games, *Math. Oper. Res.*, **1**, 197-208

[6] Bochnak J., Coste M. and Roy M.F. (1998) Real Algebraic Geometry, Springer Verlag, Berlin

[7] Coulomb J.M. (1992) Repeated Games with Absorbing States and No Signals, *Internat. J. Game Th.*, **21**, 161-174

[8] Coulomb, J.M. (1996) A Note on "Big Match", *ESAIM PS*, **1**, 89-93

[9] Coulomb J.M. (1999) Generalized Big-Match, *Math. Oper. Res.*, **24**, 795-816

[10] Coulomb J.M. (2001) Absorbing Games with a Signaling Structure, *Math. Oper. Res.*, **26**, 286-303

[11] Coulomb J.M. (2003) Stochastic Games without Perfect Monitoring, *preprint*

[12] Flesch J., Thuijsman F. and Vrieze O.J. (2000) Stochastic Games with Non-Observable Actions, *preprint*

[13] Gillette D. (1957) Stochastic Games with Zero Stop Probabilities, *Contributions to the Theory of Games*, **3**, Princeton University Press

[14] Lehrer E. (1989) Lower Equilibrium Payoffs in Two-Player Repeated Games with Nonobservable Actions, *Internat. J. Game Th.*, **18**, 57-89

[15] Lehrer E. (1990) Nash Equilibria of n-Player Repeated Games with Semi-Standard Information, *Internat. J. Game Th.*, **19**, 191- 217

[16] Lehrer E. (1992a) Correlated Equilibria in Two-Player Repeated Games with Nonobservable Actions, *Math. Oper. Res.* **17**, 175-199

[17] Lehrer E. (1992b) On the Equilibrium Payoffs Set of Two Player Repeated Games with Imperfect Monitoring, *Internat. J. Game Th.*, **20**, 211-226

[18] Mertens J.F. and Neyman A. (1981) Stochastic Games, *Internat. J. Game Th.*, **10**, 53-66

[19] Radner R. (1981) Monitoring Cooperative Agreements in Repeated Principle-Agent Relationship, *Econometrica*, **49**, 1127- 1148

[20] Rosenberg D., Solan E. and Vieille N. (2002) Blackwell Optimality in Markov Decision Processes with Partial Observation, *Ann. Stat.*, **30**, 1178-1193

[21] Rosenberg D., Solan E. and Vieille N. (2002) Stochastic games with a Single Controller and Incomplete Information, Discussion Paper 1346, Center for Mathematical Studies in Economics and Management Science, Northwestern University

[22] Rosenberg D., Solan E. and Vieille N. (2003) The Maxmin Value of Stochastic Games with Imperfect Monitoring, *preprint*

[23] Rosenberg D. and Vieille N. (2000) The Maxmin of Recursive Games with Incomplete Information on One Side, *Math. Oper. Res.*, **25**, 23-35

[24] Rubinstein A. and Yaari M. (1983) Repeated Insurance Contracts and Moral Hazard, *J. Econ. Th.*, **30**, 74-97

[25] Shapley L.S. (1953) Stochastic Games, *Proc. Nat. Acad. Sci. U.S.A.*, **39**, 1095-1100

[26] Sorin S. (1984) Big Match with Lack of Information on One Side (Part 1), *Internat. J. Game Theory*, **13**, 201-255

[27] Sorin S. (1985) Big Match with Lack of Information on One Side (Part 2), *Internat. J. Game Theory*, **14**, 173-204

[28] Sorin S. (1990) Supergames. Game theory and applications (Columbus, OH, 1987), 46–63, Econom. Theory Econometrics Math. Econom., Academic Press, San Diego, CA

[29] Sorin S. (2002) A First Course on Zero-Sum Repeated Games, *Mathematiques et applications, 37, Springer*

[30] Sorin S. and Zamir S. (1991) Big Match with lack of information on one side II, in *Stochastic Games and Related Topics*, T.E.S. Raghavan et al. (eds), Kluwer, 101-112