Commutative Stochastic Games

Xavier Venel [*]

April 18, 2016

Abstract

We are interested in the convergence of the value of n-stage games as n goes to infinity and the existence of the uniform value in stochastic games with a general set of states and finite sets of actions where the transition is commutative. This means that playing an action profile a_1 followed by an action profile a_2, leads to the same distribution on states as playing first the action profile a_2 and then a_1. For example, absorbing games can be reformulated as commutative stochastic games.

When there is only one player and the transition function is deterministic, we show that the existence of a uniform value in pure strategies implies the existence of 0-optimal strategies. In the framework of two-player stochastic games, we study a class of games where the set of states is $\mathbb{R}^m$ and the transition is deterministic and 1-Lipschitz for the L_1-norm, and prove that these games have a uniform value. A similar proof shows the existence of an equilibrium in the non zero-sum case.

These results remain true if one considers a general model of finite repeated games, where the transition is commutative and the players observe the past actions but not the state.

1 Introduction

A two-player zero-sum repeated game is a game played in discrete time. At each stage, the players independently take some decisions, which lead to an instantaneous payoff, a lottery on a new state, and a pair of signals. Each player receives one of the signals and the game proceeds to the next stage.

This model generalizes several models that have been studied in the literature. A Markov Decision Process (MDP) is a repeated game with a single player, called a decision maker, who observes the state and remembers his actions. A Partial Observation Markov Decision Processes (POMDP) is a repeated game with a single player who observes only a signal that depends on the state and his action. A stochastic game, introduced by Shapley [Sha53], is a repeated game where the players learn the state

[*] School of Mathematical Science, Tel Aviv University, email: xavier.venel@gmail.com

variable and past actions.

Given a positive integer n, the n-stage game is the game whose payoff is the expected average payoff during the first n stages. Under mild assumptions, it has a value, denoted v_n. One strand of the literature studies the convergence of the sequence of n-stage values, $(v_n)_{n\geq 1}$, as n goes to infinity.

The convergence of the sequence of n-stage values is related to the behavior of the infinitely repeated game. If the sequence of n-stage values converges to some real number v^*, one may also consider the existence of a strategy that yields a payoff close to v^* in every sufficiently long game. Let v be a real number. A strategy of player 1 *guarantees* v if for every $\eta > 0$ the expected average payoff in the n-stage game is greater than $v - \eta$ for every sufficiently large n and every strategy of player 2. Symmetrically, a strategy of player 2 *guarantees* v if for every $\eta > 0$ the expected average payoff in the n-stage game is smaller than $v + \eta$ for every sufficiently large n and every strategy of player 1. If for every $\epsilon > 0$, player 1 has a strategy that guarantees $v^* - \epsilon$ and player 2 has a strategy that guarantees $v^* + \epsilon$, then v^* is the *uniform value*. Informally, the players do not need to know the length of the game to play well, provided the game is long enough.

At each stage, the players are allowed to chose their action randomly. If each player can guarantee v^* while choosing at every stage one action instead of a probability over his actions, we say that the game has *a uniform value in pure strategies.*

Blackwell [Bla62] proved that an MDP with a finite set of states and a finite set of actions has a uniform value v^* and the decision maker has a pure strategy that guarantees v^*. Moreover, at every stage, the optimal action depends only on the current state. Dynkin and Juškevič [DJ79], and Renault [Ren11] described sufficient conditions for the existence of the uniform value when the set of states is compact, but in this more general setup there may not exist a strategy which guarantees the uniform value.

Rosenberg, Solan, and Vieille [RSV02] proved that POMDPs with a finite set of states, a finite set of actions, and a finite set of signals have a uniform value. Moreover, for any $\varepsilon > 0$, a strategy that guarantees $v^* - \varepsilon$ also yields a payoff close to v^* for other criteria of evaluation like the discounted evaluation. The existence of the uniform value was extended by Renault [Ren11] to any space of actions and signals, provided that at each stage only a finite number of signals can be realized.

In the framework of two-player games, Mertens and Neyman [MN81] showed the existence of the uniform value in stochastic games with a finite set of states and finite sets of actions. Their proof relies on an algebraic argument using the finiteness assumptions. Renault [Ren12] proved the existence of the uniform value for a two-player game where one player controls the transition and the set of states is a compact subset of a normed vector space.

There is an extensive literature about repeated games in which the players are not perfectly informed about the state or the actions played. Rosenberg, Solan and Vieille [RSV09] showed, for example, the existence of the uniform value in some class of games where the players observe neither the states nor the actions played by the other players. Another particular class that is closely related to the model we consider is repeated games with symmetric signals. At each stage, the players observe the past actions played and a public signal. Kohlberg and Zamir [KZ74] and Forges [For82] proved the existence of the uniform value, when the state is fixed once and for all at the outset of the game. Neyman and Sorin [NS98] extended this result to the non zero-sum case, and Geitner [Gei02] to an intermediate model where the game is half-stochastic and half-repeated. In these four papers, the unknown information concerns a parameter which does not change during the game. We will study models where this parameter can change during the game.

In general, the uniform value does not exist if the players have different information. For example, repeated games with incomplete information on both sides do not have a uniform value [AM95]. Nevertheless, Mertens and Zamir [MZ71] [MZ80] showed that the sequence of n-stage values converges. Rosenberg, Solan, and Vieille [RSV03] and Coulomb [Cou03] showed the existence of two quantities, called the $\max\min$ and the $\min\max$, when each player observes the state and his own actions but has imperfect monitoring of the actions of the other player. Moreover the $\max\min$ where player 2 chooses his strategy knowing the strategy of player 1, only depends on the information of player 1.

More surprisingly, the sequence $(v_n)_{n\geq 1}$ may not converge even with symmetric information. Vigeral [Vig13] provided an example of a stochastic game with a finite set of states and compact sets of actions where the sequence of n-stage values does not converge. Ziliotto [Zil13] provided an example of a repeated game with a finite set of states, finite sets of actions, and a finite set of public signals where a similar phenomenon occurs. In each case, the game has no uniform value.

In this paper, we are interested in two-player zero-sum stochastic games where the transition is commutative. In such games, given a sequence of decisions, the order is irrelevant to determine the resulting state: playing an action profile a_1 followed by an action profile a_2 leads to the same distribution over states as playing first the action profile a_2 and then a_1.

In game theory, several models satisfy this assumption. For example, Aumann and Maschler [AM95] studied repeated games with incomplete information on one side. One can introduce an auxiliary stochastic game where the new state space is the set of beliefs of the uninformed player and the sets of actions are the mixed actions of the original game. This game is commutative as we will show in Example 2.4. We will also show in Proposition 5.1 that absorbing games (see Kohlberg [Koh74]) can be reformulated as commutative stochastic games.

In Theorem 3.1 we prove that whenever a commutative MDP with a deterministic transition has a uniform value in pure strategies, the decision maker has a strategy that guarantees the value. Example 4.1 shows that to guarantee the value, the decision maker may need to choose his actions randomly. Under topological assumptions similar to Renault [Ren11], we show that the conclusion can be strengthened to the existence of a strategy without randomization that guarantees the value. By a standard argument, we deduce the existence of a strategy that guarantees the value in commutative POMDPs where the decision maker has no information on the state.

In Theorem 3.6 we prove that a two-player zero-sum stochastic game in which the set of states is a compact subset of $\mathbb{R}^m$, the sets of actions are finite, and each transition is a deterministic function that is 1-Lipschitz for the norm $\|.\|_1$, has a uniform value. We deduce the existence of the uniform value in commutative *state-blind* repeated games where at each stage the players learn the past actions played but not the state. In this case, we can define an auxiliary stochastic game on a compact set of states, which satisfies the assumptions of Theorem 3.6. Therefore, this auxiliary game has a uniform value and we deduce the existence of the uniform value in the original state-blind repeated game.

The paper is organized as follows. In Section 2, we introduce the formal definition of commutativity, the model of stochastic games, and the model of state-blind repeated games. In Section 3, we state the results. Section 4 is dedicated to several results on Markov Decision Processes. We first provide an example of a commutative deterministic MDP with a uniform value in pure strategies but no pure 0-optimal strategies. Then we prove Theorem 3.1. In Section 5, we focus on the results in the framework of stochastic games and the proof of Theorem 3.6. We first show that another widely studied class of games, called absorbing game, can be reformulated into the class of commutative games. Then we prove Theorem 3.6 and deduce the existence of the uniform value in commutative state-blind repeated games. Finally, we provide some extensions of Theorem 3.6. Especially, we show the existence of a uniform equilibrium in two-player non zero-sum state-blind commutative repeated games and m-player state-blind product-state commutative repeated games.

2 The model

When X is a non-empty set, we denote by $\Delta_f(X)$ the set of probabilities on X with finite support. When X is finite, we denote the set of probabilities on X by $\Delta(X)$ and by $\sharp X$ the cardinality of X. We will consider two types of games: stochastic games on a compact metric set X of states, denoted by $\Gamma = (X, I, J, q, g)$, and state-blind repeated games on a finite set K of states[1], denoted by $\Gamma^{sb} = (K, I, J, q, g)$. The sets of actions will always be finite. Finite sets will be given the discrete topology. We first define stochastic games and the notion of uniform value. We will then describe state-blind repeated games.

2.1 Commutative stochastic games

A two-player zero-sum stochastic game $\Gamma = (X, I, J, q, g)$ is given by a non-empty set of states X, two finite, non-empty sets of actions I and J, a reward function

[1] We use X to denote a general set of states and K to denote a finite set of states

$g : X \times I \times J \to [0,1]$ and a transition function $q : X \times I \times J \to \Delta_f(X)$.

Given an initial probability distribution $z_1 \in \Delta_f(X)$, the game $\Gamma(z_1)$ is played as follows. An initial state x_1 is drawn according to z_1 and announced to the players. At each stage $t \geq 1$, player 1 and player 2 choose simultaneously actions, $i_t \in I$ and $j_t \in J$. Player 2 pays to player 1 the amount $g(x_t, i_t, j_t)$ and a new state x_{t+1} is drawn according to the probability distribution $q(x_t, i_t, j_t)$. Then, both players observe the action pair (i_t, j_t) and the state x_{t+1}. The game proceeds to stage $t + 1$. When the initial distribution is a Dirac mass at $x_1 \in X$, denoted by δ_{x_1}, we denote by $\Gamma(x_1)$ the game $\Gamma(\delta_{x_1})$.

If for every initial state and every action pair, the image of q is a Dirac measure, q is said to be *deterministic*.

Note that we assume that the transition maps to the set of probabilities with finite support on X: given a stage, a state and an action pair, there exists a finite number of states possible at the next stage. We equip X with any σ-algebra $\mathcal{X}$ that includes all countable sets. When (X, d) is a metric space, the Borel σ-algebra suffices.

For all $i \in I$ and $j \in J$ we extend $q(\cdot, i, j)$ and $g(\cdot, i, j)$ linearly to $\Delta_f(X)$ by

$$\forall z \in \Delta_f(X),\ \widetilde{q}(z, i, j) = \sum_{x \in X} z(x) q(x, i, j) \text{ and } \widetilde{g}(z, i, j) = \sum_{x \in X} z(x) g(x, i, j).$$

Definition 2.1 *The transition q is commutative on X if for all $x \in X$, for all $i, i' \in I$ and for all $j, j' \in J$,*

$$\widetilde{q}(q(x, i, j), i', j') = \widetilde{q}(q(x, i', j'), i, j).$$

That is, the distribution over the state after two stages is equal whether action pair (i, j) is played before (i', j') or whether (i, j) is played after (i', j'). Note that if the transition q is not deterministic, $\widetilde{q}(q(x, i', j'), i, j)$ is the law of a random variable x'' computed in two steps: x' is randomly chosen with law $q(x, i', j')$, then x'' is randomly chosen with law $q(x', i, j)$; specifically (i, j) is played at the second step independently of the outcome of x'.

Remark 2.2 If the transition q does not depend on the actions, then the state process is a Markov chain and the commutativity assumption is automatically fulfilled.

Example 2.3 Let X be the set of complex numbers of modulus 1 and $\alpha : I \times J \to \Delta_f([0, 2\pi])$. Let q be defined by

$$\forall x \in X, \forall a \in I, \forall b \in J,\ q(x, a, b) = \sum_{\rho \in [0, 2\pi]} \alpha(a, b)(\rho) \delta_{x e^{i\rho}}.$$

If the state is x and the action pair (a, b) is played, then the new state is $x' = x e^{i\rho}$ with probability $f(a, b)(\rho)$. This transition is commutative by the commutativity of multiplication of complex numbers.

The next example originates in the theory of repeated games with incomplete information on one side (Aumann and Maschler [AM95]).

Example 2.4 A repeated game with incomplete information on one side, Γ, is defined by a finite family of matrices $(G^k)_{k\in K}$, two finite sets of actions I and J, and an initial probability p_1. At the outset of the game, a matrix G^k is randomly chosen with law p_1 and told to player 1 whereas player 2 only knows p_1. Then, the matrix game G^k is repeated over and over. The players observe the actions played but not the payoff.

One way to study Γ is to introduce a stochastic game on the posterior beliefs of player 2 about the state. Knowing the strategy played by player 1, player 2 updates his posterior belief depending on the actions observed. Let $\Psi = (X, A, B, \widetilde{q}, \widetilde{g})$ be a stochastic game where $X = \Delta(K)$, $A = \Delta(I)^K$ and $B = \Delta(J)$, the payoff function is given by

$$\widetilde{g}(p, a, b) = \sum_{k\in K, i\in I, j\in J} p^k a^k(i) b(j) G^k(i, j),$$

and the transition by

$$\widetilde{q}(p, a, b) = \sum_{k\in K, i\in I} a^k(i) \delta_{\hat{p}(a|i)},$$

where $a(i) = \sum_{k\in K} p^k a^k(i)$ and $\hat{p}(a|i) = \left(\frac{p^k a^k(i)}{a(i)}\right)_{k\in K} \in \Delta(K)$. Knowing the mixed action chosen by player 1 in each state, a, and having a prior belief, p, player 2 observes action i with probability $a(i)$ and updates his beliefs by Bayes rule to $\hat{p}(a|i)$. This induces the auxiliary transition $\widetilde{q}$. The payoff $\widetilde{g}$ is the expectation of the payoff under the probability generated by player 2's belief and player 1's mixed action.

We now check that the auxiliary stochastic game is commutative. Note that the second player does not influence the transition so we can ignore him. Let a and a' be two actions of player 1 and p be a belief of player 2. If player 1 plays first a and player 2 observes action i, then player 2's belief $p_2(\cdot|i)$ is given by

$$\forall k \in K,\ p_2(k|i) = \frac{p^k a^k(i)}{\sum_{k\in K} p^k a^k(i)}.$$

If now player 1 plays a' and player 2 observes i', then player 2's belief $p_3(\cdot|i, i')$ is given by

$$\forall k \in K,\ p_3(k|i, i') = \frac{p_2(k|i) a'^k(i')}{\sum_{k\in K} p_2(k|i) a'^k(i')} = \frac{p^k a'^k(i') a^k(i)}{\sum_{k\in K} p^k a'^k(i') a^k(i)}.$$

The probability that the action pair (i, i') is observed is $p^k a^k(i) a'^k(i')$. Since the belief p_3 and the probability to observe each pair (i, i') are symmetric in (a, i) and (a', i'), the transition is commutative.

Remark 2.5 Both previous examples are commutative but the transition is not deterministic.

Remark 2.6 Commutativity of the transitions implies that if we consider an initial state x and a finite sequence of actions $(i_1, j_1, ..., i_n, j_n)$, then the law of the state at stage $n+1$ does not depend on the order in which the action pairs (i_t, j_t), $t = 1, ..., n$, are played. We can thus represent a finite sequence of actions by a vector in $\mathbb{N}^{I \times J}$ counting how many times each action pair is played. Other models in which the transition along a sequence of actions is only a function of a parameter in a smaller set have been studied in the literature. For example, a transition is state independent (SIT) if it does not depend on the state. The law of the state at stage n is characterized only by the last action pair played. The law then depends on the order in which actions are played. Thuijsman [Thu92] proved the existence of stationary optimal strategies in this framework.

2.2 Uniform value

At stage t, the space of past *histories* is $H_t = (X \times I \times J)^{t-1} \times X$. Set $H_\infty = (X \times I \times J)^\infty$ to be the space of infinite *plays.* For every $t \geq 1$, we consider the product topology on H_t, $t \geq 1$ and also on H_∞.

A *(behavioral) strategy* for player 1 is a sequence $(\sigma_t)_{t \geq 1}$ of functions $\sigma_t : H_t \to \Delta(I)$. A *(behavorial) strategy* for player 2 is a sequence $\tau = (\tau_t)_{t \geq 1}$ of functions $\tau_t : H_t \to \Delta(J)$. We denote by Σ and $\mathcal{T}$, the player's respective sets of strategies.

Note that we did not make any measurability assumption on the strategies. Given $x_1 \in X$, the set of histories at stage t from state x_1 is finite since the image of the transition q is contained in the set of probabilities over X with finite support and the sets of actions are finite. It follows that any triplet (z_1, σ, τ) defines a probability over H_t without an additional measurability condition. This sequence of probabilities can be extended to a unique probability denoted $\mathbb{P}_{z_1,\sigma,\tau}$ over the set H_∞ with the infinite product topology. We denote by $\mathbb{E}_{z_1,\sigma,\tau}$ the expectation with respect to the probability $\mathbb{P}_{z_1,\sigma,\tau}$.

If for every $t \geq 1$ and every history $h \in H_t$ the image of $\sigma_t(h_t)$ is a Dirac measure, the strategy is said to be *pure.* If the initial distribution is a Dirac measure, the transition is deterministic and both players use pure strategies, then $\mathbb{P}_{z_1,\sigma,\tau}$ is a Dirac measure. The strategies induce a unique play.

The game we described is a game with perfect recall, so that by Kuhn's theorem [Kuh53] every behavior strategy is equivalent to a probability over pure strategies, called *mixed strategy*, and vice versa.

We are going to focus on two types of evaluations, the n-stage expected payoff and the expected average payoff between two stages m and n. For each positive integer n, the expected average payoff for player 1 up to stage n, induced by the strategy pair (σ, τ) and the initial distribution z_1, is given by

$$\gamma_n(z_1, \sigma, \tau) = \mathbb{E}_{z_1,\sigma,\tau}\left(\frac{1}{n}\sum_{t=1}^{n} g(x_t, i_t, j_t)\right).$$

The expected average payoff between two stages $1 \leq m \leq n$ is given by

$$\gamma_{m,n}(z_1, \sigma, \tau) = \mathbb{E}_{z_1,\sigma,\tau}\left(\frac{1}{n-m+1}\sum_{t=m}^{n} g(x_t, i_t, j_t)\right).$$

To study the infinitely repeated game $\Gamma(z_1)$, we focus on the notion of uniform value and on the notion of ε-optimal strategies.

Definition 2.7 *Let v be a real number.*

- *Player* 1 *can* guarantee v *in* $\Gamma(z_1)$ *if for all* $\varepsilon > 0$ *there exists a strategy* $\sigma^* \in \Sigma$ *of player* 1 *such that*

$$\liminf_n \inf_{\tau \in \mathcal{T}} \gamma_n(z_1, \sigma^*, \tau) \geq v - \varepsilon.$$

  We say that such a strategy σ^* guarantees $v - \varepsilon$ *in* $\Gamma(z_1)$.

- *Player* 2 *can* guarantee v *in* $\Gamma(z_1)$ *if for all* $\varepsilon > 0$ *there exists a strategy* $\tau^* \in \mathcal{T}$ *of player* 2 *such that*

$$\limsup_n \sup_{\sigma \in \Sigma} \gamma_n(z_1, \sigma, \tau^*) \leq v + \varepsilon.$$

  We say that such a strategy τ^* guarantees $v + \varepsilon$ *in* $\Gamma(z_1)$.

- *If both players can guarantee v, then v is called the* uniform value *of the game* $\Gamma(z_1)$ *and denoted by* $v^*(z_1)$. *A strategy σ (resp. τ) that guarantees $v^*(z_1) - \varepsilon$ (resp. $v^*(z_1) + \epsilon$) with $\varepsilon \geq 0$ is called* ε-optimal.

Remark 2.8 For each $n \geq 1$ the triplet $(\Sigma, \mathcal{T}, \gamma_n(z_1, ., .))$ defines a game in strategic form. This game has a value, denoted by $v_n(z_1)$. If the game $\Gamma(z)$ has a uniform value $v^*(z_1)$, then the sequence $(v_n(z_1))_{n\geq 1}$ converges to $v^*(z_1)$.

Remark 2.9 Let us make several remarks on another way to evaluate the infinite stream of payoffs. Let $\lambda \in (0, 1]$. The expected λ-discounted payoff for player 1, induced by a strategy pair (σ, τ) and the initial distribution z_1, is given by

$$\gamma_\lambda(z_1, \sigma, \tau) = \mathbb{E}_{z_1,\sigma,\tau}\left(\lambda \sum_{t=1}^{\infty} (1-\lambda)^{(t-1)} g(x_t, i_t, j_t)\right).$$

For each $\lambda \in (0, 1]$, the triplet $(\Sigma, \mathcal{T}, \gamma_\lambda(z_1, ., .))$ also defines a game $\Gamma_\lambda(z_1)$ in strategic form. The sets of strategies are compact for the product topology, and the payoff function $\gamma_\lambda(z_1, \sigma, \tau)$ is continuous. Using Kuhn's theorem, the payoff is also concave-like, convex-like and it follows therefore from Fan's minimax theorem (see [Fan53]) that the game $\Gamma_\lambda(z_1)$ has a value, denoted $v_\lambda(z_1)$. Note that there may not exist an optimal measurable strategy which depends only on the current state (Levy [Lev12]).

Some authors focus on the existence of $v(z_1)$ such that

$$\lim_{n\to\infty} v_n(z_1) = \lim_{\lambda \to 0} v_\lambda(z_1) = v(z_1).$$

When the uniform value exists, this equality is immediately true with $v(z_1) = v^*(z_1)$ since the discounted payoff can be written as a convex combination of expected average payoffs.

2.3 The model of repeated games with state-blind players

A *state-blind* repeated game $\Gamma^{sb} = (K, I, J, q, g)$ is defined by the same objects as a stochastic game. The definition of commutativity is the same. The main difference is the information that the players have, which affects their sets of strategies. We assume that at each stage, the players observe the actions played but not the state. We will restrict the discussion to a finite state space K.

Given an initial probability $p_1 \in \Delta(K)$, the game $\Gamma^{sb}(p_1)$ is played as follows. An initial state k_1 is drawn according to p_1 without being announced to the players. At each stage $t \geq 1$, player 1 and player 2 choose simultaneously an action, $i_t \in I$ and $j_t \in J$. Player 1 receives the (unobserved) payoff $g(k_t, i_t, j_t)$, player 2 receives the (unobserved) payoff $-g(k_t, i_t, j_t)$, and a new state k_{t+1} is drawn according to the probability distribution $q(k_t, i_t, j_t)$. Both players then observe only the action pair (i_t, j_t) and the game proceeds to stage $t + 1$.

Since the states are not observed, the space of public histories of length t is $H_t^{sb} = (I \times J)^{t-1}$. A strategy of player 1 in Γ^{sb} is a sequence $(\sigma_t)_{t\geq 1}$ of functions $\sigma_t : H_t^{sb} \to \Delta(I)$, and a strategy of player 2 is a sequence $\tau = (\tau_t)_{t\geq 1}$ of functions $\tau_t : H_t^{sb} \to \Delta(J)$. We denote by Σ^{sb} and $\mathcal{T}^{sb}$ the players respective sets of strategies. An initial distribution p_1 and a pair of strategies $(\sigma, \tau) \in \Sigma^{sb} \times \mathcal{T}^{sb}$ induce a unique probability over the infinite plays H_∞. For every pair of strategies (σ, τ) and initial probability p_1 the payoff is defined as in Section 2.2. Similarly, the notion of uniform value is defined as in Definition 2.7 by restricting the players to play strategies in Σ^{sb} and $\mathcal{T}^{sb}$.

Definition 2.10 *Let v be a real number.*

- *Player* 1 *can* guarantee *v in $\Gamma^{sb}(p_1)$ if for all $\varepsilon > 0$ there exists a strategy $\sigma^* \in \Sigma^{sb}$ of player* 1 *such that*

$$\liminf_n \inf_{\tau\in\mathcal{T}^{sb}} \gamma_n(p_1, \sigma^*, \tau) \geq v - \varepsilon.$$

  We say that such a strategy σ^* guarantees *$v - \varepsilon$ in $\Gamma^{sb}(p_1)$.*

- *Player* 2 *can* guarantee *v in $\Gamma^{sb}(p_1)$ if for all $\varepsilon > 0$ there exists a strategy $\tau^* \in \mathcal{T}^{sb}$ of player* 2 *such that*

$$\limsup_n \sup_{\sigma\in\Sigma^{sb}} \gamma_n(p_1, \sigma, \tau^*) \leq v + \varepsilon.$$

  We say that such a strategy τ^* guarantees *$v + \varepsilon$ in $\Gamma^{sb}(p_1)$.*

- *If both players can guarantee v, then v is called the* uniform value *of the game $\Gamma^{sb}(p_1)$ and denoted by $v^{sb}(p_1)$.*

Remark 2.11 The sets Σ^{sb} and $\mathcal{T}^{sb}$ can be seen as subsets of Σ and $\mathcal{T}$ respectively. There is no relation between $v^{sb}(p_1)$ and $v^*(p_1)$, since both players have restricted sets of strategies.

3 Results.

In this section we present the main results of the paper. Section 3.1 concerns MDPs and Section 3.2 concerns stochastic games.

3.1 Existence of 0-optimal strategies in Commutative deterministic Markov Decision Processes.

An MDP is a stochastic game, $\Gamma = (X, I, q, g)$, with a single player, that is, the set J is a singleton. Our first main result states that if an MDP with deterministic and commutative transitions has a uniform value and if the decision maker has pure ϵ-optimal strategies, then he also has a (not necessarily pure) 0-optimal strategy. We also provide sufficient topological conditions for the existence of a pure 0-optimal strategy.

Theorem 3.1 *Let $\Gamma = (X, I, q, g)$ be an MDP such that I is finite and q is deterministic and commutative.*

1. *If for all $z_1 \in \Delta_f(X)$, $\Gamma(z_1)$ has a uniform value in pure strategies, then for all $z_1 \in \Delta_f(X)$ there exists a 0-optimal strategy.*
2. *If X is a precompact metric space, $q(\cdot, i)$ is 1-Lipschitz for every $i \in I$, and $g(\cdot, i)$ is uniformly continuous for every $i \in I$, then for all $z_1 \in \Delta_f(X)$ the game $\Gamma(z_1)$ has a uniform value and there exists a 0-optimal pure strategy.*

Remark 3.2 In an MDP with a deterministic transition, a play is uniquely determined by the initial state and a sequence of actions. Thus, in the framework of deterministic MDPs we will always identify the set of pure strategies with the set of sequences of actions.

The first part of Theorem 3.1 is sharp in the sense that a commutative deterministic MDP with a uniform value in pure strategies may have no 0-optimal pure strategy. An example is described at the beginning of Section 4.

The topological assumptions of the second part of Theorem 3.1 were first introduced by Renault [Ren11] and imply the existence of the uniform value in pure strategies; by the first part of the theorem they also imply the existence of a 0-optimal strategy. Under these topological assumptions, we prove the stronger result of the existence of a 0-optimal pure strategy.

Let us now discuss the topological assumptions made in Theorem 3.1. First, if the payoff function g is only continuous or the state space is not precompact, then the uniform value may fail to exist as shown in the following example.

Example 3.3 Consider a Markov Decision Process (X, I, q, g) where there is only one action, $|I| = 1$. The set of states is the set of integers, $X = \mathbb{N}$, and the transition is given by $q(n) = n + 1$, $\forall n \in \mathbb{N}$. Note that q is commutative and deterministic.

Let $r = (r_n)_{n\in\mathbb{N}}$ be a sequence of numbers in $[0,1]$ such that the sequence of average payoffs does not converge. The payoff function is defined by $g(n) = r_n, \forall n \in \mathbb{N}$.

We consider the following metric on $\mathbb{N}$: for all $n, m \in \mathbb{N}$, $d(n,m) = \mathbb{1}_{n\neq m}$. Then $(\mathbb{N}, d)$ is not precompact, the transition q is 1-Lipschitz, and the function g is uniformly continuous. The choice of r implies that the MDP $\Gamma = (X, I, q, g)$ has no uniform value.

Consider now the following metric on $\mathbb{N}$: for all $n, m \in \mathbb{N}$, $d'(n,m) = |\frac{1}{n+1} - \frac{1}{m+1}|$. Then $(\mathbb{N}, d')$ is a precompact metric space, the transition is 1-Lipschitz and the function g is continuous. As before, the MDP $\Gamma = (X, I, q, g)$ has no uniform value. A simple computation shows that the function g is not uniformly continuous on $(\mathbb{N}, d')$. Take now g a uniformly continuous function, then $(g(n))_{n\in\mathbb{N}}$ is a Cauchy sequence in a complete space, thus converges. It follows that the sequence of Cesàro averages also converges to the same limit and the game has a uniform value.

The assumption that q is 1-Lipschitz may seem strong but turns out to be necessary in the proof of Renault [Ren11]. The reason is as follows. When computing the uniform value, one considers infinite histories. When q is 1-Lipschitz, given two states x and x' and an infinite sequence of actions $(i_1, ..., i_t, ...)$, the state at stage t on the play from x and the state at stage t on the play from x' are at a distance at most $d(x, x')$. Thus, the payoffs along both plays stay close at every stage. On the contrary, if q were say 2-Lipschitz, we only know that the distance between the state at stage t on the play from x and the state at stage t on the play from x' is at most $d(x, x')2^t$, which gives no uniform bound on the difference between the stage payoffs along the two plays. As shown in Renault [Ren11] when q is not 1-Lipschitz, the value may fail to exist. The counter-example provided by Renault is not commutative and it might be that the additional assumption of commutativity can help us in relaxing the Lipschitz requirement on q. In our proof, we use the fact that q is 1-Lipschitz at two steps: first in order to apply the result of Renault [Ren11] and then in order to concatenate strategies. It is still open whether one of these two steps can be done under the weaker assumption that q is uniformly continuous.

We list now two open problems: assume that the uniform value exists, X is precompact, g is uniformly continuous, and q is uniformly continuous, deterministic, and commutative; does there exist a 0-optimal strategy? Does an MDP with X precompact, g uniformly continuous, and q uniformly continuous, deterministic, and commutative always have a uniform value?

We deduce from Theorem 3.1 the existence of a 0-optimal strategy for commutative POMDPs with no information on the state, called *MDPs in the dark* in the literature. The auxiliary MDP associated to the POMDP is deterministic and commutative, and thus it satisfies the assumption of Theorem 3.1.

Corollary 3.4 *Let $\Gamma^{sb} = (K, I, q, g)$ be a commutative state-blind POMDP with a finite state space K and a finite set of actions I. For all $p_1 \in \Delta(K)$, the POMDP $\Gamma^{sb}(p_1)$ has a uniform value and there exists a 0-optimal pure strategy.*

We will prove Corollary 3.4 in the two-player framework.

Rosenberg, Solan, and Vieille [RSV02] asked if a 0-optimal strategy exists in POMDPs. Theorem 3.1 ensures that if the transition is commutative such a strategy exists. The following example, communicated by Hugo Gimbert, shows that it is not true without the commutativity assumption. The example also implies that there exist games that cannot be transformed into a commutative game with finite sets of actions.

Example 3.5 Consider a state-blind POMDP $\Gamma^{sb} = (X, I, g, q)$ defined as follows. There are four states $X = \{\alpha, \beta, k_0, k_1\}$ and two actions $I = \{T, B\}$. The payoff is 0 except in state k_1 where it is 1. The states k_0 and k_1 are absorbing and the transition function q is given on the other states by

$$\begin{aligned}
q(\alpha, T) &= \frac{1}{2}\delta_\alpha + \frac{1}{2}\delta_\beta, \\
q(\beta, T) &= \delta_\beta, \\
q(\alpha, B) &= \delta_{k_0}, \\
q(\beta, B) &= \delta_{k_1}.
\end{aligned}$$

This POMDP is not commutative: if the initial state is α and the decision maker plays B and then T, the state is k_0 with probability one, whereas if he plays first T and then B, the state is k_0 with probability 1/2 and k_1 with probability 1/2.

Let us check that this game has a uniform value in pure strategies, but no 0-optimal strategies. An ε-optimal strategy in $\Gamma(\alpha)$ is to play the action T until the probability to be in β is more than $1 - \varepsilon$, and then to play B. This leads to a payoff of $1 - \varepsilon$, so the uniform value in α exists and is equal to 1. The reader can verify that there is no strategy that guarantees 1 in $\Gamma(\alpha)$.

3.2 Existence of the uniform value in commutative deterministic stochastic games.

For two-player stochastic games, the commutativity assumption does not imply the existence of 0-optimal strategies. Indeed, we will prove in Proposition 5.1 that any absorbing game can be reformulated as a commutative stochastic game. Since there exist absorbing games with deterministic transitions without 0-optimal strategies, for example the Big Match (see Blackwell and Ferguson [BF68]), there exist deterministic commutative stochastic games with a uniform value and without 0-optimal strategies. In this section, we study the existence of the uniform value in one class of stochastic games on $\mathbb{R}^m$.

Theorem 3.6 *Let $\Gamma = (X, I, J, q, g)$ be a stochastic game where X is a compact subset of $\mathbb{R}^m$, I and J are finite sets, q is commutative, deterministic and 1-Lipschitz for the norm $\|.\|_1$, and g is continuous. Then for all $z_1 \in \Delta_f(X)$ the stochastic game $\Gamma(z_1)$ has a uniform value.*

Let us comment on the assumptions of Theorem 3.6. The state space is not finite yet the set of actions available to each player is the same in all states. This requirement is necessary to ensure that the commutativity property is well defined. Our proof is valid only if q is 1-Lipschitz with respect to the norm $\|.\|_1$. Thus this theorem does not apply to Example 2.3 on the circle. The proof can be adapted for polyhedral norms (i.e. such that the unit ball has a finite number of extreme points), this is further discussed in Section 5.4. Finally note that the most restrictive assumptions are on the transition.

As shown in the MDP framework, the assumption that q is 1-Lipschitz is important for the existence of a uniform value and is used in the proof at two steps. First, we use it to deduce that for all $(i, j) \in I \times J$, iterating infinitely often the action pair (i, j) leads to a limit cycle with a finite number of states. Second, it is used to prove that if a strategy guarantees w from a state x then it guarantees almost w in any game that starts at an initial state in a small neighbourhood of x.

Given a state-blind repeated game $\Gamma^{sb} = (K, I, J, q, g)$ with a commutative transition q, we define the auxiliary stochastic game $\Psi = (X, I, J, \widetilde{q}, \widetilde{g})$ where $X = \Delta(K)$, $\widetilde{q}$ is the linear extension of q, and $\widetilde{g}$ is the linear extension of g.

Since K is finite, X can be embedded in $\mathbb{R}^K$ and the transition $\widetilde{q}$ is deterministic, 1-Lipschitz for $\|.\|_1$, and commutative. Furthermore, $\widetilde{g}$ is continuous and therefore we can apply Theorem 3.6 to Ψ. It follows that for each initial state $p_1 \in X$, $\Psi(p_1)$ has a uniform value. We will check that it is the uniform value of the state-blind repeated game $\Gamma^{sb}(p_1)$ and deduce the following corollary.

Corollary 3.7 *Let $\Gamma^{sb} = (K, I, J, q, g)$ be a commutative state-blind repeated game with a finite set of states K and finite sets of actions I and J. For all $p_1 \in \Delta(K)$, the game $\Gamma^{sb}(p_1)$ has a uniform value.*

Remark 3.8 Corollary 3.7 concerns repeated games where the players observe past actions but not the state. The more general model, where the players observe past actions and have a public signal on the state, leads to the definition of an auxiliary stochastic game with a random transition. In the deterministic case given a triplet (z_1, σ, τ), the sequence of states visited along each infinite play converges to a finite cycle $\mathbb{P}_{z_1,\sigma,\tau}$-almost surely. This no longer holds if the transition is random.

We now present an example of a commutative state-blind repeated game and its auxilliary deterministic stochastic game.

Example 3.9 Let $K = \mathbb{Z}/m\mathbb{Z}$ and f be a function from $I \times J$ to $\Delta(K)$. We define the transition $q : K \times I \times J \to \Delta(K)$ as follows: given a state $k \in K$, if the players play (i, j) then for all $k' \in K$, the new state is $k + k'$ with probability $f(i, j)(k')$.

If the initial state is drawn with a distribution p and players play respectively i and j, then the new state is given by the sum of two independent random variables of respective laws p and $f(i, j)$. The addition of independent random variables is a commutative and associative operation, therefore q is commutative on K.

For example, let $m = 3$, $I = \{T, B\}$, $J = \{L, R\}$ and the function f be given by

$$\begin{array}{ccc} & L & R \\ T & \frac{1}{2}\delta_1+\frac{1}{2}\delta_2 & \delta_1 \\ B & \delta_1 & \delta_0 \end{array}.$$

If the players play (T, L) then the new state is one of the two other states with equal probability. If the players play (B, R), then the state does not change. And otherwise the state goes from state k to state $k+1$.

The extension of the transition function to the set of probabilities over K is given by

$$\begin{aligned} \widetilde{q}((p^1,p^2,p^3),T,L) &= \left(\frac{p^2+p^3}{2},\frac{p^1+p^3}{2},\frac{p^1+p^2}{2}\right), \\ \widetilde{q}((p^1,p^2,p^3),B,R) &= (p^1,p^2,p^3), \\ \widetilde{q}((p^1,p^2,p^3),B,L)=\widetilde{q}((p^1,p^2,p^3),T,R) &= (p^3,p^1,p^2). \end{aligned}$$

4 Existence of 0-optimal strategies in commutative deterministic MDPs.

In this section we focus on MDPs and Theorem 3.1. The section is divided into four parts. In the first part we provide an example showing that under the conditions of Theorem 3.1(1), there need not exist a pure 0-optimal strategy.

The rest of the section is dedicated to the proof of Theorem 3.1. In the second part we show that in a deterministic commutative MDP, for all $\varepsilon > 0$, there exist ε-optimal pure strategies such that the uniform value is constant on the play. Along these strategies, the decision maker ensures that when balancing between current payoff and future states, he is not making irreversible mistakes, in the sense that the uniform value does not decrease along the induced play.

In the third part we prove Theorem 3.1(1). To prove the existence of a (non-pure) 0-optimal strategy, we first show the existence of pure strategies such that the lim sup of the long run expected average payoffs is the uniform value. Nevertheless the payoffs may not converge along the play induced by these strategies. We show that the decision maker can choose a proper distribution over these strategies to ensure that the expected average payoff converges.

The fourth part is dedicated to the proof of Theorem 3.1(2). To construct a pure 0-optimal strategy, instead of concatenating strategies one after the other, as is often done in the literature, we define a sequence of strategies $(\sigma^l)_{l\geq 1}$ such that for every $l \geq 1$, σ^l guarantees $v^*(x^l) - \varepsilon_l$ where $x^l \in X$ and ε_l is a positive real number. We then split these strategies, seen as sequences of actions, into blocks of actions and construct a 0-optimal strategy by playing these blocks in a proper order.

4.1 An example of a commutative deterministic MDP without 0-optimal pure strategies

In this section, we provide an example of a commutative deterministic MDP with a uniform value in pure strategies that does not have a pure 0-optimal strategy. Before going into the details, we outline the structure of the example. The set of states, which is the countable set $\mathbb{N} \times \mathbb{N}$, is partitioned into countably many sets, $\{h^0, h^1, ...\}$, such that the payoff is constant on each element of the partition; the payoff is 0 on the set h^0 and $1 - \frac{1}{2^l}$ on the set h^l, for all $l \geq 1$. We will first check that for each $l \geq 1$, there exists a pure strategy from the initial state $(0,0)$ that eventually belongs to the set h^l. This will imply that the game starting at $(0,0)$ has a uniform value equal to 1. We will then prove that any 0-optimal pure strategy has to visit all sets h^l and that when switching from one set h^i to another set $h^{i'}$, the induced play has to spend many stages in the set h^0. The computation of the minimal number of stages spent in the set h^0 shows that the average expected payoff has to drop below $\frac{1}{2}$, which contradicts the optimality of the strategy. Thus, there exists no 0-optimal pure strategy in the game starting at state $(0,0)$.

Example 4.1 *The set of states is $X = \mathbb{N} \times \mathbb{N}$ and there are only two actions R and T; the action R increments the first coordinate and the action T increments the second one:*

$$q((x,y),R) = (x+1,y),$$
$$q((x,y),T) = (x,y+1).$$

Plainly the transition is deterministic and commutative.

For each $l \geq 1$, let $w^l = \sum_{m=1}^{l} \left(4^{m-1} - 1\right) = \frac{4^l - 1}{3} - l$. We define the set $h^l \subset X$ by

$$h^l = \{(w_l, 0)\} \cup \left\{(x,y),\ w^l + (y-1)\left(4^{l-1} - 1\right) \leq x \leq w^l + y\left(4^{l-1} - 1\right),\ x,y \geq 1\right\}.$$

For example

$$h^1 = \{(0,y),\ y \geq 0\},$$

and

$$h^2 = \{(3,0)\} \cup \left(\cup_{y \geq 1}\{(3y,y),(3y+1,y),(3y+2,y),(3y+3,y)\}\right).$$

For every $l \geq 1$, the set h^l is the set of states obtained along the play induced by the sequence of actions $(TR^{4^{l-1}-1})^\infty$ from state $(w^l, 0)$. We denote by h^0 the set of states not on any h^l, $l \geq 1$. Figure (1) shows the play associated to h^1, h^2, and h^3 with their respective payoffs. One can notice that the plays following these three sets separate from each other.

The payoff is $1 - \frac{1}{2^l}$ in every state on the set h^l and 0 on the set h^0 :

$$g(x,y) = \begin{cases} 1 - \frac{1}{2^l} \text{ if } x \in \left[w^l + (y-1)\left(4^{l-1} - 1\right), w^l + y\left(4^{l-1} - 1\right)\right] \\ 0 \text{ otherwise.} \end{cases}$$

The uniform value exists in every state, is equal to 1, and the decision maker has ε-optimal pure strategies: given an initial state, play R until reaching some state in h^l with $\frac{1}{2^l} \leq \varepsilon$ and then stay in h^l.

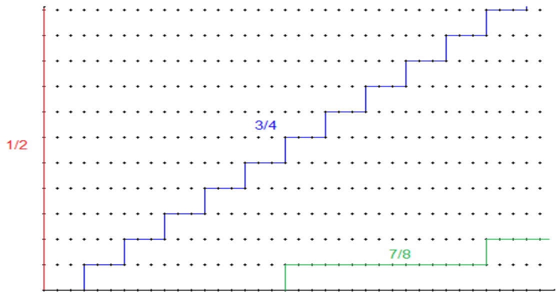

Figure 1: Payoff of the game on h^1,h^2 and h^3.

There exists no 0*-optimal pure strategy from state* $(0,0)$*. Since there is only one player, the transition is deterministic and the payoff depends only on the state, we can identify a pure strategy with the sequence of states it selects on the play that it defines. Let* $h = (z_1, ..., z_t, ...)$ *be a* 0*-optimal strategy. Since the play* h *guarantees* 1*, there exists an increasing sequence of stages* $(n_l)_{l\geq 1}$ *such that* h *crosses* h^l *at stage* n_l*.*

Let $m_l < n_{l+1}$ *be the last time before* n_{l+1} *such that* h *intersects* h^l*. The reader can check that then* $n_{l+1} - m_l \geq m_l$*, and every state of* h *between stage* $m_l + 1$ *and stage* n_{l+1} *is in* h^0*. Therefore the expected average payoff at stage* $n_{l+1} - 1$ *is less* $\frac{1}{2} \times 1 + \frac{1}{2} \times 0 = \frac{1}{2}$*.*

We now argue that there is a behavioral strategy that yields payoff 1*. We start by illustrating a strategy that yields an expected average payoff at least* $\frac{3}{8}$ *in all stages:*

- *with probability* $1/4$*, the decision maker plays* 3 *times Right in order to reach the set* h^2 *and then stays in* h^2*;*
- *with probability* $1/4$*, the decision maker stays in the set* h^1 *for* 3 *stages (*3 *times Top) then plays* 9 *times Right in order to reach the set* h^2 *and then stays in* h^2*;*
- *with probability* $1/4$*, the decision maker stays in the set* h^1 *for* $3 + 9 = 12$ *stages then plays* 36 *times Right in order to reach the set* h^2 *and then stays in* h^2*;*
- *with probability* $1/4$ *the decision maker stays in the set* h^1 *for* $3 + 9 + 36 = 48$ *stages then plays* 144 *times Right in order to reach the set* h^2 *and then stays in* h^2*.*

Note that the state at stage 192 *is on* h^2 *and more precisely equal to* $(48, 144)$ *whatever is the pure strategy chosen. The first strategy yields a payoff of* $3/4$ *except on the second and third times the decision maker is playing right (stage 2-3), the second one yields a payoff of* $1/2$ *before stage* 3 *(included) and a payoff of* $3/4$ *from stages* 13 *(included), the third one yields a payoff of* $1/2$ *before stage* 12 *(included) and a payoff of* $3/4$ *from stage* 49 *(included) and the fourth one yields a payoff of* $1/2$ *before stage* 48 *(included) and a payoff of* $3/4$ *from stages* 193 *(included).*

Thus for each stage n up to stage 192, there is at most one of the four pure strategies that gives a daily payoff of 0. The others strategies either stay in h^1 or stay in h^2, and thus yield a stage payoff at least 1/2. Therefore, the expected payoff at each stage is greater than $\frac{3}{4} * \frac{1}{2} = \frac{3}{8}$ and the expected average payoff until stage n is greater than 3/8 for every $n \geq 1$. We managed to build a strategy going from h^1 to the set h^2 such that the expected average payoff does not drop below $\frac{3}{8}$.

We can iterate and switch from h^2 to h^3 without the payoff dropping below $\frac{7}{8} * \frac{3}{4}$ by considering 8 different pure strategies. Repeating this procedure from h^l to h^{l+1} for every $l \geq 1$ will lead to a 0-optimal strategy. To define properly a strategy which ensures an expected payoff 1, we augment the strategy as in Section 4.3.

4.2 Existence of ε-optimal strategies with a constant value on the induced play

In this part, we consider a commutative deterministic MDP with a uniform value in pure strategies. We show that for all $x_1 \in X$ and all $\varepsilon > 0$ there exists an ε-optimal pure strategy in $\Gamma(x_1)$ such that the value is constant on the induced play. Lehrer and Sorin [LS92] showed that in deterministic MDPs, given a sequence of actions, the value is always non-increasing along the induced play. In particular, it is true along the play induced by an ε-optimal pure strategy. We need to define an ε-optimal pure strategy such that the value is non-decreasing.

To this end, we introduce a partial preorder on the set of states such that, if x' is greater than x, then x' can be *reached* from x, i.e. there exists a finite sequence of actions such that x' is on one play induced from x. Fix a state x_1 and let x be a state which can be reached from x_1. By commutativity, the order of actions is not relevant and we can represent the state x by a vector $m \in \mathbb{N}^I$, counting how many times each action has to be played in order to reach x from x_1. Let $M(x)$ be the set of all vectors representing the state x.

Given two vectors m and m' in $\mathbb{R}^I$, m is greater than m' if for every $i \in I$, $m(i) \geq m'(i)$. Given two states x and x', we say that x is greater than x', denoted $x \geq x'$, if there exists $m \in M(x)$ and $m' \in M(x')$, such that $m \geq m'$. By construction, $x \geq x'$ implies that x can be reached from the state x'. Indeed if x is greater than x', then there exists $m \in M(x)$ and $m' \in M(x')$ such that $m \geq m'$ in all coordinates. By playing $(m - m')(i)$ times the action i for every $i \in I$, the decision maker can reach the state x from x'.

Lemma 4.2 *Consider a commutative deterministic MDP with a uniform value in pure strategies. For all $x_1 \in X$ and all $\varepsilon > 0$ there exists an ε-optimal strategy in $\Gamma(x_1)$ such that the value is non-decreasing, thus constant, on the induced play.*

Proof: Fix $x_1 \in X$ and $\varepsilon > 0$. We construct a sequence of real numbers $(\varepsilon_l)_{l \geq 1}$ and a sequence of strategies $(\sigma^l)_{l \geq 1}$ satisfying three properties. For each $l \geq 1$, we denote by $(x_n^l)_{n\geq 1}$ the sequence of states along σ^l. First, the sequence $(\varepsilon_l)_{l\geq 1}$ is decreasing and $\varepsilon_1 \leq \varepsilon$ (property (i)). Second, for every $l \geq 1$ the strategy σ^l is ε_l-optimal in the game $\Gamma(x_1)$ (property (ii)). Finally, given any $l \geq 1$ and any stage $n \geq 1$, for every $l' \geq l$ there exists a stage n' such that $x_{n'}^{l'} \geq x_n^l$ (property (iii)). This implies that $x_{n'}^{l'}$ is reachable from x_n^l. Informally, a decision maker who follows the strategy σ^l can change his mind in order to play better: at any stage he can stop following σ^l, choose any $l' \geq l$, and play some actions such that the play merges eventually with the play induced by $\sigma^{l'}$.

Let $(\varepsilon_l)_{l\geq 1}$ be a decreasing sequence of positive numbers converging to 0 such that $\varepsilon_1 = \varepsilon$. For each $l \geq 1$, let σ^l be an ε_l-optimal pure strategy in $\Gamma(x_1)$. We identify σ^l with the sequence of actions $(i_1, i_2, ...)$ it induces. By construction, these sequences satisfy properties (i) and (ii). To satisfy property (iii), we extract a subsequence.

For all $l \geq 1$ and all $n \geq 1$, considering the strategy σ^l until stage n defines a vector $m_n(\sigma^l)$ in $M(x_n^l)$. The sequence $(m_n(\sigma^l))_{n\geq 1}$ is non-decreasing in every coordinate, so we can define the limit vector $m_\infty(\sigma^l) \in (\mathbb{N} \times \{\infty\})^I$. By definition of the limit, for any $w \in \mathbb{N}^I$ such that $w \leq m_\infty(\sigma^l)$, there exists some stage n such that $w \leq m_n(\sigma^l)$.

Since the number of actions is finite, we can choose a subsequence of $(m_\infty(\sigma^l))_{l\geq 1}$ such that each coordinate is non-decreasing in l. Informally, the closer to the value the decision maker wants the payoff to be the more he has to play each action. We keep the same notation, and denote by $(\varepsilon_l)_{l\geq 1}$ and $(\sigma^l)_{l\geq 1}$ the sequences after extraction.

After extraction ε_1 is smaller than ε. By definition, σ^l is ε_l-optimal in the game $\Gamma(x_1)$. Moreover, given two integers l, l' such that $1 \leq l \leq l'$, we have $m_\infty(\sigma^l) \leq m_\infty(\sigma^{l'})$. Let n be a positive integer, then

$$m_n(\sigma^l) \leq m_\infty(\sigma^l) \leq m_\infty(\sigma^{l'}).$$

By definition of $m_\infty(\sigma^{l'})$ as a limit, there exists a stage n' such that $m_{n'}(\sigma^{l'})$ is greater than $m_n(\sigma^l)$, and thus $x_{n'}^{l'}$ is greater than x_n^l. The subsequences $(\varepsilon_l)_{l\geq 1}$ and $(\sigma_l)_{l\geq 1}$ satisfy all the properties $(i) - (iii)$.

We now deduce that the value along σ_1 is non decreasing: for every $n \geq 1$, the uniform value in state x_n^1 is equal to the uniform value in the initial state. Fix $n \geq 1$ and $l' \geq 1$. By construction, there exists $n' \geq n$ such that $x_{n'}^{l'}$ can be reached from state x_n^1. Applying Lehrer and Sorin [LS92], we know that the value is non increasing along plays so $v^*(x_n^1) \geq v^*(x_{n'}^{l'})$. Moreover, the strategy $\sigma^{l'}$ defines a continuation strategy from $x_{n'}^{l'}$, which yields an average long-run payoff of at least $v^*(x_1) - \varepsilon_{l'}$. Thus, the uniform value along the play induced by $\sigma_{l'}$ does not drop below $v^*(x_1) - \varepsilon_{l'}$:

$$v^*(x_{n'}^{l'}) \geq v^*(x_1) - \varepsilon_{l'}.$$

Considering both results together, we obtain that

$$v^*(x_n^1) \geq v^*(x_{n'}^{l'}) \geq v^*(x_1) - \varepsilon_{l'}.$$

Since it is true for every $l' \geq 1$, we deduce that the value is non decreasing along σ_1. In order to conclude, notice that $\varepsilon_1 \leq \varepsilon$, therefore σ_1 is ε-optimal. □

4.3 Proof of Theorem 3.1(1)

In this subsection, we prove Theorem 3.1(1): in every commutative MDP with a uniform value in pure strategies, there exists a 0-optimal strategy.

A strategy σ is said to be *partially* 0*-optimal* if the limsup of the sequence of expected average payoffs is equal to the uniform value: $\limsup_n \gamma_n(x_1,\sigma) = v^*(x_1)$. We first deduce from Lemma 4.2 the existence of partially 0-optimal pure strategies. As shown in Example 4.1, expected average payoffs may not converge along partially 0-optimal strategy and, in particular, can be small infinitely often. The key point of the proof of Theorem 3.1(1) is that different partially 0-optimal strategies have bad expected average payoff at different stages. By choosing a proper mixed strategy that is supported by pure partially 0-optimal strategies, we can ensure that, at each stage, the probability to play one pure strategy with a bad expected average payoff is small.

We will first provide the formal definition of partially 0-optimal strategies and the concatenation of a sequence of strategies along a sequence of stopping times. Then, we define two specific sequences such that the concatenated strategy σ^* is 0-optimal. The proof of the optimality of σ^* is done in two steps: we check that the support of σ^* is included in the set of partially 0-optimal strategies, and that the probability to play a strategy with a bad expected average payoff at stage n converges to 0 for n sufficiently large.

We now start the proof of Theorem 3.1(1) by defining formally a partially 0-optimal strategy.

Definition 4.3 *Let $\Gamma = (X, I, q, g)$ be an MDP and $v^*(x_1)$ be the uniform value of the MDP starting at x_1. A strategy σ is* partially 0-optimal *if*

$$\limsup_n \gamma_n(x_1,\sigma) = v^*(x_1).$$

That is, for every $\varepsilon > 0$, the long run expected average payoff is greater than $v^*(x_1) - \varepsilon$ infinitely often.

We define the concatenation of strategies with respect to a sequence of stopping times [2]. Let $(u_l)_{l\geq 2}$ be a sequence of increasing stopping times and $(\sigma_l)_{l\geq 1}$ be a sequence of strategies. The concatenated strategy σ^* is defined as follows. For every $t \geq 1$ and every $h_t = (x_1, i_1, j_1, ..., x_t)$, let $l^* = l^*(h_t) = \sup\{l, u_l(h_t) \leq t\}$ and $\sigma^*(h_t) = \sigma_{l^*}(h_t^{u_l^*})$ where $h_t^{u_l^*} = (x_{u_l^*}, i_{u_l^*}, j_{u_l^*} ..., x_t)$. Informally, for every $l \geq 2$, at stage u_l the decision maker forgets the past history and follows σ_l.

<u>Definition of the 0-optimal strategy:</u> Fix $x_1 \in X$. For every $t \geq 1$, we denote by $X(t)$ the set of states which can be reached from x_1 in less than t stages. Since the transition is deterministic and the number of actions is finite, the set $X(t)$ is finite for every $t \geq 1$. We choose two specific sequences of stopping times and strategies and

[2] A stopping time u is a random variable such that the event $\{u \leq n\}$ is measurable with respect to the history up to stage n

denote by σ^* the concatenation. Let $(\varepsilon_l)_{l\geq 1}$ be a decreasing sequence of real numbers converging to 0. For each $x \in X$ and every integer $l \geq 1$, we denote by $\sigma_l(x)$ an ε_l-optimal strategy in $\Gamma(x)$ such that the uniform value is constant on the play, and let $N(l,x)$ be an integer that satisfies

$$\forall n \geq N(l,x),\ \gamma_n(x, \sigma_l(x)) \geq v^*(x) - \varepsilon_l. \tag{1}$$

In any games longer than $N(l,x)$ stages, the average expected payoff is close to the value, but the payoff in shorter games is not controlled. The strategy $\sigma_l(x)$ exists by Lemma 4.2.

We now define the sequence of stopping times. For every $l \geq 1$, we define a set of stages T_l and let u_l be a stopping time uniformly distributed over T_l. Start by setting $t_1 = 1$ and $T_1 = \{1\}$. Let $l \geq 1$ and assume that the set T_l is already defined. Denote $t_{l+1} = \left[\frac{1}{\varepsilon_{l+1}}\right] + 1$ and define the set $T_{l+1} = \{T_{l+1}^{(1)}, ..., T_{l+1}^{(t_{l+1})}\}$ by induction:

$$\begin{aligned}
T_{l+1}^{(1)} &= T_l^{(t_l)} + \max_{x \in X(T_l^{(t_l)})} N(l,x) + \left[\frac{1}{\varepsilon_l} + 1\right] T_l^{(t_l)}, \\
T_{l+1}^{(2)} &= T_{l+1}^{(1)} + \max_{x \in X(T_{l+1}^{(1)})} N(l+1,x), \\
&..., ... \\
T_{l+1}^{(t_{l+1})} &= T_{l+1}^{(t_{l+1}-1)} + \max_{x \in X(T_{l+1}^{(t_{l+1}-1)})} N\left(l+1, x\right).
\end{aligned}$$

Let $t \in T_{l+1}$, we call the smallest integer strictly greater than t in $T_{l+1} \cup T_{l+2}$, the *successor* of t. Formally, there exists $c_{l+1} \leq t_{l+1}$ such that $t = T_{l+1}^{(c_{l+1})}$. If c_{l+1} is strictly smaller than t_{l+1}, the successor of t is $T_{l+1}^{(c_{l+1}+1)}$; if $c_{l+1} = t_{l+1}$, then the successor of t is $T_{l+2}^{(1)}$.

We make few comments on the definition of the set T_{l+1}. First, the number of stages between two different integers t and t' in T_{l+1} is such that a strategy, which starts playing like $\sigma_{l+1}(x_t)$ at stage t yields an expected average payoff between stage t and stage $t'-1$ greater than $v^*(x_1) - \varepsilon_{l+1}$. Second, the weight of the first $T_l^{t_l}$ stages in a game of length T_{l+1}^1 is small.

We prove that the strategy $\sigma^*(x_1)$ is 0-optimal. We consider here $\sigma^*(x_1)$ as a mixed strategy, i.e. a probability over pure strategies. More precisely, let Ω be the set of pure strategies defined as concatenation of a sequence of integers $(n_l)_{l\geq 2}$ with $n_l \in T_l$ for every $l \geq 2$ and the sequence of strategy $(\sigma_l)_{l\geq 1}$. σ^* is a probability over Ω.

We show that every pure strategy in Ω is partially 0-optimal.

Lemma 4.4 *Let $(n_l)_{l\geq 2}$ be a sequence of integers such that for every $l \geq 2$, $n_l \in T_l$. Denote by σ the concatenated strategy induced by $(n_l)_{l\geq 2}$ and $(\sigma_l)_{l\geq 1}$.*

The strategy σ is partially 0-optimal. Moreover, we have explicit lower bounds for specific stages. For every $l \geq 2$, let n'_l be the successor of n_l. Then

$$\forall l \geq 2,\ \forall n \in [n'_l - 1, n_{l+1} - 1],\ \gamma_n(x_1, \sigma) \geq v^*(x_1) - 2\varepsilon_{l-1}.$$

Proof: We first show that the sequence $\gamma_{n_{l+1}-1}(x_1, \sigma)$ converges to the uniform value $v^*(x_1)$ when l goes to ∞. At stage n_l, the strategy σ starts to follow an ε_l-optimal strategy from the current state. By definition, $n'_l - n_l \geq N(l, x_{n_l})$, and thus by Equation (1)

$$\gamma_{n_l, n'_l - 1}(x_1, \sigma) = \gamma_{n'_l - n_l}(x_{n_l}, \sigma_l(x_{n_l})) \geq v^*(x_{n_l}) - \varepsilon_l \geq v^*(x_1) - \varepsilon_l.$$

More generally, for every $n \in [n'_l - 1, n_{l+1} - 1]$, we have

$$\gamma_{n_l, n}(x_1, \sigma) = \gamma_{n - n_l + 1}(x_{n_l}, \sigma_l(x_{n_l})) \geq v^*(x_1) - \varepsilon_l.$$

In particular we have

$$\gamma_{n_l, n_{l+1}-1}(x_1, \sigma) \geq v^*(x_1) - \varepsilon_l. \qquad (2)$$

The expected average payoff between stage n_l and $n_{l+1} - 1$ is greater than $v^*(x_1) - \varepsilon_l$. It follows that the sequence $(\gamma_{n_{l+1}-1}(x_1, \sigma))_{l \geq 1}$ converges to $v^*(x_1)$ and therefore the strategy σ is partially 0-optimal.

We now prove the second part of the lemma, giving some explicit subsequences and bounds on the rate of convergence: for all $l \geq 2$, for all n between $n'_l - 1$ and $n_{l+1} - 1$, we prove that

$$\gamma_n(x_1, \sigma) \geq v^*(x_1) - 2\varepsilon_{l-1}. \qquad (3)$$

Fix $l \geq 2$. We first prove this lower bound for the expected average payoff until stage $n_l - 1$ (which is before $n'_l - 1$). By definition of T^1_l, the weight of the n_{l-1} first stages is small in the MDP of length $n_l - 1$:

$$\begin{aligned}\frac{n_{l-1} - 1}{n_l - 1} \leq \frac{n_{l-1} - 1}{T^{(1)}_l - 1} &\leq \frac{T^{(t_{l-1})}_{l-1}}{T^{(t_{l-1})}_{l-1} + N(l-1, x_{T^{(t_{l-1})}_{l-1}}) - 1 + \left[\frac{1}{\varepsilon_{l-1}} + 1\right] T^{(t_{l-1})}_{l-1}} \\ &\leq \frac{T^{(t_{l-1})}_{l-1}}{\left[\frac{1}{\varepsilon_{l-1}} + 1\right] T^{(t_{l-1})}_{l-1}} \leq \varepsilon_{l-1}.\end{aligned}$$

Using Equation (2) for $l' = l - 1$ and the previous equation, it follows that

$$\begin{aligned}\gamma_{n_l - 1}(x_1, \sigma) &= \left[\frac{n_{l-1} - 1}{n_l - 1} \gamma_{n_{l-1}-1}(x_1, \sigma) + \frac{n_l - n_{l-1}}{n_l - 1} \gamma_{n_{l-1}, n_l - 1}(x_1, \sigma)\right], \\ &\geq \left[\gamma_{n_{l-1}, n_l - 1}(x_1, \sigma) - \frac{n_{l-1} - 1}{n_l - 1}\right], \\ &\geq v^*(x_1) - 2\varepsilon_{l-1}.\end{aligned}$$

Let n be a positive integer between $n'_l - 1$ and $n_{l+1} - 1$. The expected average payoff until stage n is the convex combination of the expected average payoff until stage $n_l - 1$ (before $n'_l - 1$) and the average expected payoff between stages n_l and n. Both of these quantities are greater than $v^*(x_1) - 2\varepsilon_{l-1}$, and therefore their convex combination is greater than $v^*(x_1) - 2\varepsilon_{l-1}$ as well. □

Remark 4.5 Following the notation of Lemma 4.4, if $n \in \cup_{l \geq 2}[n_l, n'_l - 2]$ then we only know that the n-stage expected average payoff is greater than 0.

Lemma 4.6 *$\sigma^*(x_1)$ is a 0-optimal strategy.*

Proof: We consider here $\sigma^*(x_1)$ as a mixed strategy. Lemma 4.4 showed that with probability one the pure strategies in the support of σ^* are partially 0-optimal.

Let $l \geq 2$ and fix n an integer in $[T^1_l, T^1_{l+1} - 1]$. We show that, with probability higher than $1 - \varepsilon_l$, the decision maker is following a pure strategy giving an expected average payoff until stage n higher than $v^*(x_1) - 2\varepsilon_{l-2}$.

By definition, there exists a unique stage n^*_l in T_l such that n is between n^*_l and $n'^*_l - 1$ where n'^*_l is the successor of n^*_l:

$$n^*_l \leq n \leq n'^*_l - 1. \tag{4}$$

Let σ be a pure strategy with positive probability under σ^*. There exists a sequence $(n_d)_{d \geq 2}$ such that for all $d \geq 2$, $n_d \in T_d$ and σ is the concatenated strategy induced by $(n_d)_{d \geq 2}$ and $(\sigma_d)_{d \geq 1}$. We follow the previous notation and denote for every $d \geq 2$, the successor of n_d by n'_d. Since $n \in [T^1_l, T^1_{l+1} - 1]$, by construction of the sets T_{l-1},T_l, and T_{l+1}, we have

$$n'_{l-1} \leq n \leq n_{l+1} - 1. \tag{5}$$

We now use the inequalities (4) and (5) to handle the three different cases depending on the respective places of n^*_l, the beginning of the block containing n, and n_l, the stage where the strategy σ is switching from an ε_{l-1} strategy to an ε_l-optimal strategy: $n_l > n^*_l$, $n_l < n^*_l$, and $n_l = n^*_l$.

If $n_l > n^*_l$, then at stage n the pure strategy σ is still following the ε_{l-1}-optimal strategy from state $x_{n_{l-1}}$ and therefore yields a high expected average payoff. Formally, we have $n'_{l-1} \leq n \leq n'^*_l - 1 \leq n_l - 1$, so that by Lemma 4.4 applied to $l' = l - 1$,

$$\gamma_n(x_1, \sigma) \geq v^*(x_1) - 2\varepsilon_{l-2}.$$

If $n_l < n^*_l$, then at stage n, the pure strategy σ has already followed the ε_l-optimal strategy from state x_{n_l} for a long time and thus yields a high expected average payoff. Formally, we have $n'_l \leq n^*_l \leq n \leq n_{l+1} - 1$, so that by Lemma 4.4 applied to l,

$$\gamma_n(x_1, \sigma) \geq v^*(x_1) - 2\varepsilon_{l-1} \geq v^*(x_1) - 2\varepsilon_{l-2}.$$

Finally if $n_l = n_l^*$, we do not control the expected average payoff but by definition of the stopping time u_l the probability of the event $\{n_l = n_l^*\}$ is smaller than ε_l under σ^*.

We can now conclude. We denote by $\mathbb{P}_{\sigma^*}$ the probability distribution induced by σ^* on the set of pure strategy and $\mathbb{E}_{\sigma^*}$ the corresponding expectation. Since the payoffs are in $[0,1]$, it follows that

$$\gamma_n(x_1, \sigma^*) = \mathbb{E}_{\sigma^*}\left(\gamma_n(x_1, \sigma)\right) \geq (1 - \varepsilon_l)(v^*(x_1) - 2\varepsilon_{l-2}) \geq v^*(x_1) - 3\varepsilon_{l-2}.$$

This is true for every $l \geq 1$ and every integer $n \in [T_l^1, T_{l+1}^1 - 1]$, therefore the expected average payoff converges to the uniform value: the strategy σ^* is 0-optimal. □

4.4 Proof of Theorem 3.1(2)

In this section, we prove Theorem 3.1(2): namely, if the set of states X is a precompact metric space, the transition is 1-Lipschitz, deterministic, and commutative, and the payoff function is uniformly continuous, then there exists a pure 0-optimal strategy.

We will first justify the existence of the uniform value and that we can assume that the set of states is compact. Then, we will define recursively a sequence of states $(x^l)_{l\geq 1}$ such that $x^1 = x_1$ and x^{l+1} is a limit point of states along an ε_l-optimal pure strategy $\sigma^l(x^l)$ starting from x^l where the value is constant on the induced play. Therefore, the value in all these states is equal to $v^*(x_1)$.

For each $l \geq 1$ we will define by induction a sequence of stages $(n_k^l)_{k\geq 1}$ such that the sequence of states induced by σ^l at stages n_k^l converges to the limit point x^{l+1}. We impose in addition conditions on n_{l+1}^l and on the speed of convergence. This sequence of stages splits the strategy σ^l into a finite sequence of streaks of actions. Given $k \geq 1$, we call an elementary block the streak of actions played between stage n_{k-1}^l and n_k^l. Note that it has $n_k^l - n_{k-1}^l$ actions. By convention, the first block starts at stage $n_0^l = 1$.

We will define the 0-optimal strategy σ^* by playing these elementary blocks in a specific order. The strategy σ^* is defined as a succession of two types of blocks $(A_l)_{l\geq 1}$ and $(B_l)_{l\geq 1}$ such that for all $l \geq 1$, A_l is composed of $l+1$ consecutive elementary blocks from $\sigma^l(x^l)$ and B_l is composed of $l-1$ elementary blocks, one from each $\sigma^{l'}(x^{l'})$ for $1 \leq l' \leq l-1$:

$$\sigma^* = (A_1, B_1, A_2, B_2, A_3, ...).^3$$

Block B_{l-1} ensures that the distance between the state at the beginning of block A_l and x^l is small. Block A_l guarantees an expected average payoff close to the value up to a function of ε_l. Moreover, block A_l is long enough for the total expected average payoff of σ^* at the end of A_l to be close to the value. It will follow that the strategy σ^* is partially 0-optimal. The rest of the proof consists in showing that the expected average payoff does not drop between these stages, neither during block B_{l+1} nor during the first stages of A_{l+1}. It follows that the strategy is 0-optimal.

[3] Recall that a pure strategy is identified with the sequence of actions it selects on the play path.

Let $\Gamma = (X, I, J, q, g)$ be a deterministic commutative MDP with a precompact metric space, a uniformly continuous payoff function and a 1-Lipschitz transition. We first justify the existence of the uniform value. We follow Section 6.1 in Renault [Ren11]. Let $\Psi = (Z, F, r)$ be an auxilliary dynamic programming problem. The set of states is $Z = X \times I$, the correspondence is given by

$$\forall (x,i) \in Z,\ F(x,i) = \{(q(x,a),a),\ a \in I\},$$

and the payoff function is for all $(x,i) \in Z$, $r(x,i) = g(x,i)$. We consider on Z the following metric $D((x,i),(x',i') = \max(d(x,x'), \delta_{i \neq i'})$. The set of states (Z, D) is precompact metric, the correspondence is 1-Lipschtiz (i.e. for all $z, z' \in Z, z_1 \in F(z)$ there exists $z'_1 \in F(z)$ such that $D(z_1, z'_1) \leq D(z,z')$) and the payoff function is equicontinuous. By Corollary 3.9 of the same paper [Ren11], Ψ has a uniform value for any initial state. We can deduce immediatly that $\Gamma(x_1)$ has a uniform value for every $x_1 \in X$.

We now prove that we can assume that X is compact. Define an MDP $\hat{\Gamma}(\hat{X}, I, \hat{q}, \hat{g})$ as follows: $\hat{X}$ is the Cauchy completion of X, $\hat{q}$ is the 1-Lipschitz extension of q to $\hat{X}$, $\hat{g}$ is the uniformly continuous extension[4] of g to $\hat{X}$. By Renault [Ren11] and previous paragraph, both MDPs Γ and $\hat{\Gamma}$ have a uniform value for any initial state.

Moreover the previous construction defines the transition on new states but does not change its value whenever it was already defined: for any state x_1 in X, $\hat{q}$ and q coincides, as well as g and $\hat{g}$. Therefore the MDPs $\Gamma(x_1)$ and $\hat{\Gamma}(x_1)$ are the same MDP on X. It follows that they have the same value.

In the following we assume that X is compact. Let $x_1 \in X$ and let $(\varepsilon_l)_{l \geq 1}$ be a decreasing sequence of positive real numbers that converges to 0. For each $x \in X$ and $l \geq 1$ denote by $\sigma_l(x)$ an ε_l-optimal pure strategy in $\Gamma(x)$ such that the value along the induced play is constant, and by $N(l,x)$ an integer such that

$$\forall n \geq N(l,x),\ \gamma_n(x, \sigma_l(x)) \geq v^*(x) - \varepsilon_l.$$

Since g is uniformly continuous, there exists $(\eta_l)_{l \geq 1}$ such that

$$\forall x, x' \in X,\ d(x,x') \leq \eta_l,\ \forall a \in \Delta(I),\ |g(x,a) - g(x',a)| \leq \varepsilon_l.$$

Let $\sigma = (i_t)_{t \geq 1} \in I^\infty$ be an infinite sequence of actions and let x_1 and x'_1 be two initial states. For every $n \geq 1$, the distance between x_n, the state at stage n obtained along the play induced by x_1 and σ, and x'_n, the state at stage n obtained along the play induced by x'_1 and σ, is smaller than $d(x_1, x'_1)$. It follows that

$$\forall x_1, x'_1 \in X,\ s.t.\ d(x_1,x'_1) \leq \eta_l,\ \forall \sigma = (i_t)_{t \geq 1} \in I^\infty,\ \forall n \geq 1,\ |\gamma_n(x_1,\sigma) - \gamma_n(x'_1,\sigma)| \leq \varepsilon_l.$$

Definition of the strategy σ^*: Let $x^1 = x_1$. Given $(x^j)_{1 \leq j \leq l}$ define x^{l+1} to be a limit point of the play $(x^l, \sigma_l(x^l))$. Since the value is constant on the play induced

[4] Note that an extension is not possible if the underlying function is only continuous.

by $(x^l, \sigma_l(x^l))$, the uniform value in x^{l+1} is also equal to $v^*(x_1)$. To construct the 0-optimal strategy, we split each play $\sigma_j(x^j)$ into blocks by induction on j.

Let us assume that $(n^j_k)_{k\geq 1}$ have been defined for every $j \leq l-1$, i.e. the splittings of all strategies $\{\sigma_1(x^1), ..., \sigma_{l-1}(x^{l-1}\}$ have been defined. We now split the sequence $\sigma_l(x^l)$.

Define $L_l = 1 + \sum_{j\leq l-1}(n^j_l - 1)$, which depends only on the sequences for $j \leq (l-1)$. We denote by $(x^l_n)_{n\geq 1}$ the sequence of states along $(x^l, \sigma_l(x^l))$. Let us define the sequence of stages $(n^l_k)_{k\geq 1}$ such that it satisfies four properties. The three first properties are restriction on n^l_{l+1} and the last one is a restriction on the rate of convergence to x^{l+1}. First the strategy $\sigma_l(x^l)$ guarantees in $\Gamma(x^l)$ the value with an error less than ε_l in all games longer than n^l_{l+1}:

$$n^l_{l+1} \quad \geq N(l, x^l). \tag{6}$$

Second, L_l is small compared to n^l_{l+1}:

$$\frac{L_l}{n^l_{l+1}} \leq \varepsilon_l. \tag{7}$$

Third,

$$\frac{N(l+1, x^{l+1}) + \sum_{j=1}^{l-1}\left(n^j_{l+1} - n^j_l\right)}{n^l_{l+1}} \leq \varepsilon_l. \tag{8}$$

Finally, at the beginning of the k-th block of this decomposition the state is close to the limit point

$$d(x^l_{n^l_k}, x^{l+1}) \leq \frac{\eta_k}{k-1}. \tag{9}$$

Fix $l \geq 1$. We define A_l to be the finite sequence of actions given by $\sigma^l(x^l)$ between stage 1 and stage n^l_{l+1}. In term of elementary blocks, it is composed of the first $l+1$ elementary blocks of $\sigma^l(x^l)$ and is composed of $n^l_{l+1} - 1$ actions. We define B_l as the sequence of actions where the decision maker is playing, for each $l' < l$, the elementary block of $\sigma^{l'}(x^{l'})$ between stages $n^{l'}_l$ and $n^{l'}_{l+1}$. Thus B_l is the concatenation of $l-1$ elementary blocks. Moreover the number of actions in B_l is $b_l = \sum_{j=1}^{l-1}\left(n^j_{l+1} - n^j_l\right)$, which appeared in (8). The strategy σ^* is the sequence of actions given by the alternating sequence $(A_l, B_l)_{l\geq 1}$.

We now show that the strategy σ^* is 0-optimal.

We first prove that the state at the beginning of A_l is close to x^l. Therefore the expected average payoff of σ^* at the end of A_l is bigger than $v^*(x_1) - 3\varepsilon_l$ and σ^* is partially 0-optimal.

Lemma 4.7 *The payoff at the end of A_l is greater than $v^*(x_1) - 3\varepsilon_l$:*

$$\gamma_{L_l + n^l_{l+1} - 1}(x_1, \sigma^*) \geq v^*(x_1) - 3\varepsilon_l$$

Corollary 4.8 *The strategy σ^* is partially* 0*-optimal.*

Proof of Lemma 4.7: Let us denote by $(x_n)_{n\geq 1}$ the sequence of states on the play induced by σ^*.

We first prove that the state at the beginning of A_l is close to x^l. One can verify that the first stage of A_l is the stage $L_l = 1 + \sum_{j\leq l-1}(n_l^j - 1)$. By definition, at stage L_l for each $l' \leq l-1$, all first l elementary blocks of $\sigma^{l'}(x^{l'})$ have been played: all of the first $l'+1$ on block $A_{l'}$ and then one after each other in the blocks B_j for $j \in [l'+1, l-1]$. By commutativity, the state does not depend on the order of actions and the state is the same as after the sequence σ' where the decision maker plays $\sigma_1(x^1)$ for $n_l^1 - 1$ stages, $\sigma_2(x^2)$ for $n_l^2 - 1$ stages,..., and $\sigma_{l-1}(x^{l-1})$ for $n_l^{l-1} - 1$ stages.

For each strategy σ_j, Equation (9) implies that the distance between x^{j+1} and the state at stage n_l^j on the play from x^j is less than $\frac{\eta_l}{l-1}$ for each $j \in \{1, ..., l-1\}$. The map q is 1-Lipschitz, so the distances sum up and an immediate induction implies that

$$d(x_{L_l}, x^l) \leq \eta_l. \tag{10}$$

Let us now compute the payoff in the MDP of length $L_l + n_{l+1}^l - 1$, i.e. until the end of A_l. Equation (7) ensures that the payoff is almost equal to the payoff between stages L_l and $L_l + n_{l+1}^l - 1$:

$$\begin{aligned}
\gamma_{L_l+n_{l+1}^l-1}(x_1,\sigma^*) &= \frac{L_l - 1}{L_l + n_{l+1}^l - 1}\gamma_{L_l-1}(x_1,\sigma^*) + \frac{n_{l+1}^l}{L_l + n_{l+1}^l - 1}\gamma_{L_l, L_l+n_{l+1}^l-1}(x_1,\sigma^*) \\
&\geq \frac{n_{l+1}^l}{L_l + n_{l+1}^l - 1}\gamma_{L_l, L_l+n_{l+1}^l-1}(x_1,\sigma^*) \\
&\geq \gamma_{L_l, L_l+n_{l+1}^l-1}(x_1,\sigma^*) - \frac{L_l - 1}{L_l + n_{l+1}^l - 1} \\
&\geq \gamma_{L_l, L_l+n_{l+1}^l-1}(x_1,\sigma^*) - \varepsilon_l.
\end{aligned}$$

Moreover σ^* plays like an ε_l-optimal strategy in $\Gamma(x^l)$ between stages L_l and $L_l + n_{l+1}^l - 1$, and the distance between x_{L_l} and x^l is less than η_l by Equation (10). Therefore, by Equation (6) we have

$$\begin{aligned}
\gamma_{L_l+n_{l+1}^l-1}(x_1,\sigma^*) &\geq \gamma_{n_{l+1}^l}(x_{L_l}, \sigma_l(x^l)) - \varepsilon_l \\
&\geq \gamma_{n_{l+1}^l}(x^l, \sigma_l(x^l)) - 2\varepsilon_l \\
&\geq v^*(x_1) - 3\varepsilon_l.
\end{aligned}$$

□

We now check that the average expected payoff does not drop between these stages. We distinguish between two different cases: if $n \in [L_l + n_{l+1}^l - 1, L_{l+1} + N(l+1, x^{l+1})]$ or if $n \in [L_{l+1} + N(l+1, x^{l+1}), L_{l+1} + n_{l+2}^{l+1} - 1]$.

In the first case, the MDP ends at a stage in B_l or in the beginning of block A_{l+1}. Equation (8) implies that the length of the game is almost equal to $L_l + n_{l+1}^l - 1$, therefore the expected average payoff is greater than $v^*(x_1) - 4\varepsilon_l$.

In the second case, the MDP ends in the middle of block A_{l+1}. The expected average payoff is the convex combination of the expected average payoff until $L_{l+1}-1$ and the average expected payoff between L_{l+1} and n. We check that both of them are high and we deduce that the expected average payoff is greater than $v^*(x_1) - 4\varepsilon_l$.

Lemma 4.9 *Let $n \in [L_l + n^l_{l+1} - 1, L_{l+1} + N(l+1, x^{l+1})]$. Then*

$$\gamma_n(x_1, \sigma^*) \geq v^*(x_1) - 4\varepsilon_l.$$

The expected average payoff in any n-stage MDP such that n is in the middle of block B_l or at the beginning of block A_{l+1} is greater than $v^*(x_1) - 4\varepsilon_l$.

Proof: The key point is that the number of stages is close to the case of Lemma 4.7. Let $n \in [L_l + n^l_{l+1} - 1, L_{l+1} + N(l+1, x^{l+1})]$. By equation (8), we have

$$\begin{aligned} n - L_l - n^l_{l+1} + 1 &\leq N(l+1, x^{l+1}) + \sum_{j=1}^{l-1} (n^j_{l+1} - n^j_l) \\ &\leq \varepsilon_l n^l_{l+1}. \end{aligned}$$

It follows that

$$\begin{aligned} \gamma_n(x_1, \sigma^*) &= \frac{L_l + n^l_{l+1} - 1}{n} \gamma_{L_l + n^l_{l+1} - 1}(x_1, \sigma^*) + \frac{n - L_l - n^l_{l+1} + 1}{n} \gamma_{L_l + n^l_{l+1}, n}(x_1, \sigma^*) \\ &\geq \frac{L_l + n^l_{l+1} - 1}{n} \gamma_{L_l + n^l_{l+1} - 1}(x_1, \sigma^*) \\ &\geq \gamma_{L_l + n^l_{l+1} - 1}(x_1, \sigma^*) - \frac{n - L_l - n^l_{l+1} + 1}{n} \\ &\geq v^*(x_1) - 3\varepsilon_l - \frac{n - L_l - n^l_{l+1} + 1}{n^l_{l+1}} \\ &\geq v^*(x_1) - 4\varepsilon_l. \square \end{aligned}$$

Lemma 4.10 *Let $n \in [L_{l+1} + N(l+1, x^{l+1}), L_{l+1} + n^{l+1}_{l+2} - 1]$. Then*

$$\gamma_n(x_1, \sigma^*) \geq v^*(x_1) - 4\varepsilon_l.$$

The payoff in any n-stage MDP stopping in the middle of block A_{l+1} is greater than $v^*(x_1) - 4\varepsilon_l$.

Proof: Let $n \in [L_{l+1} + N(l+1, x^{l+1}), L_{l+1} + n^{l+1}_{l+2} - 1]$. The expected average payoff is the convex combination of the expected average payoff until $L_l + n^l_{l+1} - 1$ and the expected average payoff between $L_l + n^l_{l+1} - 1$ and n. It follows that

$$\begin{aligned} \gamma_n(x_1, \sigma^*) &= \frac{L_{l+1} - 1}{n} \gamma_{L_{l+1} - 1}(x_1, \sigma^*) + \frac{n - (L_{l+1} - 1)}{n} \gamma_{L_{l+1}, n}(x_1, \sigma^*) \\ &= \frac{L_{l+1} - 1}{n} \gamma_{L_{l+1} - 1}(x_1, \sigma^*) + \frac{n - (L_{l+1} - 1)}{n} \gamma_{n - L_{l+1} + 1}(x_{L_{l+1}}, \sigma_{l+1}(x_{L_{l+1}})) \\ &\geq \frac{L_{l+1} - 1}{n} (v^*(x_1) - 4\varepsilon_l) + \frac{n - (L_{l+1} - 1)}{n} (v^*(x^{l+1}) - 2\varepsilon_{l+1}) \\ &\geq v^*(x_1) - 4\varepsilon_l. \end{aligned}$$

The expected average payoff is greater than $v^*(x_1) - 4\varepsilon_l$. □

Lemma 4.9 and Lemma 4.10 are true for every $l \geq 1$, therefore the strategy σ^* is pure and 0-optimal at x_1, which concludes the proof.

5 Commutative stochastic games.

In this section, we focus on commutative stochastic games and state-blind repeated games. In Section 5.1, we show that the class of absorbing games is in fact a subclass of commutative stochastic games. We show that each absorbing state can be replaced by a non-absorbing state leading to some new states, which are useless from a strategic point of view but designed in order to fulfill the commutativity assumption. In Section 5.2, we prove the existence of the uniform value in stochastic games with a deterministic commutative 1-Lipschitz transition (Theorem 3.6). In Section 5.3, we deduce the existence of the uniform value in state blind commutative repeated games (Corollary 3.7). In Section 5.4, we provide some generalizations.

5.1 Absorbing games

Absorbing games were introduced by Kohlberg [Koh74]. They are stochastic games with a single non-absorbing state. An absorbing game is thus given by $\Gamma = (\{\alpha\} \cup X, I, J, q, g)$ where α is the unique non-absorbing state and all states $x \in X$ are absorbing: $q(x,i,j)(x) = 1\ \forall x \in X, i \in I, j \in J$. The state α is the only state where the players have an influence on the payoff and on future states. For each action pair $(i,j) \in I \times J$, we denote by $q(\alpha,i,j)(X)$ the total probability to reach an absorbing state by playing the action pair (i,j).

Proposition 5.1 *Let $\Gamma = (\{\alpha\} \cup X, I, J, q, g)$ be an absorbing game. There exists a commutative game $\Gamma' = (X', I, J, q', g')$ and a state $\alpha'_2 \in X'$ such that for all $n \geq 1$, $v_n(\alpha) = v'_n(\alpha'_2)$. Moreover a player can guarantee w in $\Gamma'(\alpha'_2)$ if and only if he can guarantee w in $\Gamma(\alpha)$.*

Proof: Let $q(\alpha,i,j|X)$ be the conditional probability on X if the action pair (i,j) is played and there has been absorption. Define an auxiliary commmutative game $\Gamma' = (X', I', J', q', g')$ as follows. The action spaces are $I' = I$ and $J' = J$. For each $i \in I$ (resp. $j \in J$), we define a new state x_i (resp. x_j). The state space is given by $X' = X_I \times X_J$, where $X_I = \{\alpha'\} \cup \{x_i,\ i \in I\} \cup \{\omega\}$ and $X_J = \{\alpha'\} \cup \{x_j,\ j \in J\} \cup \{\omega\}$. In the following, we denote (α', α') by α'_2. The payoff function is defined by

$$\forall i, i' \in I, \forall j, j' \in J, \qquad \begin{aligned} g'(\alpha'_2, i, j) &= g(\alpha, i, j), \\ g'((x_{i'}, x_{j'}), i, j) &= \mathbb{E}_{q(\alpha, i', j'|X)}(g(x)), \\ g'((x_{i'}, \omega), i, j) &= 1, \\ g'((\omega, x_{j'}), i, j) &= 0, \\ g'((\omega, \omega), i, j) &= 1/2. \end{aligned}$$

The payoff function in Γ' reflects the role of the different states. The state α'_2 is a substitute of the state α, and for each pair (i', j'), the state $(x_{i'}, x_{j'})$ replaces the absorption occurring in state α by playing the action pair (i', j'). This state will not be absorbing but an equilibrium at $(x_{i'}, x_{j'})$ is to stay in this state. If player 1 deviates, then with some probability the state will remain $(x_{i'}, x_{j'})$ and with the remaining probability the new state will be $(\omega, x_{j'})$, where player 2 can guarantee a payoff of 0. Similarly, if player 2 deviates, then the new state will remain $(x_{i'}, x_{j'})$ with some probability and with the remaining probability it will be $(x_{i'}, \omega)$ where player 1 can guarantee a payoff of 1.

The transition q' is defined in three steps: we define two transitions s_I on X_I controlled only by player 1 and s_J on X_J controlled only by player 2. We then consider the product transition corresponding to the absorbing part of q, and finally we define q'. At each step, we check that the transition is commutative. We define s_I and s_J by

$$\begin{aligned} \forall i, i' \in I, \quad s_I(\alpha', i) &= x_i, & \forall j, j' \in J, \quad s_J(\alpha', j) &= x_j, \\ s_I(x_{i'}, i) &= \begin{cases} x_{i'} & \text{if } i = i', \\ \omega & \text{if } i \neq i', \end{cases} & s_J(x_{j'}, j) &= \begin{cases} x_{j'} & \text{if } j = j', \\ \omega & \text{if } j \neq j', \end{cases} \\ s_I(\omega, i) &= \omega, & s_J(\omega, j) &= \omega. \end{aligned}$$

We now verify that s_I is commutative. A similar argument shows that s_J is commutative. Let i and i' be two actions of player 1. It is sufficient to check that s_I commutes when $i \neq i'$. However, if player 1 plays i and i', the state after two stages is ω regardless of the initial state and of the order in which he plays these actions.

Let s be the transition on $X_I \times X_J$ defined by $s((x, y), (i, j)) = (s_I(x, i), s_J(y, j))$. The reader can verify that s is commutative; it is depicted graphically in Figure 2.

Let q' be defined as follows: $q'(x, i, j) = q(\alpha, i, j)(\alpha)\delta_x + q(\alpha, i, j)(X)\delta_{s(x,i,j)}$ for all $x \in X'$, for all $i \in I$, and for all $j \in J$. Thus for all $x \in X$, for all $i, i' \in I$ and for all $j, j' \in J$ we have

$$\begin{aligned} \widetilde{q}'(q'(x, i, j), i', j') = {} & q(\alpha, i, j)(\alpha) q(\alpha, i', j')(\alpha)\delta_x + q(\alpha, i, j)(\alpha) q(\alpha, i', j')(X)\delta_{s(x,i',j')} \\ & + q(\alpha, i, j)(X) q(\alpha, i', j')(\alpha)\delta_{s(x,i,j)} + q(\alpha, i, j)(X) q(\alpha, i', j')(X)\delta_{s(s(x,i,j),i',j')}. \end{aligned} \tag{11}$$

The right hand side of Equation (11) is symmetric between (i, j) and (i', j') except the last term that involves s. Since s is commutative, so is q'. Note that $\widetilde{q}$ may not be the product of one function depending on I and one function depending on J.

Fix $n \geq 1$. We prove that the n-stage values of $\Gamma(\alpha)$ and the n-stage values of $\Gamma'(\alpha'_2)$ are equal. Since the state (ω, ω) is absorbing, the value is equal to $1/2$, the stage payoff. For all i' in I, the state $(x_{i'}, \omega)$ is controlled by player 1. His optimal action is i' which guarantees him a payoff of 1. The situation is symmetric for $(\omega, x_{j'})$, so for all $j' \in J$, $v'_n((\omega, x_{j'})) = 0$. Fix $(i', j') \in I \times J$. The action i'(resp. j') is optimal for player 1 (resp. 2) in state $(x_{i'}, x_{j'})$ thus $v'_n(x_{i',j'}) = \mathbb{E}_{q(\alpha,i',j'|X)}(g(x))$. The stage payoffs and the continuation values are equal in both the game $\Gamma(\alpha)$ and the game

$$
\begin{array}{cc}
 & \begin{array}{ccc} & j & \\ & \downarrow & \end{array} \\
\xrightarrow{i} & \left(\begin{array}{c|c|c} (\omega,\omega) & (\omega,x_j) & (\omega,\omega) \\ \hline (x_i,\omega) & \circlearrowright & (x_i,\omega) \\ \hline (\omega,\omega) & (\omega,x_j) & (\omega,\omega) \end{array}\right) \\
 & (x_i,x_j)
\end{array}
\quad
\begin{array}{c}
\begin{array}{ccc} & j & \\ & \downarrow & \end{array} \\
\left(\begin{array}{ccc} (\omega,\omega) & (\omega,\omega) & (\omega,\omega) \\ \hline \circlearrowright & \circlearrowright & \circlearrowright \\ \hline (\omega,\omega) & (\omega,\omega) & (\omega,\omega) \end{array}\right) \\
(x_i,\omega)
\end{array}
$$

$$
\begin{array}{cc}
\xrightarrow{i} & \left(\begin{array}{c|c|c} (\omega,\omega) & \circlearrowright & (\omega,\omega) \\ (\omega,\omega) & \circlearrowright & (\omega,\omega) \\ (\omega,\omega) & \circlearrowright & (\omega,\omega) \end{array}\right) \\
 & (\omega,x_j)
\end{array}
\quad
\begin{array}{c}
\left(\begin{array}{ccc} \circlearrowright & \circlearrowright & \circlearrowright \\ \circlearrowright & \circlearrowright & \circlearrowright \\ \circlearrowright & \circlearrowright & \circlearrowright \end{array}\right) \\
(\omega,\omega)
\end{array}
$$

Figure 2: A graphic depiction of s.

$\Gamma'(\alpha_2)$ so the values in α and in α'_2 are equal.

Finally there is a correspondence between strategies. Given a strategy σ for player 1 in the absorbing game Γ that guarantees w, define σ' in Γ' by $\sigma'(\alpha'_2) = \sigma(\alpha)$ and for all $i' \in I$, $\sigma'(x_{i'}, .) = i'$. For all $i' \in I$ and $j' \in J$, this strategy guarantees the payoff $\mathbb{E}_{q(\alpha,i',j'|X)}(g(x))$ in the state $(x_{i'}, x_{j'})$, so it guarantees w from state α'_2. Reciprocally given σ' a strategy in Γ' that guarantees w' from α'_2, then σ'^* the strategy in Γ' such that $\sigma'^*(\alpha'_2) = \sigma'(\alpha'_2)$ and $\sigma'^*(x_i, .) = i$ also guarantees w' in $\Gamma'(\alpha'_2)$. The strategy σ defined by $\sigma(\alpha) = \sigma'(\alpha'_2)$ guarantees the same payoff in the absorbing game. From a strategic point of view the two games are completely equivalent. □

5.2 Proof of Theorem 3.6

In this section we prove Theorem 3.6. Let $\Gamma = (X, I, J, q, g)$ be a stochastic game where X is a compact subset of $\mathbb{R}^m$, I and J are finite sets, q is commutative, deterministic, and 1-Lipschitz for $\|.\|_1$, and g is continuous. We will prove that for all $z_1 \in \Delta_f(X)$, the stochastic game $\Gamma(z_1)$ has a uniform value. It is sufficient to prove that for all $x_1 \in X$, $\Gamma(x_1)$ has a uniform value.

The outline of the proof is the following. For each $x \in X$ we separate the action pairs into two different sets. An action pair $(i, j) \in I \times J$ is *cyclic* at x if the play that is obtained by repeating (i, j) starting from x, comes back to x after a finite number of stages. If (i, j) does not satisfy this property, we say that it is *non-cyclic*. Denote by $\mathcal{C}(x)$ the set of cyclic action pairs at x and by $\mathcal{NC}(x) = (I \times J)\backslash\mathcal{C}(x)$ the set of non-cyclic action pairs at x.

We denote by $\Phi_k = \{x; |\mathcal{C}(x)| \geq k\}$ the set of states with more than k cyclic action pairs. We will prove by induction on the number of cyclic action pairs that the uniform value exists for all initial points $x_1 \in X$.

We first argue that $\Phi_{|I\times J|}$ is non-empty and every state $x_1 \in \Phi_{|I\times J|}$ has a uniform value. To this end we will note that whatever the players play, only finitely many states can be reached from x_1, so that $\Gamma(x_1)$ is in essence a game with a finite number of states. By Mertens and Neyman [MN81] the game has a uniform value.

For the induction step, given a state x_1 with $k-1$ cyclic action pairs, we study a family $(\dot{\Gamma}(\varepsilon, x_1))_{\varepsilon>0}$ of games, which approximate $\Gamma(x_1)$ more and more precisely, and that have a uniform value. Assume by induction that for all states x in Φ_l, for $l \geq k$, the game $\Gamma(x)$ has a uniform value. For each $\varepsilon > 0$, let η be defined by uniform continuity of g. The game $\dot{\Gamma}(\varepsilon, x_1)$ is defined as follows: every state x such that there exists $l \geq k$ and $x' \in \Phi_l$ with $\|x - x'\|_1 \leq \eta$ is turned into an absorbing state with payoff the uniform value at x'. We will show that $\dot{\Gamma}(\varepsilon, x_1)$ can be written with a finite number of states, depending on x_1. By Mertens and Neyman [MN81], it has a uniform value at the initial state x_1 denoted $v(\varepsilon)(x_1)$. Finally, we prove that $v(\varepsilon)(x_1)$ converges when ε goes to 0 and that the limit is the uniform value of $\Gamma(x_1)$.

We now turn to the formal proof. We first prove an auxiliary Lemma studying the play induced by iterating the same action pair in Section 5.2.1. In Section 5.2.2, we focus on the initial step of the induction. In Section 5.2.3, we prove the inductive step and conclude the proof.

Denote by $q_{i,j}$ the operator from X to X defined by $q_{i,j}(x) = q(x, i, j)$. The map q is deterministic, so we can define the play along a sequence of actions. Fix $n \geq 1$ and $h = (i_1, j_1, ..., i_n, j_n) \in (I \times J)^n$. For all integers $l \leq n$ set $x_{l+1}(h) = q_{i_l, j_l} ... q_{i_1, j_1} x_1 = \prod_{t=1}^{l} q_{i_t, j_t} x_1$. We say that x is *reachable* from x_1 if there exists a play from x_1 to x.

5.2.1 Asymptotic behavior of the play induced by one action pair

Let $x \in X$. If the action pair (i, j) is cyclic at x then the sequence of states induced by repeating (i, j) from x is periodic. We focus on a non-cyclic action pair at x and we will prove that the set of states along the play induced by repeating (i, j) converges to a periodic orbit of states with strictly more cyclic action pairs than x. In order to prove this result, we use the following lemma (Sine [Sin90]).

Lemma 5.2 *Let $m \geq 1$, there exists $f(m) \geq 1$ such that for all maps M from $X \subset \mathbb{R}^m$ to X, 1-Lipschtiz for $\|.\|_1$, there exists an integer $L \leq f(m)$ and a family of maps $B_0, \cdots, B_{L-1}$ such that*

$$\forall l \in \{0, ..., L-1\}, \lim_{t \to +\infty} M^{tL+l} = B_l.$$

A classic example is the case where M is the transition of a Markov chain on a finite set. If λ is a complex eigenvalue of M then $|\lambda| \leq 1$ since the map is 1-Lipschitz. Moreover the theorem of Perron-Frobenius ensures that if $|\lambda| = 1$ then there exists $l \leq m$ such that $\lambda^l = 1$. The integer L is then the smallest common multiple of all such l and we can take $f(m) = m!$.

Applied to our framework, we deduce that, by iterating a non cyclic action pair (i, j) from x, the induced play has a finite number of limit points with strictly more cyclic action pairs than x.

Lemma 5.3 *Let $x \in X$, $(i,j) \in \mathcal{NC}(x)$ be a non-cyclic action pair at x, and $\varepsilon > 0$. There exist an integer n and a finite set $S_x \subset X$ such that*

$$\forall t \geq n, \exists x' \in S_x,\ \|q_{i,j}^t x - x'\|_1 \leq \varepsilon \text{ and } \sharp\mathcal{C}(x') \geq \sharp\mathcal{C}(x) + 1.$$

Proof: Let $x \in X$, $(i,j) \in \mathcal{NC}(x)$ be a non cyclic action pair and ε be a positive real. We show three properties: first the sequence $(q_{i,j}^t x)_{t \geq 1}$ has a finite number of limit points, then a cyclic action pair at x is still cyclic at the limit points and finally the pair (i,j) becomes cyclic at the limit points. Therefore, the number of cyclic action pairs strictly increases.

By Lemma 5.2 applied to $Q = q_{i,j}$, there exist an integer L and some operators B_0,..., B_{L-1} such that

$$\forall l \in \{0, ..., L-1\} \lim_{t \to +\infty} Q^{tL+l} = B_l.$$

In addition, for every $l \in \{0, ..., L-1\}$, $Q^l B_0 = B_0 Q^l = B_l$. By compactness of X, $B_0 x$ is in X and there exists an integer n such that

$$\forall t \geq n, \|Q^{tL} x - B_0 x\|_1 \leq \varepsilon.$$

Since Q is 1-Lipschitz for the norm 1, $\|Q^{tL+l} x - B_l x\|_1 \leq \varepsilon$. Denoting $n' = n(L+1)$ and $S_x = \{B_l x,\ l = 0, \ldots, L-1\}$, we have

$$\forall t \geq n',\ \exists x' \in S_x,\ \|Q^t x - x'\|_1 \leq \varepsilon.$$

The play has a finite number of limit points.

Let (i', j') be a cyclic action pair in x and d an integer such that $q_{i',j'}^d x = x$. We check that (i', j') is still cyclic at the limit points. For all $l \in \{0, ..., L-1\}$, we have

$$\begin{aligned} q_{i',j'}^d B_l x &= \lim_t q_{i',j'}^d Q^{tL+l} x \\ &= \lim_t Q^{tL+l} q_{i',j'}^d x = \lim_t Q^{tL+l} x = B_l x. \end{aligned}$$

The commutation assumption implies the second equality. Therefore (i', j') is still a cyclic action pair on the set S_x.

Moreover the iterated action pair (i,j), which was non-cyclic at x, becomes cyclic at x' for all $x' \in S_x$. For all $l \in \{0, ..., L-1\}$, we have

$$Q^L B_l x = \lim_t Q^L Q^{tL+l} x = \lim_t Q^{(t+1)L+l} x = B_l x.$$

All cyclic action pairs at x are still cyclic on S_x and (i,j) becomes cyclic, so the number of cycling action pairs is strictly increasing. □

Example 5.4 Consider a stochastic game with state space $X = \Delta(\mathbb{Z}/2\mathbb{Z})$, initial state $x_1 = (1,0)$, trivial sets of actions $I = \{i_1\}$, $J = \{j_1\}$, and transition

$$Q = q_{i_1,j_1} = \begin{pmatrix} 1/4 & 3/4 \\ 3/4 & 1/4 \end{pmatrix}.$$

Then for all $t \in \mathbb{N}$, $Q^t x_1$ has no cyclic action pairs but it converges to $x_\infty = (1/2, 1/2)$ where the action pair (i_1, j_1) is cyclic.

5.2.2 Initialization of the induction

Proposition 5.5 *The set* $\Phi_{|I\times J|}$ *is non-empty.*

The proposition is an immediate corrolary of Lemma 5.3. Starting from any initial state $x^1 \in X$, we apply Lemma 5.3 to one non-cyclic action pair and we get a state $x^2 \in X$ with more cyclic action pairs. Then, we can repeat from this new state and iterate the lemma until all the action pairs are cyclic.

Proposition 5.6 $\forall x_1 \in \Phi_{|I\times J|}$*, the game* $\Gamma(x_1)$ *has a uniform value.*

Proof: Fix $x_1 \in \Phi_{|I\times J|}$. Let $M \geq 1$ be such that for all action pairs (i,j), the play that starts at x_1 and in which the players repeatedly play (i,j) returns to x after at most M stages. We argue by contradiction that all states reachable from x_1 can be reached in less than $(M-1)\sharp(I\times J)$ stages. By contradiction let x^* be a state, which is not reached in $(M-1)\sharp(I\times J)$ stages. We define

$$t^* = \inf_{t\geq 1} \left\{t,\ \exists h = (i_l, j_l)_{l=1\dots t} \in (I\times J)^t,\ x_t(h) = x^*\right\}$$

the minimum number of stages needed to reach x^*. By assumption, $t^* > (M-1)\sharp(I\times J)$ and

$$\sum_{(i,j)\in\mathcal{C}(x_1)} \sharp\{l, (i_l, j_l) = (i,j)\} = t^*$$

$$\Rightarrow \exists (i^*, j^*) \in \mathcal{C}(x_1)\ \sharp\{l, (i_l, j_l) = (i^*, j^*)\} \geq \frac{t^*}{\sharp(I\times J)}$$

$$\Rightarrow \exists (i^*, j^*) \in \mathcal{C}(x_1)\ \sharp\{l, (i_l, j_l) = (i^*, j^*)\} \geq M.$$

So one action pair is repeated more than M times. By definition, there exists $d^* \leq M$ such that $q_{i,j}^{d^*} x_1 = x_1$. Hence the state at stage $t^* - d^*$ along the sequence of actions deduced from h, by deleting d^* times the action pairs (i^*, j^*), is x^*. This contradicts the definition of t^*. Therefore, all states are reached in less than $(M-1)\sharp(I\times J)$ stages and since I and J are finite, the game $\Gamma(x_1)$ can be defined only with a finite number of states.

Formally, the game $\Gamma(x_1)$ is a stochastic game with a finite set of states and finite sets of actions, so it has a uniform value by the theorem of Mertens and Neyman [MN81]. □

5.2.3 Inductive step

We now prove the inductive step. Fix $0 < k \leq |I \times J|$ and assume that for all $x \in \cup_{l=k}^{|I\times J|}\Phi_l$, the game $\Gamma(x)$ has a uniform value. Fix $x_1 \in \Phi_{k-1}$.

First, we check that the 1-Lipschitz transition and the uniform continuity of the payoff imply the continuity of the payoff that a player can guarantee, then we describe the family of auxiliary games and conclude the proof.

Lemma 5.7 *Given $\varepsilon > 0$, there exists $\eta > 0$ such that if $x \in X$ and player 1 guarantees w in $\Gamma(x)$ then, for all x', such that $\|x - x'\|_1 \leq \eta$, he guarantees $w - \varepsilon$ in $\Gamma(x')$.*

Proof: Given $\varepsilon > 0$, for all $(i, j) \in I \times J$, the map $g(\cdot, i, j)$ is uniformly continuous. Moreover, the number of maps is finite, so there exists $\eta > 0$ such that for all $x, x' \in X$ with $\|x - x'\|_1 \leq \eta$, we have

$$\forall (i,j) \in (I \times J),\ |g(x,i,j) - g(x',i,j)| \leq \varepsilon.$$

We first check the result for pure strategies. Fix $x \in X$. Let $\sigma \in \Sigma$ be a pure strategy, we define $\widetilde{\sigma}(x)$ to be the strategy which plays as if the game were $\Gamma(x)$ no matter what the initial state is. In particular, this strategy does not depend on the state and only on the sequence of actions. Let $\tau \in J^{\mathbb{N}}$ be a sequence of actions of player 2.

We denote by x_t the state at stage t along (x, σ, τ) and x'_t the state at stage t along $(x', \widetilde{\sigma}(x), \tau)$. For all $(i, j) \in I \times J$, q is a 1-Lipschtiz function so for all $t \geq 1$, $\|x_t - x'_t\|_1 \leq \|x - x'\|_1 \leq \eta$, and for all $n \geq 1$,

$$\begin{aligned}|\gamma_n(x,\sigma,\tau) - \gamma_n(x',\widetilde{\sigma}(x),\tau)| &\leq \frac{1}{n}\sum_{t=1}^{n} |g(x_t,i_t,j_t) - g(x'_t,i_t,j_t)| \\ &\leq \varepsilon.\end{aligned}$$

Let σ^* be a mixed strategy[5]. We denote by $\mathbb{P}_{\sigma^*}$ the probability distribution induced by σ^* on the set of pure strategies and $\mathbb{E}_{\sigma^*}$ the corresponding expectation. We define the mixed strategy $\widetilde{\sigma^*}$ by associating to each pure strategy σ the strategy $\widetilde{\sigma}(x)$. It is measurable and we have

$$\begin{aligned}|\gamma_n(x,\sigma^*,\tau) - \gamma_n(x',\widetilde{\sigma^*},\tau)| &\leq |\mathbb{E}_{\sigma^*}\left(\gamma_n(x,\sigma,\tau) - \gamma_n(x',\widetilde{\sigma}(x),\tau)\right)| \\ &\leq \mathbb{E}_{\sigma^*}\left(|\gamma_n(x,\sigma,\tau) - \gamma_n(x',\widetilde{\sigma}(x),\tau)|\right) \\ &\leq \varepsilon.\end{aligned}$$

If player 1 guarantees w in $\Gamma(x)$ then he guarantees $w - \varepsilon$ in the game $\Gamma(x')$ for every x' such that $\|x - x'\|_1 \leq \eta$. □

[5] Recall that by Kuhn's theorem, a behavioral strategy is equivalent to a mixed strategy.

Let ε be a positive real and η be associated to ε by Lemma 5.7. We denote by $\Phi(\eta)$ the set of states reachable from x_1 such that there is no state $x \in \cup_{l=k}^{|I\times J|}\Phi_l$ in the η-neighbourhood,

$$\Phi(\eta) = \{x \text{ reachable from } x_1,\ \forall x' \in X\ x' \notin \cup_{l=k}^{|I\times J|}\Phi_l \text{ or } \|x - x'\|_1 > \eta\}.$$

Proposition 5.8 *The set $\Phi(\eta)$ is finite.*

Proof: We first prove that there exists M such that any state in $\Phi(\eta)$ can be reached in less than M stages and then we deduce that $\Phi(\eta)$ is finite.

For each action pair (i,j) in $\mathcal{NC}(x_1)$, we denote by $u(i,j)$ the integer given by Lemma 5.3. Since there is a finite number of action pairs, there exists M' an integer such that for all $(i,j) \in \mathcal{NC}(x_1)$, $u(i,j) \leq M'$ and for all $(i,j) \in \mathcal{C}(x_1)$, the minimal period of (i,j) is smaller than M'. Set $M = M'\sharp(I \times J)$.

We prove that for all $x \in \Phi(\eta)$, $t^*(x) = \inf\{t|\ \exists h \in (I \times J)^t\ x_t(h) = x\}$, the least number of stages necessary to reach x, is smaller than M.

By contradiction, let $x \in \Phi(\eta)$ such that $t^* = t^*(x) \geq M$ and h be an history associated to x and t^*, then one action pair (i^*, j^*) is repeated more than M' times. This action pair is either cyclic or non-cyclic at x_1. If this action pair is cyclic, the history can be shortened, as in the proof of Proposition 5.6, which is absurd with respect to the definition of t^*. If this action pair is non-cyclic at x_1, there exists $\overline{x} \in X$ such that

$$\|q^{M'}_{i^*,j^*}x_1 - \overline{x}\|_1 \leq \varepsilon,$$
$$\text{and } \sharp\mathcal{C}(\overline{x}) > k-1.$$

Denote by h' the sequence of action pairs where (i^*, j^*) has been deleted M' times from h and x' the state obtained from $\overline{x}$ by playing h'. The transition is 1-Lipschitz and $\mathcal{C}$ is non-decreasing, therefore we have

$$\|x - x'\|_1 \leq \varepsilon,$$
$$\text{and } \sharp\mathcal{C}(x') > k-1,$$

which contradicts the definition of x.

To conclude notice that there exists a finite number of actions, therefore the set $\Phi(\eta)$ is finite. □

By Proposition 5.8, the set of states, reachable from x_1, and at a distance at least η from any state with more than k cyclic action pairs, i.e. $\Phi(\eta)$, is finite. We denote by $q(\Phi(\eta))$ the set of all states obtained by one transition from one of these states and, which are not already in $\Phi(\eta)$. The set $q(\Phi(\eta))$ is finite and for each $x \in q(\Phi(\eta))$, there exists $\xi(x) \in \cup_{l=k}^{|I\times J|}\Phi_l$ such that $d(x, \xi(x)) \leq \eta$. The induction assumption implies therefore that the game $\Gamma(\xi(x))$ has a uniform value denoted by $v^*(\xi(x))$. We define the auxiliary game $\dot{\Gamma}(\varepsilon, x_1)$ as follows: the initial state is x_1, the sets of actions are I and J, the transition function and reward functions are given by:

$$\dot{q}(x,i,j) = \begin{cases} q_{i,j}x & \text{if } x \in \Phi(\eta) \\ x & \text{if } x \in q(\Phi(\eta)) \\ x & \text{otherwise,} \end{cases}$$

$$\text{and } \dot{g}(x,i,j) = \begin{cases} g(x,i,j) & \text{if } x \in \Phi(\eta) \\ v^*(\xi(x)) & \text{if } x \in q(\Phi(\eta)) \\ 0 & \text{otherwise.} \end{cases}$$

The sets of strategies for players 1 and 2 are the same as in the game Γ. In the game starting at x_1, all the states are in $\Phi(\eta)$ or $q(\Phi(\eta))$. Since both sets are finite, this game is formally a stochastic game with a finite set of states and finite sets of actions. Therefore $\dot{\Gamma}(\varepsilon, x_1)$ has a uniform value by the theorem of Mertens and Neyman [MN81].

Proposition 5.9 *$\dot{\Gamma}(\varepsilon, x_1)$ has a uniform value in x_1 denoted by $v^*(\varepsilon)(x_1)$.*

We now prove that when ε goes to 0, the value $v^*(\varepsilon)(x_1)$ has to converge and the limit is the uniform value of the game $\Gamma(x_1)$. We first prove that the value of the auxiliary game is a good approximation to what the players can guarantee in $\Gamma(x_1)$.

Proposition 5.10 *If player 1 can guarantee w in $\dot{\Gamma}(\varepsilon, x_1)$ then he can guarantee $w - 3\varepsilon$ in $\Gamma(x_1)$.*

Proof: By assumption, there exists $\dot{\sigma}$ a strategy of player 1 in $\dot{\Gamma}(\varepsilon, x_1)$ and a stage $\dot{N} \geq 1$ such that

$$\forall n \geq \dot{N},\ \forall \dot{\tau},\ \dot{\gamma}_n(x_1, \dot{\sigma}, \dot{\tau}) \geq w - \varepsilon.$$

For each state $x \in q(\Phi(\eta))$, we denote by $\sigma^{\xi,x}$ the strategy given by Lemma 5.7 with respect to the point $\xi(x)$ and to an ε-optimal strategy in $\Gamma(\xi(x))$ such that

$$\exists N(x) \geq 1,\ \forall n \geq N(x),\ \forall \tau,\ \gamma_n(x, \sigma^{\xi,x}, \tau) \geq v^*(\xi(x)) - 2\varepsilon.$$

Let $\overline{N} = \max(N(x), x \in \Phi(\eta))$ be an upper bound.

Given an infinite play $h \in (X \times I \times J)^{\mathbb{N}}$, we denote by $\theta(h)$ the first stage where the state is at a distance less than η from a state in $\cup_{l=k}^{|I\times J|}\Phi_l$:

$$\theta(h) = \inf_{t \geq 1}\{t | x_t(h) \in q(\Phi(\eta))\}.$$

We define the strategy σ which plays optimally in $\dot{\Gamma}$ until a state $x' \in q(\Phi(\eta))$ is reached, and then optimally as if the remaining game was $\Gamma(\xi(x'))$. Formally, we have

$$\forall n \geq 1,\ \sigma_n(h) = \begin{cases} \dot{\sigma}_n(h) & \text{if } n \leq \theta(h) - 1 \\ \sigma^{\xi, x_{\theta(h)}(h)}_{n-\theta(h)+1} & \text{if } n \geq \theta(h). \end{cases}$$

We prove that σ guarantees $w-3\varepsilon$. Let τ be a strategy of player 2, we denote by x_t the state at stage t. Let $N^* \in \mathbb{N}$ such that $N^* \geq \dot{N}$ and $\frac{\overline{N}}{N^*} \leq \varepsilon$. Fix $n \geq N^*$, we separate the histories in two groups depending on whether $n-\theta(h)+1 > \overline{N}$ or $n-\theta(h)+1 \leq \overline{N}$.

We first focus on the set of histories $\{h \in H_\infty,\ n-\theta(h)+1 > \overline{N}\}$ and notice that on these histories the expected average payoff between $\theta(h)$ and n is close to the uniform value at $\xi(x_\theta(h))$.

We denote by σ^{h_n} and τ^{h_n} the strategies induced by σ and τ given the finite history h_n. Since $\|x_{\theta(h)} - \xi(x_{\theta(h)})\| \leq \eta$, we have

$$\begin{aligned}
&\mathbb{E}_{x_1,\sigma,\tau}\left(\sum_{t=\theta(h)}^{n} g(x_t,i_t,j_t)\mathbb{1}_{n-\theta(h)+1>\overline{N}}\right)\\
=&\mathbb{E}_{x_1,\sigma,\tau}\left(\gamma_{n-\theta(h)+1}(x_{\theta(h)},\sigma^{h_{\theta(h)}},\tau^{h_{\theta(h)}})(n-\theta(h)+1)\mathbb{1}_{n-\theta(h)+1>\overline{N}}\right)\\
\geq&\mathbb{E}_{x_1,\sigma,\tau}\left(\left(v^*(\xi(x_{\theta(h)}))-2\varepsilon\right)(n-\theta(h)+1)\mathbb{1}_{n-\theta(h)+1>\overline{N}}\right).
\end{aligned}$$

Therefore

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_{x_1,\sigma,\tau}\left(\sum_{t=1}^{n} g(x_t,i_t,j_t)\mathbb{1}_{n-\theta(h)+1>\overline{N}}\right)\\
&=\frac{1}{n}\mathbb{E}_{x_1,\sigma,\tau}\left(\left(\sum_{t=1}^{\theta(h)-1} g(x_t,i_t,j_t)+\sum_{t=\theta(h)}^{n} g(x_t,i_t,j_t)\right)\mathbb{1}_{n-\theta(h)+1>\overline{N}}\right)\\
&\geq \mathbb{E}_{x_1,\sigma,\tau}\left(\frac{1}{n}\left(\sum_{t=1}^{\theta(h)-1} g(x_t,i_t,j_t)+v^*(\xi(x_{\theta(h)}))(n-\theta(h)+1)\right)\mathbb{1}_{n-\theta(h)+1\geq\overline{N}}-2\varepsilon\mathbb{1}_{n-\theta(h)+1\geq\overline{N}}\right).
\end{aligned}$$

We now consider the set of histories $\{h \in H_\infty,\ n-\theta(h)+1 \leq \overline{N}\}$ and notice that the payoff between $\theta(h)$ and n has a small weight. By definition on this set of histories

$$\frac{n-\theta(h)+1}{n} \leq \frac{\overline{N}}{N^*} \leq \varepsilon.$$

Moreover we have

$$\forall x \in X,\ \forall x' \in q(\Phi(\eta)),\ \forall i \in I,\ \forall j \in J,\ g(x,i,j) \geq -1 \geq v(\xi(x')) - 2.$$

It follows that

$$
\begin{aligned}
&\frac{1}{n}\mathbb{E}_{x_1,\sigma,\tau}\left(\sum_{t=1}^{n} g(x_t,i_t,j_t)\mathbb{1}_{n-\theta(h)\leq \overline{N}}\right)\\
&= \mathbb{E}_{x_1,\sigma,\tau}\left(\frac{1}{n}\left(\sum_{n=1}^{\theta(h)-1} g(x_t,i_t,j_t)+\sum_{t=\theta(h)}^{n} g(x_t,i_t,j_t)\right)\mathbb{1}_{n-\theta(h)+1\leq \overline{N}}\right)\\
&\geq \mathbb{E}_{x_1,\sigma,\tau}\left(\frac{1}{n}\left(\sum_{t=1}^{\theta(h)-1} g(x_t,i_t,j_t)+v^*(\xi(x_{\theta(h)}))(n-\theta(h)+1)-2(n-\theta(h)+1)\right)\mathbb{1}_{n-\theta(h)+1\leq \overline{N}}\right)\\
&\geq \mathbb{E}_{x_1,\sigma,\tau}\left(\frac{1}{n}\left(\sum_{t=1}^{\theta(h)-1} g(x_t,i_t,j_t)+v^*(\xi(x_{\theta(h)}))(n-\theta(h)+1)\right)\mathbb{1}_{n-\theta(h)+1\leq \overline{N}}-2\varepsilon\mathbb{1}_{n-\theta(h)+1\leq \overline{N}}\right).
\end{aligned}
$$

Therefore by summing the two inequalities we get the result

$$\gamma_n(x_1,\sigma,\tau)\geq \dot{\gamma}_n(x_1,\dot{\sigma},\tau)-2\varepsilon\geq w-3\varepsilon.$$

□

It follows from Proposition 5.10 that for all $\varepsilon>0$, player 1 can guarantee $v(\varepsilon)(x_1)-3\varepsilon$ in the game $\Gamma(x_1)$. So player 1 can guarantee the superior limit when ε converges to 0: for all $\delta>0$, there exists n_1 and a strategy $\sigma^*\in\Sigma$ such that for all $\tau\in\mathcal{T}$, for all $n'\geq n_1$,

$$\gamma_{n'}(x_1,\sigma^*,\tau)\geq \limsup_{\varepsilon\to 0} v(\varepsilon)(x_1)-\delta.$$

The same argument shows that player 2 can guarantee the inferior limit. Therefore, for all $\delta>0$, there exist n_2 and a strategy $\tau^*\in\mathcal{T}$ such that for all $\sigma\in\Sigma$, for all $n'\geq n_2$,

$$\gamma_{n'}(x_1,\sigma,\tau^*)\leq \liminf_{\varepsilon\to 0} v(\varepsilon)(x_1)+\delta.$$

Given $\delta>0$ and $n'\geq\max(n_1,n_2)$, we have

$$\limsup_{\varepsilon\to 0} v(\varepsilon)(x_1)-\delta\leq \gamma_{n'}(x_1,\sigma^*,\tau^*)\leq \liminf_{\varepsilon\to 0} v(\varepsilon)(x_1)+\delta.$$

Therefore $v(\varepsilon)(x_1)$ converges when ε goes to 0 and the limit is the uniform value of the game $\Gamma(x_1)$. This proves the induction hypothesis at the next step and concludes the proof. For all $x_1\in X$, the game $\Gamma(x_1)$ has a uniform value.

5.3 Proof of Corollary 3.7

In this section, we provide a short proof of Corollary 3.7. Recall that given a state-blind repeated game $\Gamma^{sb}=(K,I,J,q,g)$ with a commutative transition q, we define the auxiliary stochastic game $\Psi=(X,I,J,\widetilde{q},\widetilde{g})$ where $X=\Delta(K)$, $\widetilde{q}$ is the linear extension of q, and $\widetilde{g}$ is the linear extension of g.

In this framework deducing the existence of the uniform value in the original repeated game from the existence of the uniform value in the auxiliary game is easy since the sets of strategies are almost the same in the two games. A player can use a strategy of the repeated game Γ in Ψ by looking only at the actions played and reciprocally a player can use a strategy of the stochastic game Ψ in the repeated game Γ by completing the sequence of actions with the unique sequence of compatible beliefs.
Proof: The set of strategies in the game Γ^{sb} are respectively denoted by Σ^{sb} and $\mathcal{T}^{sb}$. We will denote in this proof the payoff in the n stage game by γ_n^{sb} and the value of the n-stage game by $v_n^{sb}(p_1)$ for all $n \geq 1$.

We denote by $\widetilde{H}_t$ the set of histories in Ψ of length t, by $\widetilde{\Sigma}$ the set of strategies of player 1, and by $\widetilde{\mathcal{T}}$ the set of strategies of player 2. Let $p_1 \in \Delta(K)$, $\widetilde{\sigma} \in \widetilde{\Sigma}$ and $\widetilde{\tau} \in \widetilde{\mathcal{T}}$. The payoff in the n-stage game, starting from p_1 and given that the players follow $\widetilde{\sigma}$ and $\widetilde{\tau}$, is denoted by $\widetilde{\gamma_n}(\delta_{p_1}, \widetilde{\sigma}, \widetilde{\tau})$ and the value by $w_n(p_1)$. The set X is compact, $\widetilde{g}$ is continuous and the transition $\widetilde{q}$ is commutative and deterministic, so we can apply Theorem 3.6 to Ψ. We denote by $w^*(p_1)$ the uniform value. The values of both games coincide since the payoff and strategy sets coincide up to the following identification.

We focus on the case of player 1 since the situation is symmetric for player 2. Let σ^{sb} be a strategy in Σ^{sb}, then it defines naturally a strategy $\widetilde{\sigma}$ in $\widetilde{\Sigma}$ by forgetting the states. If we denote by Π^t the projection from $\widetilde{H}_t$ on H_t^{sb} that keeps only the actions: for all $t \geq 1$, we define

$$\widetilde{\sigma}(\widetilde{h}_t) = \sigma^{sb}(\Pi^t(\widetilde{h}_t)).$$

Reciprocally for all $t \geq 1$, given a sequence of actions $h_t^{sb} = (i_1, j_1, ...i_t, j_t)$, the completion $\Xi^t(h^{sb})$ in $\widetilde{H}_t$ is the unique sequence such that p_1 is fixed and for all $t \geq 1$, $q(p_t, i_t, j_t) = p_{t+1}$. Let $\widetilde{\sigma}$ be a strategy in $\widetilde{\Sigma}$, then we define the strategy σ^{sb} by completing the history: for all $t \geq 1$

$$\sigma^{sb}(h_t^{sb}) = \widetilde{\sigma}(\Xi^t(h_t^{sb})).$$

A similar procedure gives two functions between the sets of strategies of player 2.

Given $\widetilde{\sigma} \in \widetilde{\Sigma}$ and $\tau^{sb} \in \mathcal{T}^{sb}$, set $\sigma^{sb} \in \Sigma^{sb}$ and $\widetilde{\tau} \in \widetilde{\mathcal{T}}$ as in the previous paragraph. By definition of $\widetilde{q}$, the state at stage t in Ψ under $\mathbb{P}_{\delta_p, \widetilde{\sigma}, \widetilde{\tau}}$ is equal to the law of the state in Γ^{sb} under $\mathbb{P}_{p, \sigma^{sb}, \tau^{sb}}$. Therefore for all $n \geq 1$, we have

$$\gamma_n^{sb}(p_1, \sigma^{sb}, \tau^{sb}) = \widetilde{\gamma_n}(\delta_{p_1}, \widetilde{\sigma}, \widetilde{\tau}).$$

Finally, let $\varepsilon > 0$, $\widetilde{\sigma}$ be an ε-optimal strategy in Ψ and $N \geq 1$ an integer such that for all $\widetilde{\tau} \in \widetilde{\mathcal{T}}$,

$$\widetilde{\gamma_n}(\delta_{p_1}, \widetilde{\sigma}, \widetilde{\tau}) \geq w^*(p_1) - \varepsilon,$$

then for all $\tau^{sb} \in \mathcal{T}^{sb}$, we have

$$\begin{aligned} \gamma_n^{sb}(p_1, \sigma^{sb}, \tau^{sb}) &= \widetilde{\gamma_n}(\delta_{p_1}, \widetilde{\sigma}, \widetilde{\tau}) \\ &\geq w^*(p_1) - \varepsilon. \end{aligned}$$

The strategy σ^{sb} guarantees $w^*(p_1) - \varepsilon$ and therefore player 1 guarantees $w^*(p_1)$. By symmetry, player 2 guarantees $w^*(p_1)$ and the game $\Gamma^{sb}(p_1)$ has a uniform value equal to $w^*(p_1)$. □

5.4 Extensions.

The proof of Theorem 3.6 can be extended by replacing some of the lemmas with more general results. The result of Sine [Sin90], for example, applies to more general norms than the norm $\|.\|_1$.

Definition 5.11 *A norm on $\mathbb{R}^n$ is polyhedral if the unit ball has a finite number of extreme points.*

For example the norm $\|.\|_1$ and the sup norm are polyhedral norms but not the Euclidean norm. For polyhedral norm, the application of the theorem of Sine [Sin90] to compact sets gives the following results,

Lemma 5.12 *Let $N(.)$ be a polyhedral norm and $K \subset \mathbb{R}^m$ be a compact set. There exists $\varphi(N, m) \in \mathbb{N}$ such that for all functions T, 1-Lipschtiz for N, there exists $t \leq \varphi(N, m)$ such that $(T^{tn})_{n\in\mathbb{N}}$ converges.*

We deduce the following theorem.

Theorem 5.13 *Let $\Gamma = (X, I, J, q, g)$ be a stochastic game, such that X is a compact set of $\mathbb{R}^m$, I and J are finite sets, q is commutative deterministic 1-Lipschitz for a polyhedral norm, and g is continuous. For all $z_1 \in \Delta_f(X)$, the stochastic game $\Gamma(z_1)$ has a uniform value.*

This theorem does not apply to Example 2.3 on the circle and the existence of a uniform value in this model is still an open question.

We can obtain new results on non zero-sum stochastic games by replacing the theorem from Mertens and Neyman [MN81] with other existence results. First, Vieille [Vie00a][Vie00b] proves the existence of an equilibrium payoff in every two-player stochastic games. So our proof, adapted to the non zero-sum case leads to the following result.

Theorem 5.14 *Let $\Gamma = (X, I, J, q, g_1, g_2)$ be a two-player non zero-sum stochastic game such that X is a compact subset of $\mathbb{R}^m$, I and J are finite sets of actions, q is commutative deterministic 1-Lipschitz for $\|.\|_1$ and g_1 and g_2 are continuous. Then, for all $z_1 \in \Delta_f(X)$, the stochastic game $\Gamma(z_1)$ has an equilibrium payoff.*

Secondly, there exist some specific classes of m-player stochastic games where the existence of an equilibrium has been proven. For example, Flesch, Schoenmakers and Vrieze [FSV08][FSV09] prove the existence of an equilibrium for m-player stochastic games where each player controls a finite Markov chain and the payoffs depend on the m states and the m actions at stage n. Note that the commutativity assumption here is reduced to a condition player by player. As in our proof, the commutativity assumption implies that we can study deterministic transitions 1-Lipschitz for the norm $\|.\|_1$.

Theorem 5.15 *Let $\Gamma = \big((X_j, I_j, q_j)_{j\in\{1,...,m\}}, g\big)$ be a m-player product-state space stochastic game such that for all $j \in \{1, ..., m\}$, X_j is a compact subset of $\mathbb{R}^{m_j}$, I_j is a finite set of actions, q_j is commutative deterministic 1-Lipschitz for $\|.\|_1$ and $g : \prod(X_j \times I_j) \to [0,1]^m$ is continuous. For all $z_1 \in \Delta_f(\prod_j X_j)$, the stochastic game $\Gamma(z_1)$ has an equilibrium payoff.*

Acknowledgments

I thank J.Flesch, S.Gaubert, J.Renault, E. Solan and S.Sorin as well as three referees. The suggestions they provided were extremely helpful.

Part of this work has being done in *C&O* (University Paris 6) and GREMAQ (University Toulouse 1 Capitole).

The author gratefully acknowledges the support of the Agence Nationale de la Recherche, under grant ANR JEUDY, ANR-10-BLAN 0112, 41 as well as the PEPS project Interactions INS2I Propriétés des Jeux Stochastiques de Parité à Somme Nulle avec Signaux and the Israel Science Foundation, under grant ISF Grant #1517/11.

References

[AM95] R.J. Aumann, M. Maschler, and R.E. Stearns. *Repeated games with incomplete information.* The MIT press, 1995.

[Bla62] D. Blackwell. Discrete dynamic programming. *Ann. Math. Statist.*, 33:719–726, 1962.

[BF68] D. Blackwell, T. S. Ferguson The Big Match *Ann. Math. Statist.*, 39:159-163, 1968.

[Cou03] J.M. Coulomb. Stochastic games without perfect monitoring. *International Journal of Game Theory*, 32(1):73–96, 2003.

[DJ79] E.B. Dynkin and A.A. Juškevič. *Controlled markov processes*, volume 235. Springer New York, 1979.

[Fan53] K. Fan. Minimax theorems. *Proceedings of the National Academy of Sciences of the United States of America*, 39:42–47, 1953.

[For82] F. Forges. Infinitely repeated games of incomplete information: Symmetric case with random signals. *International Journal of Game Theory*, 11(3):203–213, 1982.

[FSV08] J. Flesch, G. Schoenmakers, and K. Vrieze. Stochastic games on a product state space. *Mathematics of Operations Research*, 33(2):403–420, 2008.

[FSV09] J. Flesch, G. Schoenmakers, and K. Vrieze. Stochastic games on a product state space: The periodic case. *International Journal of Game Theory*, 38(2):263–289, 2009.

[Gei02] J. Geitner. Note Equilibrium payoffs in stochastic games of incomplete information: the general symmetric case. *International Journal of Game Theory*, 30(3):449–452, 2002.

[Gil57] D. Gillette. Stochastic games with zero stop probabilities. *Ann. Math. Stud*, 39:178–187, 1957.

[Koh74] E. Kohlberg. Repeated games with absorbing states. *The Annals of Statistics*, 2(4):724–738, 1974.

[KZ74] E. Kohlberg and S. Zamir. Repeated games of incomplete information: The symmetric case. *The Annals of Statistics*, 2(5):1040–1041, 1974.

[Kuh53] , H.W. Kuhn. Extensive games and the problem of information, *Contributions to the Theory of Games*,2(28):193–216, 1953.

[Lev12] L. Yehuda. A discounted stochastic game with no stationary Nash equilibrium Hebrew University of Jerusalem-Center for the Study of Rationality, 2012.

[LS92] E. Lehrer and S. Sorin. A uniform Tauberian theorem in dynamic programming. *Mathematics of Operations Research*, 303–307, 1992.

[MZ71] J.-F. Mertens and S. Zamir. The Value of Two-Person Zero-Sum Repeated Games the extensive case *International Journal of Game Theory*, 1(1):217–227,1971.

[MZ80] J.-F. Mertens and S. Zamir. Minmax and maxmin of repeated games with incomplete information *International Journal of Game Theory*, 9(4):201–215, 1980.

[MN81] J.-F. Mertens and A. Neyman. Stochastic games. *Internat. J. Game Theory*, 10(2):53–66, 1981.

[NS98] A. Neyman and S. Sorin. Equilibria in repeated games of incomplete information: the general symmetric case. *Internat. J. Game Theory*, 27(2):201–210, 1998.

[Ren12] J. Renault. The value of Repeated Games with an informed controller. *arXiv:0803.3345v2, preprint*, 2009.

[Ren11] J. Renault. Uniform value in dynamic programming. *J. Eur. Math. Soc.*, 13:309-330, 2011.

[Rhe74] , D. Rhenius. Incomplete information in Markovian decision models *The Annals of Statistics*, 6:1327–1334, 1974.

[RSV02] D. Rosenberg, E. Solan, and N. Vieille. Blackwell optimality in Markov decision processes with partial observation. *Ann. Statist.*, 30(4):1178–1193, 2002.

[RSV03] D. Rosenberg, E. Solan, and N. Vieille. The Max Min value of stochastic games with imperfect monitoring. *Internat. J. Game Theory*, 32:133–150, 2003.

[RSV09] D. Rosenberg, E. Solan, and N. Vieille. Protocol with no acknowledgement. *Oper. Res.*, 57:905–915, 2009.

[Sha53] L.S. Shapley. Stochastic games. *Proc. of the National Academy of Sciences of the United States of America*, 39:1095, 1953.

[SY70] , Y. Sawaragi and T. Yoshikawa. Discrete-time Markovian decision processes with incomplete state observation. *The Annals of Mathematical Statistics*, 41:78–86, 1970

[Sin90] R. Sine A nonlinear Perron-Frobenius theorem. *Proc. Amer. Math. Soc.*, 109(2):331–336, 1990.

[Thu92] F. Thuijsman Optimality and equilibria in stochastic games. *CWI Tracts*, 82:1–107, 1992.

[Vie00a] N. Vieille. Two player stochastic games. I. A reduction. *Israel J. Math.*, 119:55–91, 2000.

[Vie00b] N. Vieille. Two-player stochastic games II: The case of recursive games. *Israel Journal of Mathematics*, 119(1):93–126, 2000.

[Vig13] G. Vigeral. A zero-sum stochastic game with compact action sets and no asymptotic value. *Dynamic Games and Applications*, to appear , 2013.

[Zil13] B. Ziliotto. Zero-sum repeated games: counterexamples to the existence of the asymptotic value and the conjecture $maxmin = limv(n)$. , preprint.