Pathwise uniform value in gambling houses and Partially Observable Markov Decision Processes

Xavier Venel, Bruno Ziliotto

▶ **To cite this version:**

Xavier Venel, Bruno Ziliotto. Pathwise uniform value in gambling houses and Partially Observable Markov Decision Processes. 2016. hal-01302567

HAL Id: hal-01302567
https://hal.science/hal-01302567

Preprint submitted on 14 Apr 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:1505.07495v2 [math.OC] 8 Sep 2015

Pathwise uniform value in gambling houses and Partially Observable Markov Decision Processes

Xavier Venel [*], Bruno Ziliotto[†]

September 9, 2015

Abstract

In several standard models of dynamic programming (gambling houses, MDPs, POMDPs), we prove the existence of a robust notion of value for the infinitely repeated problem, namely the *pathwise uniform value*. This solves two open problems. First, this shows that for any $\epsilon > 0$, the decision-maker has a pure strategy σ which is ϵ-optimal in any n-stage game, provided that n is big enough (this result was only known for behavior strategies, that is, strategies which use randomization). Second, the strategy σ can be chosen such that under the long-run average payoff criterion, the decision-maker has more than the limit of the n-stage values.

Keywords: Dynamic programming, Markov decision processes, Partial Observation, Uniform value, Long-run average payoff.

MSC2010: *Primary:* 90C39, *Secondary:* 90C40, 37A50, 60J20.

Introduction

The standard model of Markov Decision Process (or Controlled Markov chain) was introduced by Bellman [4] and has been extensively studied since then. In this model, at the beginning of every stage, a decision-maker perfectly observes the current state, and chooses an action accordingly, possibly randomly. The current state and the selected action determine a stage payoff and the law of the next state. There are two standard ways to aggregate the stream of payoffs. Given a strictly positive integer n, in the n-stage MDP, the total payoff is the Cesaro mean $n^{-1}\sum_{m=1}^{n} g_m$, where g_m is the payoff at stage m. Given $\lambda \in (0,1]$, in the λ-discounted MDP, the total payoff is the λ-discounted sum $\lambda\sum_{m\geqslant 1}(1-\lambda)^{m-1}g_m$. The maximum payoff that the decision-maker can obtain in the n-stage problem (resp. λ-discounted problem) is denoted by v_n (resp. v_λ).

A huge part of the literature investigates *long-term* MDPs, that is, MDPs which are repeated a large number of times. In the n-stage problem (resp. λ-discounted problem), this corresponds to n being large (resp. λ being small). A first approach is to determine whether (v_n) and (v_λ) converge when n goes to infinity and λ goes to 0, and whether the two limits coincide. When this is the case, the MDP is said to have an *asymptotic value*. The asymptotic value represents the long-term payoff outcome. When the asymptotic value exists, a second approach consists in determining if for any $\epsilon > 0$, there exists a behavior (resp. pure) strategy that is optimal up to ϵ in any n-stage and λ-discounted problem, provided that n is big and λ is small. When this is the case, the MDP is said to have a *uniform value* in behavior (resp. pure) strategies.

A third approach is to define the payoff in the infinite problem as being the expectation of $\liminf_{n\to+\infty} n^{-1}\sum_{m=1}^{n} g_m$: in literature, this is referred as the *long-run average payoff*

[*]CES, Université Paris 1 Panthéon Sorbonne, Paris. France. Email: xavier.venel@univ-paris1.fr
[†]TSE (GREMAQ, Université Toulouse 1 Capitole), 21 allée de Brienne, 31000 Toulouse, France.

criterion[1] (AP criterion, see Arapostathis et al. [3] for a review of the subject). We denote by w_∞ the maximal payoff that the decision-maker can guarantee under this criterion. Clearly, under this criterion, the decision-maker cannot have more than $\liminf_{n\to+\infty} v_n$. A natural question is whether he can obtain $\liminf_{n\to+\infty} v_n$.

When the set space and action sets are finite, Blackwell [6] has proved the existence of a pure strategy that is optimal for every discount factor close to 0, and one can deduce that the uniform value exists in pure strategies, and that under the AP criterion, the decision-maker can have $\lim_{n\to+\infty} v_n$.

In many situations, the decision-maker may not be perfectly informed of the current state variable. For instance, if the state variable represents a resource stock (like the amount of oil in an oil field), the quantity left, which represents the state, can be evaluated, but is not exactly known. This motivates the introduction of the more general model of Partially Observable Markov Decision Process (POMDP). In this model, at each stage, the decision-maker does not observe the current state, but instead receives a signal which is correlated to it. Rosenberg, Solan and Vieille [18] have proved that any POMDP has a uniform value in behavior strategies, when the state space, the action set and the signal set are finite. In the proof, the authors highlight the necessity that the decision-maker resort to behavior strategies, and ask whether the uniform value exists in pure strategies. They also raise the question of the behavior of the time averages of the payoffs, which is linked to the AP criterion. Renault [15] and Renault and Venel [16] have provided two alternative proofs of the existence of the uniform value in behavior strategies in POMDPs, and also ask whether the uniform value exists in pure strategies.

One of the main contributions of this paper is to solve this question positively. We prove that POMDPs have a uniform value in pure strategies. Moreover, for all $\epsilon > 0$, under the AP criterion, the decision-maker can have $\lim_{n\to+\infty} v_n - \epsilon$. In fact, we prove this result in a much more general framework, as we shall see now.

The result of Rosenberg, Solan and Vieille [18] (existence of the uniform value in behavior strategies in POMDPs) has been generalized in several dynamic programming models with infinite state space and action set. The first one is to consider the model of gambling house. Introduced by Dubins and Savage [10], a gambling house is defined by a correspondence from a metric space X to the set of probabilities on X. At every stage, the decision-maker chooses a probability on X which is compatible with the correspondence and the current state. A new state is drawn from this probability, and this new state determines the stage payoff. When the state space is compact, and the correspondence is 1-Lipschitz, and the payoff function is continuous (for suitable metrics), the existence of the uniform value in behavior strategies stems from the main theorem in [15]. One can deduce from this result the existence of the uniform value in behavior strategies in MDPs and POMDPs, for a finite state space and any action and signal sets. Renault and Venel [16] have extended the results of [15] to more general payoff evaluations.

The proofs in Renault [15] and Renault and Venel [16] are quite different from the one of Rosenberg, Solan and Vieille [18]. Still, they heavily rely on the use of behavior strategies for the decision-maker, and they do not provide any results concerning the AP criterion.

In this paper, we consider a gambling house with compact state space, closed graph correspondence and continuous payoff function. We show that if the family $\{v_n, n \geqslant 1\}$ is equicontinuous and w_∞ is continuous, the gambling house has a uniform value in pure strategies. Moreover, for all $\epsilon > 0$, the decision-maker can guarantee $\lim_{n\to+\infty} v_n - \epsilon$ under the AP criterion. This result especially applies to 1-Lipschitz gambling houses. We deduce the same result for compact MDPs with 1-Lipschitz transition, and POMDPs with finite set space, compact action set and finite signal set.

Note that under an ergodic assumption on the transition function, like assuming that from any state, the decision-maker can make the state go back to the initial state (see Altman [2]), or assuming that the law of the state variable converges to an invariant measure (see Borkar

[1] In some papers, the decision-maker minimizes the cost: in this case, the long-run average payoff criterion corresponds to the *long-run average cost criterion*.

[7, 8]), these results were already known. One remarkable feature of our proof is that we are able to use ergodic theory without any ergodic assumptions.

The paper is organized as follows. The first part presents the model of gambling house and recalls usual notions of value. The second part defines pathwise uniform value and states our results, that is, the existence of the pathwise uniform value in gambling houses, MDPs and POMDPs. The last three parts are dedicated to the proof of these results.

1 Gambling houses

1.1 Model of gambling house

Let us start with a few notations. We denote by $\mathbb{N}^*$ the set of strictly positive integers. If A is a measurable space, we denote by $\Delta(A)$ the set of probability measures over A. If (A, d) is a compact metric space, we will always equip (A, d) with the Borelian algebra, and denote by $\mathcal{B}(A)$ the set of Borel subsets of A. The set of continuous functions from A to $[0, 1]$ is denoted by $\mathcal{C}(A, [0, 1])$. The set $\Delta(A)$ is compact metric for the Kantorovich-Rubinstein distance d_{KR}, which metrizes the weak* topology. Recall that the distance d_{KR} is defined for all z and z' in $\Delta(A)$ by

$$d_{KR}(z, z') := \sup_{f \in E_1} \left| \int_A f(x)z(dx) - \int_A f(x)z'(dx) \right| = \inf_{\pi \in \Pi(z,z')} \int_{A\times A} d(x, y)\pi(dx, dy),$$

where $E_1 \subset \mathcal{C}(A, [0, 1])$ is the set of 1-Lipschitz functions from A to $[0, 1]$ and $\Pi(z, z') \subset \Delta(A\times A)$ is the set of measures on $A \times A$ with first marginal z and second marginal z'. Because A is compact, the infimum is a minimum. For $f \in \mathcal{C}(A, [0, 1])$, the linear extension of f is the function $\hat{f} \in \mathcal{C}(\Delta(A), [0, 1])$, defined for $z \in \Delta(A)$ by

$$\hat{f}(z) := \int_A f(x)z(dx).$$

A *gambling house* $\Gamma = (X, F, r)$ is defined by the following elements:

- X is the *state space*, which is assumed to be compact metric for some distance d.
- $F : (X, d) \rightrightarrows (\Delta(X), d_{KR})$ is a correspondence with a closed graph and nonempty values.
- $r : X \to [0, 1]$ is the *payoff function*, which is assumed to be continuous.

Remark 1. *Because the state space is compact, F is a closed graph correspondence if and only if it is an upper hemicontinuous correspondence with closed values.*

Let $x_0 \in X$ be an initial state. The gambling house starting from x_0 proceeds as follows. At each stage $m \geqslant 1$, the decision-maker chooses $z_m \in F(x_{m-1})$. A new state x_m is drawn from the probability distribution z_m, and the decision-maker gets the payoff $r(x_m)$.

For the definition of strategies, we follow Maitra and Sudderth [14, Chapter 2]. First, we need the following definition (see [9, Chapter 11, section 1.8]):

Definition 1. *Let $\nu \in \Delta(\Delta(X))$. The* barycenter *of ν is the probability measure $\mu = \mathrm{Bar}(\nu) \in \Delta(X)$ such that for all $f \in \mathcal{C}(X, [0, 1])$,*

$$\hat{f}(\mu) = \int_{\Delta(X)} \hat{f}(z)\nu(dz).$$

Given M a closed subset of $\Delta(X)$, we denote by $\mathrm{Sco}\, M$ *the* strong convex hull *of the set M, that is,*

$$\mathrm{Sco}\, M := \{\mathrm{Bar}(\nu), \nu \in \Delta(M)\}.$$

Equivalently, $\mathrm{Sco}\, M$ *is the closure of the convex hull of M.*

For every $m \geqslant 1$, we denote by $H_m := X^m$ the set of possible histories before stage m, which is compact for the product topology.

Definition 2. *A* behavior (resp. pure) strategy σ *is a sequence of mappings* $\sigma := (\sigma_m)_{m\geqslant 1}$ *such that for every* $m \geqslant 1$*,*

- $\sigma_m : H_m \to \Delta(X)$ *is (Borel) measurable,*
- *for all* $h_m = (x_0, ..., x_{m-1}) \in H_m$, $\sigma_m(h_m) \in \mathrm{Sco}(F(x_{m-1}))$ *(resp.* $\sigma_m(h_m) \in F(x_{m-1})$*).*

We denote by Σ *(resp.* Σ_p*) the set of behavior (resp. pure) strategies.*

Note that $\Sigma_p \subset \Sigma$. The following proposition ensures that Σ_p is nonempty. This is a special case of Kuratowski-Ryll-Nardzewski theorem (see [1, Theorem 18.13, p. 600].

Proposition 1. *Let* K_1 *and* K_2 *be two compact metric spaces, and* $\Phi : K_1 \rightrightarrows K_2$ *be a closed graph correspondence with nonempty values. Then* Φ *admits a measurable selector, that is, there exists a measurable mapping* $\varphi : K_1 \to K_2$ *such that for all* $k \in K_1$, $\varphi(k) \in K_2$*.*

Proof. In [1], the theorem is stated for weakly measurable correspondences. By [1, Theorem 18.10, p. 598] and [1, Theorem 18.20, p. 606], any correspondence satisfying the assumptions of the proposition is weakly measurable, thus the proposition holds. □

Definition 3. *A strategy* $\sigma \in \Sigma$ *is* Markov *if there exists a measurable mapping* $f : \mathbb{N}^* \times X \to \Delta(X)$ *such that for every* $h_m = (x_0, ..., x_{m-1}) \in H_m$, $\sigma(h_m) = f(m, x_{m-1})$*. When this is the case, we identify* σ *with* f*.*

A strategy σ *is* stationary *if there exists a measurable mapping* $f : X \to \Delta(X)$ *such that for every* $h_m = (x_0, ..., x_{m-1}) \in H_m$, $\sigma(h_m) = f(x_{m-1})$*. When this is the case, we identify* σ *with* f*.*

Let $H_\infty := X^{\mathbb{N}}$ be the set of all possible plays in the gambling house Γ. By the Kolmogorov extension theorem, an initial state $x_0 \in X$ and a behavior strategy σ determine a unique probability measure over H_∞, denoted by $\mathbb{P}_\sigma^{x_0}$.

Let $x_0 \in X$ and $n \geqslant 1$. The payoff in the n-stage problem starting from x_0 is defined for $\sigma \in \Sigma$ by

$$\gamma_n(x_0, \sigma) := \mathbb{E}_\sigma^{x_0}\left(\frac{1}{n}\sum_{m=1}^{n} r_m\right),$$

where $r_m := r(x_m)$ is the payoff at stage $m \in \mathbb{N}^*$. The value $v_n(x_0)$ of this problem is the maximum expected payoff with respect to behavior strategies:

$$v_n(x_0) := \sup_{\sigma\in\Sigma} \gamma_n(x_0, \sigma).$$

By Feinberg [11, Theorem 5.2], any behavior strategy can be assimilated to a probability measure on the set of pure strategies. It follows that the above supremum is reached at a pure strategy.

Remark 2. *For* $\mu \in \Delta(X)$*, one can also define the gambling house with initial distribution* μ*, where the initial state is drawn from* μ *and announced to the decision-maker. The definition of strategies and values are the same, and for all* $n \in \mathbb{N}^*$*, the value of the* n*-stage gambling house starting from* μ *is equal to* $\hat{v}_n(\mu)$*.*

1.2 Long-term gambling houses

1.2.1 Uniform value

Definition 4. *Let $x_0 \in X$. The gambling house $\Gamma(x_0)$ has an* asymptotic value $v_\infty(x_0) \in [0,1]$ *if the sequence $(v_n(x_0))_{n\geqslant 1}$ converges to $v_\infty(x_0)$.*

Definition 5. *Let $x_0 \in X$. The gambling house $\Gamma(x_0)$ has a* uniform value $v_\infty(x_0) \in [0,1]$ *in behavior (resp. pure) strategies if it has an asymptotic value $v_\infty(x_0)$ and for every $\varepsilon > 0$, there exists $n_0 \in \mathbb{N}^*$ and a behavior (resp. pure) strategy σ such that for all $n \geqslant n_0$,*

$$\gamma_n(x_0, \sigma) \geqslant v_\infty(x_0) - \varepsilon.$$

Definition 6. *A gambling house Γ is 1-Lipschitz if its correspondence F is 1-Lipschitz, that is, for every $x \in X$, every $u \in F(x)$ and every $y \in X$, there exists $w \in F(y)$ such that $d_{KR}(u,w) \leqslant d(x,y)$.*

Renault and Venel [16] have proved that any 1-Lipschitz gambling house has a uniform value in behavior strategies[2]. They asked about the existence of the uniform value in pure strategies. This is a recurring open problem in the literature. In the framework of POMDPs, this open problem already appeared in Rosenberg, Solan and Vieille [18] and in Renault [15].

1.2.2 The long-run average payoff criterion

To study long-term dynamic programming problems, an alternative to the uniform approach is to associate a payoff to each infinite history. Given an initial state $x_0 \in X$, the *infinitely repeated* gambling house $\Gamma_\infty(x_0)$ is the problem with strategy set Σ, and payoff function γ_∞ defined for all $\sigma \in \Sigma$ by

$$\gamma_\infty(x_0, \sigma) := \mathbb{E}_\sigma^{x_0}\left(\liminf_{n\to+\infty} \frac{1}{n}\sum_{m=1}^{n} r_m\right).$$

In the literature, the above payoff is often referred as the *long-run average payoff criterion* (see [3]). The value of $\Gamma_\infty(x_0)$ is

$$w_\infty(x_0) := \sup_{\sigma\in\Sigma} \gamma_\infty(x_0, \sigma).$$

Remark 3. *The above supremum may not be reached: there may not exist 0-optimal strategies in $\Gamma_\infty(x_0)$ (see for example Rosenberg, Solan and Vieille [18]).*

The following proposition plays a key role in this paper:

Proposition 2. *For all $\epsilon > 0$, there exists ϵ-optimal pure strategies in $\Gamma_\infty(x_0)$.*

Proof. Exactly like for the n-stage game, this result is a direct consequence of Theorem 5.2 in Feinberg [11]. □

If $\Gamma(x_0)$ has a uniform value $v_\infty(x_0)$, we have $w_\infty(x_0) \leqslant v_\infty(x_0)$ by the dominated convergence theorem. A natural question is to ask whether the equality holds. When this is the case, it significantly strengthens the notion of uniform value, as shown by the following example.

Example 1. *There are two states, x and x^*, and $F(x) = F(x^*) = \{x, x^*\}$. Moreover, $r(x) = 0$ and $r(x^*) = 1$. Thus, at each stage, the decision-maker has to choose between having a payoff 0 and having a payoff 1. Obviously, this problem has a uniform value equal to 1. Let $\epsilon > 0$. Let σ be the strategy such that for all $n \in \mathbb{N}$, at stage $2^{2^n} - 1$, the decision-maker chooses x with probability $\epsilon/2$, and sticks to this choice until stage $2^{2^{n+1}} - 1$; with probability $1-\epsilon/2$, he chooses*

[2] In fact, their model of gambling house is slightly different: they do not assume that F is closed-valued, but instead assume that it takes values in the set of probability measures on X with finite support.

x^*, and sticks to this choice until stage $2^{2^{n+1}} - 1$. The strategy σ is uniformly ϵ-optimal: there exists $n_0 \in \mathbb{N}^*$ such that for all $n \geqslant n_0$,

$$\gamma_n(x, \sigma) \geqslant 1 - \epsilon.$$

Nonetheless, by the law of large numbers, for any $n_0 \in \mathbb{N}^*$, there exists a random time T such that $\mathbb{P}_\sigma^x$ almost surely, $T \geqslant n_0$ and

$$\frac{1}{T} \sum_{m=1}^{T} r_m \leqslant \epsilon.$$

Therefore, the strategy σ does not guarantee more than ϵ in the game $\Gamma_\infty(x)$.

2 Main results

2.1 Gambling houses

We introduce a stronger notion of uniform value, which allows us to deal with the two open questions mentioned in the previous section at the same time.

Definition 7. *Let $x_0 \in X$. The gambling house $\Gamma(x_0)$ has a* pathwise uniform value *in behavior (resp. pure) strategies if*

- *The gambling house $\Gamma(x_0)$ has an asymptotic value $v_\infty(x_0)$.*
- *For all $\epsilon > 0$, there exists a behavior (resp. pure) strategy σ such that*

$$\gamma_\infty(x_0, \sigma) \geqslant v_\infty(x_0) - \epsilon.$$

A strategy σ satisfying the above equation is called pathwise ϵ-optimal strategy. When for all $x_0 \in X$, $\Gamma(x_0)$ has a pathwise uniform value in behavior (resp. pure) strategies, we say that Γ has a pathwise uniform value in behavior (resp. pure) strategies.

Proposition 2 implies that there exists a pathwise uniform value in behavior strategies if and only if there exists a pathwise uniform value in pure strategies. The following proposition shows that the concept of pathwise uniform value is more general than the concept of uniform value.

Proposition 3. *Assume that $\Gamma(x_0)$ has a pathwise uniform value (in behavior or pure strategies). Then it has a uniform value in pure strategies.*

Proof. By Proposition 2, $\Gamma(x_0)$ has a pathwise uniform value in pure strategies. Let $\epsilon > 0$, and σ be a pathwise ϵ-optimal pure strategy. We have

$$\mathbb{E}_\sigma^{x_0} \left(\liminf_{n \to +\infty} \frac{1}{n} \sum_{m=1}^{n} r_m \right) \geqslant v_\infty(x_0) - \epsilon.$$

By Fatou's lemma, it follows that

$$\liminf_{n \to +\infty} \mathbb{E}_\sigma^{x_0} \left(\frac{1}{n} \sum_{m=1}^{n} r_m \right) \geqslant v_\infty(x_0) - \epsilon,$$

and the gambling house $\Gamma(x_0)$ has a uniform value in pure strategies.

□

We can now state our main theorem concerning gambling houses:

Theorem 1. *Let Γ be a gambling house such that $\{v_n, n \geqslant 1\}$ is uniformly equicontinuous and w_∞ is continuous. Then Γ has a pathwise uniform value in pure strategies. Consequently, it has a uniform value in pure strategies, and*

$$w_\infty = v_\infty.$$

In particular, we obtain the following result.

Theorem 2. *Let Γ be a 1-Lipschitz gambling house. Then Γ has a pathwise uniform value in pure strategies. Consequently, it has a uniform value in pure strategies, and*

$$w_\infty = v_\infty.$$

In the two next subsections, we present similar results for MDPs and POMDPs.

2.2 MDPs

A Markov Decision Process (MDP) is a 4-uple $\Gamma = (K, I, g, q)$, where (K, d_K) is a compact metric state space, (I, d_I) is a compact metric action set, $g : K \times I \to [0, 1]$ is a continuous payoff function, and $q : K \times I \to \Delta(K)$ is a continuous transition function. As usual, the set $\Delta(K)$ is equipped wih the KR metric, and we assume that for all $i \in I$, $q(., i)$ is 1-Lipschitz. Given an initial state $k_1 \in K$ known by the decision-maker, the MDP $\Gamma(k_1)$ proceeds as follows. At each stage $m \geqslant 1$, the decision-maker chooses $i_m \in I$, and gets the payoff $g_m := g(k_m, i_m)$. A new state k_{m+1} is drawn from $q(k_m, i_m)$, and is announced to the decision-maker. Then, $\Gamma(k_1)$ moves on to stage $m + 1$. A behavior (resp. pure) strategy is a measurable map $\sigma : \cup_{m\geqslant 1} K \times (I \times K)^{m-1} \to \Delta(I)$ (resp. $\sigma : \cup_{m\geqslant 1} K \times (I \times K)^{m-1} \to I$). An initial state k_1 and a strategy σ induce a probability measure $\mathbb{P}_\sigma^{k_1}$ on the set of plays $H_\infty = (K \times I)^{\mathbb{N}^*}$.

The notion of uniform value is defined in the same way as in gambling houses. We prove the following theorem:

Theorem 3. *The MDP Γ has a pathwise uniform value in pure strategies, that is, for all $k_1 \in K$, the two following statements hold:*

- *The sequence $(v_n(k_1))$ converges when n goes to infinity to some real number $v_\infty(k_1)$.*
- *For all $\epsilon > 0$, there exists a pure strategy σ such that*

$$\mathbb{E}_\sigma^{k_1}\left(\liminf_{n\to+\infty} \frac{1}{n}\sum_{m=1}^{n} g(k_m, i_m)\right) \geqslant v_\infty(k_1) - \epsilon.$$

Consequently, the MDP Γ has a uniform value in pure strategies.

2.3 POMDPs

A Partially Observable Markov Decision Process (POMDP) is a 5-uple $\Gamma = (K, I, S, g, q)$, where K is a finite set space, I is a compact metric action set, S is a finite signal set, $g : K \times I \to [0, 1]$ is a continuous payoff function, and $q : K \times I \to \Delta(K \times S)$ is a continuous transition function. Given an initial distribution $p_1 \in \Delta(K)$, the POMDP $\Gamma(p_1)$ proceeds as follows. An initial state k_1 is drawn from p_1, and the decision-maker is not informed about it. At each stage $m \geqslant 1$, the decision-maker chooses $i_m \in I$, and gets the (unobserved) payoff $g(k_m, i_m)$. A pair (k_{m+1}, s_m) is drawn from $q(k_m, i_m)$, and the decision-maker receives the signal s_m. Then the game proceeds to stage $m + 1$. A behavior strategy (resp. pure strategy) is a measurable map $\sigma : \cup_{m\geqslant 1}(I\times S)^{m-1} \to \Delta(I)$ (resp. $\sigma : \cup_{m\geqslant 1}(I\times S)^{m-1} \to I$). An initial distribution $p_1 \in \Delta(K)$ and a strategy σ induce a probability measure $\mathbb{P}_\sigma^{p_1}$ on the set of plays $H_\infty := (K \times I \times S)^{\mathbb{N}^*}$.

The notion of uniform value is defined in the same way as in gambling houses. We prove the following theorem:

Theorem 4. *The POMDP Γ has a pathwise uniform value in pure strategies, that is, for all $p_1 \in \Delta(K)$, the two following statements hold:*

- *The sequence $(v_n(p_1))$ converges when n goes to infinity to some real number $v_\infty(p_1)$.*
- *For all $\epsilon > 0$, there exists a pure strategy σ such that*

$$\mathbb{E}_\sigma^{p_1}\left(\liminf_{n\to+\infty}\frac{1}{n}\sum_{m=1}^{n} g(k_m, i_m)\right) \geqslant v_\infty(p_1) - \epsilon.$$

Consequently, the POMDP Γ has a uniform value in pure strategies.

In particular, this theorem solves positively the open question mentioned in [18], [15] and [16]: finite POMDPs have a uniform value in pure strategies.

3 Proof of Theorem 1

Let $\Gamma = (X, F, r)$ be a gambling house such that $\{v_n, n \geqslant 1\} \cup \{w_\infty\}$ is uniformly equicontinuous. Let $v : X \to [0,1]$ be defined by $v := \limsup_{n\to+\infty} v_n$.

Let $x_0 \in X$ be an initial state. By Proposition 2, in order to prove Theorem 1, it is sufficient to prove that for all $\epsilon > 0$, there exists a behavior strategy σ such that

$$\gamma_\infty(x_0, \sigma) = \mathbb{E}_\sigma^{x_0}\left(\liminf_{n\to+\infty}\frac{1}{n}\sum_{m=1}^{n} r_m\right) \geqslant v(x_0) - \epsilon.$$

Let us first give the structure and the intuition of the proof. It builds on three main ideas, each of them corresponding to a lemma.

First, Lemma 1 associates to x_0 a probability measure $\mu^* \in \Delta(X)$, such that:

- Going from x_0, for all $\epsilon > 0$ and $n_0 \in \mathbb{N}^*$, there exists a strategy σ_0 and $n \geqslant n_0$ such that the occupation measure $\frac{1}{n}\sum_{m=1}^{n} z_m \in \Delta(X)$ is close to μ^* up to ϵ (for the KR distance).
- $\hat{r}(\mu^*) = \hat{v}(\mu^*) = v(x_0)$
- If the initial state is drawn according to μ^*, the decision-maker has a behavior stationary strategy σ^* such that for all $m \geqslant 1$, z_m is distributed according to μ^* (μ^* is an invariant measure for the gambling house).

Let x be in the support of μ^*. Building on a pathwise ergodic theorem, Lemma 2 shows that

$$\frac{1}{n}\sum_{m=1}^{n} r_m \to v(x) \quad \mathbb{P}_{\sigma^*}^{x} \text{ a.s.}$$

Let $y \in X$ be close to x. Lemma 3 shows that, if $y \in X$ is close to x, then there exists a behavior strategy σ such that $\gamma_\infty(y, \sigma)$ is close to $v(y)$.

These lemmas are put together in the following way. Lemma 1 implies that, going from x_0, the decision-maker has a strategy σ_0 such that there exists a (deterministic) stage $m \geqslant 1$ such that with high probability, the state x_m is close to the support of μ^*, and such that the expectation of $v(x_m)$ is close to $v(x_0)$. Let x be an element in the support of μ^* such that x_m is close to x. By Lemma 3, going from x_m, the decision-maker has a strategy σ such that $\gamma_\infty(x_m, \sigma)$ is close to $v(x_m)$. Let $\widetilde{\sigma}$ be the strategy that plays σ_0 until stage m, then switches to σ. Then $\gamma_\infty(x_0, \widetilde{\sigma})$ is close to $v(x_0)$, which concludes the proof of Theorem 1.

3.1 Preliminary results

Let $\Gamma = (X, F, r)$ be a gambling house. We define a relaxed version of the gambling house, in order to obtain a deterministic convex gambling house $H : \Delta(X) \rightrightarrows \Delta(X)$. The interpretation of $H(z)$ is the following: if the initial state is drawn according to z, $H(z)$ is the set of all possible measures on the next state that the decision-maker can generate by using behavior strategies.

First, we define $G : X \rightrightarrows \Delta(X)$ by

$$\forall x \in X \quad G(x) := \mathrm{Sco}(F(x)).$$

By [1, Theorem 17.35, p.573], the correspondence G has a closed graph, which is denoted by $\mathrm{Graph}\, G$. Note that a behavior strategy in the gambling house Γ corresponds to a pure strategy in the gambling house (X, G, r). For every $z \in \Delta(X)$, we define $H(z)$ by

$$H(z) := \Bigg\{ \mu \in \Delta(X) \mid \exists\, \sigma : X \to \Delta(X) \text{ measurable s.t. } \forall x \in X,\ \sigma(x) \in G(x) \text{ and}$$
$$\forall f \in \mathcal{C}(X, [0,1]),\ \hat{f}(\mu) = \int_X \hat{f}(\sigma(x)) z(dx) \Bigg\}.$$

Note that replacing "$\forall x \in X,\ \sigma(x) \in G(x)$" by "$\forall x \in X,\ \sigma(x) \in G(x)\ z-a.s.$" does not change the above definition (throughout the paper, "a.s." stands for "almost surely").

By Proposition 1, H has nonempty values. We now check that the correspondence H has a closed graph.

Proposition 4. *The correspondence H has a closed graph.*

Proof. Let $(z_n, \mu_n)_{n\in\mathbb{N}} \in (\mathrm{Graph}\, H)^{\mathbb{N}}$ such that $(z_n, \mu_n)_{n\in\mathbb{N}}$ converges to some $(z, \mu) \in \Delta(X) \times \Delta(X)$. Let us show that $\mu \in H(z)$. For this, we construct $\sigma : X \to \Delta(X)$ associated to μ in the definition of $H(z)$.

By definition of H, for every $n \in \mathbb{N}$, there exists $\sigma_n : X \to \Delta(X)$ a measurable selector of G such that for every $f \in \mathcal{C}(X, [0,1])$,

$$\hat{f}(\mu_n) = \int_X \hat{f}(\sigma_n(x)) z_n(dx).$$

Let $\pi_n \in \Delta(\mathrm{Graph}\, G)$ such that the first marginal of π_n is z_n, and the conditional distribution of π_n knowing $x \in X$ is $\delta_{\sigma_n(x)} \in \Delta(\Delta(X))$. By definition, for every $f \in \mathcal{C}(X, [0,1])$, we have

$$\begin{aligned}
\int_{X\times\Delta(X)} \hat{f}(p)\pi_n(dx, dp) &= \int_X \left(\int_{\Delta(X)} \hat{f}(p)\delta_{\sigma_n(x)}(dp) \right) z_n(dx) \\
&= \int_X \hat{f}(\sigma_n(x)) z_n(dx) \\
&= \hat{f}(\mu_n).
\end{aligned}$$

The set $\Delta(\mathrm{Graph}\, G)$ is compact, thus there exists π a limit point of the sequence $(\pi_n)_{n\in\mathbb{N}}$. By definition of the weak* topology on $\Delta(X)$ and on $\Delta(\mathrm{Graph}\, G)$, the previous equation yields

$$\int_{X\times\Delta(X)} \hat{f}(p)\pi(dx, dp) = \hat{f}(\mu). \tag{1}$$

To conclude, let us disintegrate π. Let z' be the first marginal of π. The sets X and $\Delta(X)$ are compact metric spaces, thus there exists a probability kernel $K : X \times \mathcal{B}(\Delta(X)) \to [0,1]$ such that

- for every $x \in X$, $K(x,.) \in \Delta(\Delta(X))$,
- for every $B \in \mathcal{B}(\Delta(X))$, $K(.,B)$ is measurable,
- for every $h \in \mathcal{C}(X \times \Delta(X), [0,1])$,

$$\int_{X\times\Delta(X)} h(x,p)\pi(dx,dp) = \int_X \left(\int_{\Delta(X)} h(x,p)K(x,dp) \right) z'(dx). \tag{2}$$

Note that the second condition is equivalent to: "The mapping $x \to K(x,.)$ is measurable" (see [5, Proposition 7.26, p.134]). For every $n \geqslant 1$, the first marginal of π_n is equal to z_n that converges to z, thus $z' = z$. Define a measurable mapping $\sigma : X \to \Delta(X)$ by $\sigma(x) := \mathrm{Bar}(K(x,.)) \in \Delta(X)$. Because $\pi \in \Delta(\mathrm{Graph}\, G)$, we have $\sigma(x) \in G(x)$ $z-$a.s. Let $f \in \mathcal{C}(X,[0,1])$. Using successively (1) and (2) yield

$$\begin{aligned} \hat{f}(\mu) &= \int_{X\times\Delta(X)} \hat{f}(p)\pi(dx,dp) \\ &= \int_X \left(\int_{\Delta(X)} \hat{f}(p)K(x,dp) \right) z(dx) \\ &= \int_X \hat{f}(\sigma(x))z(dx). \end{aligned}$$

Thus, $\mu \in H(z)$, and H has a closed graph. □

Let $\mu, \mu' \in \Delta(X)$. Denote $\lambda \cdot \mu + (1-\lambda)\cdot\mu'$ the probability measure $\mu'' \in \Delta(X)$ such that for all $f \in \mathcal{C}(X,[0,1])$,

$$\hat{f}(\mu'') = \lambda\hat{f}(\mu) + (1-\lambda)\hat{f}(\mu').$$

For $(\mu_m)_{m\in\mathbb{N}^*} \in \Delta(X)^{\mathbb{N}^*}$ and $n \in \mathbb{N}^*$, the measure $\dfrac{1}{n}\displaystyle\sum_{m=1}^{n} \mu_m$ is defined in a similar way.

Proposition 5. *The correspondence H is linear on $\Delta(X)$:*

$$\forall z, z' \in \Delta(X),\ \forall \lambda \in [0,1],\ H(\lambda \cdot z + (1-\lambda)\cdot z') = \lambda \cdot H(z) + (1-\lambda)\cdot H(z').$$

Proof. Let $z, z' \in \Delta(X)$ and $\lambda \in [0,1]$, then the inclusion

$$H(\lambda \cdot z + (1-\lambda)\cdot z') \subset \lambda \cdot H(z) + (1-\lambda)\cdot H(z')$$

is immediate. We now prove the converse inclusion. Let $\mu \in \lambda \cdot H(z) + (1-\lambda)\cdot H(z')$. By definition, there exists $\sigma : X \to \Delta(X)$ and $\sigma' : X \to \Delta(X)$ two measurable selectors of G such that for every $f \in \mathcal{C}(X,[0,1])$,

$$\hat{f}(\mu) = \lambda \int_X \hat{f}(\sigma(x))z(dx) + (1-\lambda)\int_X \hat{f}(\sigma'(x))z'(dx).$$

Denote by π (resp. π'), the probability distribution on $X \times \Delta(X)$ generated by z and σ (resp. z' and σ'). Let $\pi'' := \lambda\cdot\pi + (1-\lambda)\cdot\pi'$, then π'' is a probability on $X \times \Delta(X)$ such that $\pi''(\mathrm{Graph}(G)) = 1$, and the marginal on X is $\lambda \cdot z + (1-\lambda)\cdot z'$. Let $\sigma'' : X \to \Delta(X)$ given by the disintegration of π'' with respect to the first coordinate. Let $f \in \mathcal{C}(X,[0,1])$. As in the proof of Proposition 4 (see Equation (1)), we have

$$\begin{aligned} \hat{f}(\mu) &= \lambda \int_{X\times\Delta(X)} \hat{f}(p)\pi(dx,dp) + (1-\lambda)\int_{X\times\Delta(X)} \hat{f}(p)\pi'(dx,dp) \\ &= \int_{X\times\Delta(X)} \hat{f}(p)\pi''(dx,dp) \\ &= \int_X \hat{f}(\sigma''(x))z(dx), \end{aligned}$$

thus $\mu \in H(\lambda \cdot z + (1-\lambda) \cdot z')$.

□

3.2 Invariant measure

The first lemma associates a fixed point of the correspondence H to each initial state:

Lemma 1. *Let $x_0 \in X$. There exists a distribution $\mu^* \in \Delta(X)$ such that*

- *μ^* is H-invariant: $\mu^* \in H(\mu^*)$,*
- *for every $\varepsilon > 0$ and $N \geqslant 1$, there exists a (pure) strategy σ_0 and $n \geqslant N$ such that σ is 0-optimal in $\Gamma_n(x_0)$, $v_n(x_0) \geqslant v(x_0) - \epsilon$ and*

$$d_{KR}\left(\frac{1}{n}\sum_{m=1}^{n} z_m(x_0,\sigma), \mu^*\right) \leqslant \varepsilon,$$

  where $z_m(x_0,\sigma_0) \in \Delta(X)$ is the distribution of x_m, the state at stage m, given the initial state x_0 and the strategy σ_0.

- *$\hat{r}(\mu^*) = \hat{v}(\mu^*) = v(x_0)$.*

Proof. The proof builds on the same ideas as in Renault and Venel [16, Proposition 3.24, p. 28]. Let $n \in \mathbb{N}^*$ and σ_0 be a pure optimal strategy in the n-stage problem $\Gamma_n(x_0)$.

Let

$$z_n := \frac{1}{n}\sum_{m=1}^{n} z_m(x_0,\sigma_0),$$

and

$$z'_n := \frac{1}{n}\sum_{m=2}^{n+1} z_m(x_0,\sigma_0).$$

By construction, for every $m \in \{1,2,...,n\}$, $z_{m+1}(x_0,\sigma_0) \in H(z_m(x_0,\sigma_0))$, therefore by linearity of H (see Proposition 5)

$$z'_n \in H(z_n).$$

Moreover, we have

$$d_{KR}(z_n, z'_n) \leqslant \frac{2}{n}\operatorname{diam}(X), \tag{3}$$

where $\operatorname{diam}(X)$ is the diameter of X.

The set $\Delta(X)$ is compact. Up to taking a subsequence, there exists $\mu^* \in \Delta(X)$ such that $(v_n(x_0))$ converges to $v(x_0)$ and (z_n) converges to μ^*. By inequality (3), (z'_n) also converges to μ^*. Because H has a closed graph, we have $\mu^* \in H(\mu^*)$, and μ^* is H-invariant. By construction, the second property is immediate.

Finally, we have a series of inequalities that imply the third property.

- v is decreasing in expectation along trajectories: the sequence $(\hat{v}(z_m(x_0,\sigma_0)))_{m\geqslant 1}$ is decreasing, thus for every $n \geqslant 1$,

$$v(x_0) \geqslant \frac{1}{n}\sum_{m=1}^{n} \hat{v}(z_m(x_0,\sigma_0)) = \hat{v}(z_n).$$

  Taking n to infinity, by continuity of $\hat{v}$, we obtain that $v(x_0) \geqslant \hat{v}(\mu^*)$.

- We showed that $\mu^* \in H(\mu^*)$. Let $\sigma^* : X \to \Delta(X)$ be the corresponding measurable selector of G. Let us consider the gambling house $\Gamma(\mu^*)$, where the initial state is drawn from μ^* and announced to the decision-maker (see Remark 2). The map σ^* is a stationary strategy in $\Gamma(\mu^*)$, and for all $m \geqslant 1$, $z_m(\mu^*, \sigma^*) = \mu^*$. Consequently, for all $n \in \mathbb{N}^*$, the strategy σ^* guarantees $\hat{r}(\mu^*)$ in $\Gamma_n(\mu^*)$. Thus, we have

$$\hat{v}(\mu^*) \geqslant \hat{r}(\mu^*).$$

- By construction, the payoff is linear on $\Delta(X)$ and $\hat{r}(z_n) = v_n(x_0)$. By continuity of $\hat{r}$, taking n to infinity, we obtain

$$\hat{r}(\mu^*) = v(x_0).$$

□

In the next section, we prove that in $\Gamma(\mu^*)$, under the strategy σ^*, the average payoffs converge almost surely to $v(x)$, where x is the initial (random) state.

3.3 Pathwise ergodic theorem

We recall here the ergodic theorem in Hernández-Lerma and Lasserre [13, Theorem 2.5.1, p. 37].

Theorem 5 (pathwise ergodic theorem)**.** *Let $(X, \mathcal{B})$ be a measurable space, and ξ be a Markov chain on $(X, \mathcal{B})$, with transition probability function P. Let μ be an invariant probability measure for P. For every f an integrable function with respect to μ, there exist a set $B_f \in \mathcal{B}$ and a function f^* integrable with respect to μ, such that $\mu(B_f) = 1$, and for all $x \in B_f$,*

$$\frac{1}{n} \sum_{m=1}^{n} f(\xi_m) \to f^*(\xi_0) \quad P_x - a.s.$$

Moreover,

$$\int_X f^*(x)\mu(dx) = \int_X f(x)\mu(dx).$$

Lemma 2. *Let $x_0 \in X$ and $\mu^* \in \Delta(X)$ be the corresponding invariant measure (see Lemma 1). There exist a measurable set $B \subset \Delta(X)$ such that $\mu^*(B) = 1$ and a stationary strategy $\sigma^* : X \to \Delta(X)$ such that for all $x \in B$,*

$$\frac{1}{n} \sum_{m=1}^{n} r_m \to v(x) \quad \mathbb{P}^x_{\sigma^*} - a.s.$$

Proof. Because μ^* is a fixed point of H, there exists $\sigma^* : X \to \Delta(X)$ a measurable selector of G (thus, a behavior stationary strategy in Γ) such that for all $f \in \mathcal{C}(X, [0, 1])$,

$$\hat{f}(\mu^*) = \int_X \hat{f}(\sigma^*(x))\mu^*(dx).$$

Consider the gambling house $\Gamma(\mu^*)$. Under σ^*, the sequence of states $(x_m)_{m\in\mathbb{N}}$ is a Markov chain with invariant measure μ^*. From Theorem 5, there exist a measurable set $B_0 \subset X$ such that $\mu^*(B_0) = 1$, and a measurable map $w : X \to [0, 1]$ such that for all $x \in B_0$, we have

$$\frac{1}{n} \sum_{m=1}^{n} r(x_m) \underset{n\to+\infty}{\to} w(x) \quad \mathbb{P}^x_{\sigma^*} - \text{almost surely},$$

and

$$\hat{w}(\mu^*) = \hat{r}(\mu^*).$$

We now prove that $w = v \quad \mathbb{P}_{\sigma^*}^{\mu^*}$ – a.s.. First, we prove that $w \leqslant v \quad \mathbb{P}_{\sigma^*}^{\mu^*}$ – a.s.. Let $x \in B_0$. Using first the dominated convergence theorem, then the definition of $v_n(x)$, we have

$$\begin{aligned} w(x) &= \mathbb{E}_{\sigma^*}^x \left(\lim_{n\to+\infty} \frac{1}{n} \sum_{m=1}^n r(x_m) \right) \\ &= \lim_{n\to+\infty} \mathbb{E}_{\sigma^*}^x \left(\frac{1}{n} \sum_{m=1}^n r(x_m) \right) \\ &\leqslant \limsup_{n\to+\infty} v_n(x) = v(x). \end{aligned}$$

Moreover, we know by Lemma 1 that $\hat{r}(\mu^*) = \hat{v}(\mu^*)$, therefore

$$\hat{w}(\mu^*) = \hat{r}(\mu^*) = \hat{v}(\mu^*).$$

This implies that $w = v \quad \mathbb{P}_{\sigma^*}^{\mu^*}$ – a.s., and the lemma is proved. □

3.4 Junction lemma

By assumption, $\{v_n, n \geqslant 1\} \cup \{w_\infty\}$ is uniformly equicontinuous. Therefore, there exists an increasing modulus of continuity $\eta : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\forall x, y \in X,\ |w_\infty(x) - w_\infty(y)| \leqslant \eta(d(x,y)),$$

and for all $n \geqslant 1$,

$$\forall x, y \in X,\ |v_n(x) - v_n(y)| \leqslant \eta(d(x,y)).$$

Then, v is also uniformly continuous with the same modulus of continuity.

Lemma 3. *Let $\varepsilon > 0$, $x, y \in X$ and σ^* be a strategy such that*

$$\frac{1}{n} \sum_{m=1}^n r_m \to v(x) \quad \mathbb{P}_{\sigma^*}^x \text{ a.s.}$$

Then there exists a strategy σ such that

$$\mathbb{E}_\sigma^y \left(\liminf_{n\to+\infty} \frac{1}{n} \sum_{m=1}^n r_m \right) \geqslant v(y) - 2\eta(d(x,y)) - \varepsilon.$$

Proof. By assumption, we have

$$\mathbb{E}_{\sigma^*}^x \left(\liminf_{n\to+\infty} \frac{1}{n} \sum_{m=1}^n r_m \right) = \mathbb{E}_{\sigma^*}^x \left(v(x) \right) = v(x),$$

therefore $v(x) \geqslant w_\infty(x)$. Moreover, by Fatou's lemma, $w_\infty(x) \leqslant v(x)$. Thus, $w_\infty(x) = v(x)$.

Let $\varepsilon > 0$. By definition of $w_\infty(y)$, there exists a strategy σ such that

$$\begin{aligned} \mathbb{E}_\sigma^y \left(\liminf_{n\to+\infty} \frac{1}{n} \sum_{m=1}^n r_m \right) &\geqslant w_\infty(y) - \varepsilon, \\ &\geqslant w_\infty(x) - \eta(d(x,y)) - \varepsilon, \\ &= v(x) - \eta(d(x,y)) - \varepsilon, \\ &\geqslant v(y) - 2\eta(d(x,y)) - \varepsilon. \end{aligned}$$

□

We can now finish the proof of Theorem 1.

3.5 Conclusion of the proof

Proof of Theorem 1. We can now put Lemma 1, 2 and 3 together to finish the proof of Theorem 1. Fix an initial state $x_0 \in X$ and $\epsilon > 0$. We will define a strategy $\widetilde{\sigma}$ as follows: start by following a strategy σ_0 until some stage n_3, then switch to another strategy depending on the state x_{n_3}. We first define the stage n_3, then build the strategy $\widetilde{\sigma}$ and finally check that this strategy indeed guarantees a good long-run average payoff.

By assumption, the family $(v_n)_{n\geqslant 1}$ is uniformly equicontinuous. Consequently, there exists $n_0 \in \mathbb{N}^*$ such that for all $n \geqslant n_0$ and for all $x \in X$,

$$v_n(x) \leqslant v(x) + \epsilon.$$

We first consider Lemma 1 for x_0, $\epsilon' = \epsilon^3$ and $N = 2n_0$. There exists μ^* an invariant measure, σ_0 a (pure) strategy and $n_1 \geqslant 2n_0$ such that μ^* satisfies the conclusion of Lemma 1 and

$$d_{KR}\left(\frac{1}{n_1}\sum_{m=1}^{n_1} z_m(x_0,\sigma_0), \mu^*\right) \leqslant \epsilon^3.$$

Let B be given by Lemma 2. In general, there is no hope to prove the existence of a stage m such that $z_m(x_0,\sigma_0)$ is close to μ^*. Instead, we prove the existence of a stage n_3 such that under the strategy σ_0, x_{n_3} is with high probability close to B, and $v(z_{n_3}(x_0,\sigma_0))$ is close to $v(x_0)$.

Let $n_2 = \lfloor \epsilon n_1 \rfloor + 1$, $A = \{x \in X | d(x,B) \leqslant \varepsilon\}$ and $A^c = \{x \in X | d(x,B) > \varepsilon\}$. We denote $\mu_{n_1} = \frac{1}{n_1}\sum_{m=1}^{n_1} z_m(x_0,\sigma_0)$. By property of the KR distance, there exists a coupling $\gamma \in \Delta(X \times X)$ such that the first marginal of γ is μ_{n_1}, the second marginal is μ^*, and

$$d_{KR}(\mu_{n_1},\mu^*) \quad = \quad \int_{X^2} d(x,x')\gamma(dx,dx').$$

By definition of A, for all $(x,x') \in A^c \times B$, we have $d(x,x') > \epsilon$. Thus, Markov inequality yields

$$\begin{aligned}\int_{X^2} d(x,x')\gamma(dx,dx') &\geqslant \epsilon\gamma(A^c \times B)\\ &= \varepsilon\mu_{n_1}(A^c).\end{aligned}$$

We deduce that $\mu_{n_1}(A^c) \leqslant \varepsilon^2$. Because the n_2 first stages have a weight of order ε in μ_{n_1}, we deduce the existence of a stage m such that $z_m(A^c) \leqslant \epsilon$:

$$\begin{aligned}\mu_{n_1}(A^c) &= \frac{1}{n_1}\sum_{m=1}^{n_1} z_m(A^c)\\ &= \frac{1}{n_1}\sum_{m=1}^{n_2} z_m(A^c) + \frac{1}{n_1}\sum_{m=n_2+1}^{n_1} z_m(A^c)\\ &\geqslant \epsilon \min_{1\leqslant m\leqslant n_2} z_m(A^c),\end{aligned}$$

and thus

$$z_{n_3}(A^c) := \min_{1\leqslant m\leqslant n_2} z_m(A^c) \leqslant \epsilon. \tag{4}$$

Moreover, $\hat{v}(z_{n_3}(x_0,\sigma_0))$ is greater than $v(x_0)$ up to a margin ε. Indeed we have

$$\begin{aligned}\hat{v}(z_{n_3}(x_0,\sigma_0)) &\geqslant v_{n_1-n_3+1}(z_{n_3}(x_0,\sigma_0)) - \varepsilon\\ &\geqslant v_{n_1}(x_0) - \frac{n_3-1}{n_1} - \varepsilon\\ &\geqslant v(x_0) - 2\varepsilon - \varepsilon.\\ &\geqslant v(x_0) - 3\varepsilon.\end{aligned}$$

Using Equation (4) and the last inequality, we deduce that

$$\mathbb{E}_{\sigma_0}^{x_0}(1_A v(x_{n_3})) \geqslant \mathbb{E}_{\sigma_0}^{x_0}(v(x_{n_3})) - z_{n_3}(A^c) \geqslant v(x_0) - 4\varepsilon.$$

We have defined both the initial strategy σ_0 and the switching stage n_3. To conclude, we use Lemma 3 in order to define the strategy from stage n_3. Note that in Lemma 3, we did not prove that the strategy σ could be selected in a measurable way with respect to the state. Thus, we need to use a finite approximation. The set X is a compact metric set, thus there exists a partition $\{\mathcal{P}^1, ..., \mathcal{P}^L\}$ of X such that for every $l \in \{1, ..., L\}$, $\mathcal{P}^l$ is measurable and $\text{diam}(P^l) \leqslant \epsilon$. It follows that there exists a finite subset $\{x^1, ..., x^L\}$ of B such that for every $x \in A \cap \mathcal{P}^l$, $d(x, x^l) \leqslant 3\varepsilon$. We denote by ψ the application which associates to every $x \in A \cap \mathcal{P}^l$ the state x^l.

We define the strategy $\widetilde{\sigma}$ as follows:

- Play σ_0 until stage n_3.
- If $x_{n_3} \in A$, then there exists $l \in \{1, ..., L\}$ such that $x_{n_3} \in \mathcal{P}^l$. Play the strategy given by Lemma 3, with $x = x^l$ and $y = x_{n_3}$. If $x_{n_3} \notin A$, play any strategy.

Let us check that the strategy $\widetilde{\sigma}$ guarantees a good payoff with respect to the long-run average payoff criterion. By definition, we have

$$\begin{aligned}
\gamma_\infty(x_0, \widetilde{\sigma}) &= \mathbb{E}_{\widetilde{\sigma}}^{x_0}\left(\liminf_{n\to+\infty} \frac{1}{n}\sum_{m=1}^{n} r_m\right) \\
&= \mathbb{E}_{\widetilde{\sigma}}^{x_0}\left(\mathbb{E}_{\widetilde{\sigma}}^{x_0}\left(\liminf_{n\to+\infty} \frac{1}{n}\sum_{m=1}^{n} r_m \middle| x_{n_3}\right)\right) \\
&\geqslant \mathbb{E}_{\sigma_0}^{x_0}\left([v(x_{n_3}) - 2\eta(d(x_{n_3}, \psi(x_{n_3}))) - \varepsilon]1_A\right) \\
&\geqslant v(x_0) - 5\varepsilon - 2\eta(3\varepsilon).
\end{aligned}$$

Because $\eta(0) = 0$ and η is continuous at 0, the gambling house $\Gamma(x_0)$ has a pathwise uniform value, and Theorem 1 is proved. □

4 Proofs of Theorem 2, Theorem 3 and Theorem 4

This section is dedicated to the proofs of Theorem 2, Theorem 3 and Theorem 4. Theorem 2 and Theorem 3 stem from Theorem 1. Theorem 4 is not a corollary of Theorem 1. Indeed, applying Theorem 1 to the framework POMDPs, would only yield the existence of the uniform value in pure strategies and not the existence of the pathwise uniform value.

4.1 Proof of Theorem 2

Let $\Gamma := (X, F, r)$ be a gambling house such that F is 1-Lipschitz. Without loss of generality, we can assume that r is 1-Lipschitz. Indeed, any continuous payoff function can be uniformly approximated by Lipschitz payoff functions, and dividing the payoff function by a constant does not change the decision problem.

In order to prove Theorem 2, it is sufficient to prove that for all $n \geqslant 1$, v_n is 1-Lipschitz, and w_∞ is 1-Lipschitz. Indeed, it implies that the family $\{v_n, n \geqslant 1\}$ is uniformly equicontinuous and w_∞ is continuous. Theorem 2 then stems from Theorem 1.

Recall that $G : X \rightrightarrows \Delta(X)$ is defined for all $x \in X$ by $G(x) := \text{Sco}\, F(x)$.

Lemma 4. *The correspondence G is 1-Lipschitz.*

Proof. Let x and x' be two states in X. Fix $\mu \in G(x)$. Let us show that there exists $\mu' \in G(x')$ such that $d_{KR}(\mu, \mu') \leqslant d(x, x')$.

By definition of $G(x)$, there exists $\nu \in \Delta(F(x))$ such that for all $g \in \mathcal{C}(X, [0,1])$,

$$\hat{g}(\mu) = \int_{\Delta(X)} \hat{g}(z)\nu(dz).$$

Let $M = F(x) \subset \Delta(X)$. We consider the correspondence $\Phi : M \rightrightarrows \Delta(X)$ defined for $z \in M$ by

$$\Phi(z) := \{z' \in F(x') \mid d_{KR}(z, z') \leqslant d(x, x')\}.$$

Because F is 1-Lipschitz, Φ has nonempty values. Moreover, Φ is the intersection of two correspondences with a closed graph, therefore it is a correspondence with a closed graph. Applying Proposition 1, we deduce that Φ has a measurable selector $\varphi : M \to \Delta(X)$.

Let $\nu' \in \Delta(\Delta(X))$ be the image measure of ν by φ. Throughout the paper, we use the following notation for image measures:

$$\nu' := \nu \circ \varphi^{-1}.$$

By construction, $\nu'(F(x')) = 1$ and for all $h \in \mathcal{C}(\Delta(X), [0,1])$,

$$\int_{\Delta(X)} h(\varphi(z))\nu(dz) = \int_{\Delta(X)} h(u)\nu'(du).$$

Let $\mu' := \mathrm{Bar}(\nu')$ and $f \in E_1$. The function $\hat{f}$ is 1-Lipschitz, and

$$\begin{aligned}
\left|\hat{f}(\mu) - \hat{f}(\mu')\right| &= \left|\int_{\Delta(X)} \hat{f}(z)\nu(dz) - \int_{\Delta(X)} \hat{f}(u)\nu'(du)\right| \\
&= \left|\int_{\Delta(X)} \hat{f}(z)\nu(dz) - \int_{\Delta(X)} \hat{f}(\varphi(z))\nu(dz)\right| \\
&\leqslant \int_{\Delta(X)} \left|\hat{f}(z) - \hat{f}(\varphi(z))\right| \nu(dz) \\
&\leqslant d(x, x').
\end{aligned}$$

□

Because G is 1-Lipschitz, given $(x, u) \in \mathrm{Graph}\, G$ and $y \in X$, there exists $w \in G(y)$ such that $d_{KR}(u, w) \leqslant d(x, y)$. For our purpose, we need that the optimal coupling between u and w can be selected in a measurable way. This is the aim of the following lemma:

Lemma 5. *There exists a measurable mapping $\psi : \mathrm{Graph}\, G \times X \to \Delta(X \times X)$ such that for all $(x, u) \in \mathrm{Graph}\, G$, for all $y \in X$,*

- *the first marginal of $\psi(x, u, y)$ is u,*
- *the second marginal of $\psi(x, u, y)$ is in $G(y)$,*
- $\displaystyle\int_{X \times X} d(s, t)\psi(x, u, y)(ds, dt) \leqslant d(x, y).$

Proof. Let $S := \mathrm{Graph}(G) \times X$, $X' := \Delta(X \times X)$ and $\Xi : S \rightrightarrows X'$ the correspondence defined for all $(x, u, y) \in S$ by

$$\Xi(x, u, y) = \{\pi \in \Delta(X \times X) \mid \pi_1 = u, \pi_2 \in G(y)\},$$

where π_1 (resp. π_2) denotes the first (resp. second) marginal of π. The correspondence Ξ has a closed graph. Let $f : X' \to \mathbb{R}$ defined by

$$f(\pi) := \int_{X\times X} d(s,t)\pi(ds,dt).$$

The function f is continuous. Applying the measurable maximum theorem (see [1, Theorem 18.19, p.605]), we obtain that the correspondence $s \to \underset{\pi\in\Xi(s)}{\operatorname{argmin}} f(\pi)$ has a measurable selector, which proves the lemma. □

Proposition 6. *Let $x, y \in X$ and σ be a strategy. Then there exist a probability measure $\mathbb{P}_\sigma^{x,y}$ on $H_\infty \times H_\infty$, and a strategy τ such that:*

- *$\mathbb{P}_\sigma^{x,y}$ has first marginal $\mathbb{P}_\sigma^x$,*
- *$\mathbb{P}_\sigma^{x,y}$ has second marginal $\mathbb{P}_\tau^y$,*
- *The following inequalities holds: for every $n \geqslant 1$*

$$\mathbb{E}_\sigma^{x,y}\left(\frac{1}{n}\sum_{m=1}^{n}|r(X_m)-r(Y_m)|\right) \leqslant d(x,y),$$

and

$$\mathbb{E}_\sigma^{x,y}\left(\limsup_{n\to+\infty}\frac{1}{n}\sum_{m=1}^{n}|r(X_m)-r(Y_m)|\right) \leqslant d(x,y),$$

where X_m (resp. Y_m) is the m-th coordinate of the first (resp. second) infinite history.

Proof. Define the stochastic process $(X_m, Y_m)_{m\geqslant 0}$ on $(X \times X)^{\mathbb{N}}$ such that the conditional distribution of (X_m, Y_m) knowing $(X_l, Y_l)_{0\leqslant l\leqslant m-1}$ is

$$\psi(X_{m-1}, \sigma(X_0, ..., X_{m-1}), Y_{m-1}),$$

with ψ defined as in Lemma 5. Let $\mathbb{P}_\sigma^{x,y}$ be the law on H_∞^2 induced by this stochastic process and the initial distribution $\delta_{(x,y)}$. By construction, the first marginal of $\mathbb{P}_\sigma^{x,y}$ is $\mathbb{P}_\sigma^x$.

For $m \in \mathbb{N}^*$ and $(y_0, ..., y_{m-1}) \in X^m$, define $\tau_m(y_0, ..., y_{m-1}) \in \Delta(X)$ as being the law of Y_m, conditional to $Y_0 = y_0, ..., Y_{m-1} = y_{m-1}$. By convexity of G, this defines a (behavior) strategy τ in the game Γ. Moreover, the probability measure $\mathbb{P}_\tau^y$ is equal to the second marginal of $\mathbb{P}_\sigma^{x,y}$.

For all $m \in \mathbb{N}^*$, we have $\mathbb{P}_\sigma^{x,y}$-almost surely

$$\begin{aligned}\mathbb{E}_\sigma^{x,y}\left(d(X_m,Y_m)|X_{m-1},Y_{m-1}\right) &= \int_{X\times X} d(s',t')\psi(X_{m-1},\sigma(X_0,...,X_{m-1}),Y_{m-1})(ds',dt'),\\ &\leqslant d(X_{m-1},Y_{m-1}).\end{aligned}$$

The random process $(d(X_m, Y_m))_{m\geqslant 0}$ is a positive supermartingale. Therefore, we have

$$\begin{aligned}\mathbb{E}_\sigma^{x,y}\left(\frac{1}{n}\sum_{m=1}^{n}|r(X_m)-r(Y_m)|\right) &\leqslant \mathbb{E}_\sigma^{x,y}\left(\frac{1}{n}\sum_{m=1}^{n}d(X_m,Y_m)\right),\\ &= \frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_\sigma^{x,y}\left(d(X_m,Y_m)\right),\\ &\leqslant d(x,y).\end{aligned}$$

Moreover, the random process $(d(X_m, Y_m))_{m\geqslant 0}$ converges $\mathbb{P}_\sigma^{x,y}$-almost surely to a random variable D, such that $\mathbb{E}_\sigma^{x,y}(D) \leqslant d(x,y)$. For every $n \geqslant 1$, we have

$$\frac{1}{n}\sum_{m=1}^{n} |r(X_m) - r(Y_m)| \quad \leqslant \quad \frac{1}{n}\sum_{m=1}^{n} d(X_m, Y_m)$$

and the Cesàro theorem yields

$$\limsup_{n\to+\infty} \frac{1}{n}\sum_{m=1}^{n} |r(X_m) - r(Y_m)| \leqslant D \quad \mathbb{P}_\sigma^{x,y} \text{ a.s.}$$

Integrating the last inequality yields the proposition. □

Proposition 6 implies that for all $n \geqslant 1$, v_n is 1-Lipschitz, and that w_∞ is 1-Lipschitz. Thus, Theorem 2 holds.

4.2 Proof of Theorem 3 for MDPs

In this subsection, we consider a MDP $\Gamma = (K, I, g, q)$, as described in Subsection 2.2: the state space (K, d_K) and the action set (I, d_I) are compact metric, and the transition function q and the payoff function g are continuous. As in the previous section, without loss of generality we assume that the payoff function g is in fact 1-Lipschitz.

In the model of gambling house, there is no explicit set of actions. In order to apply Theorem 1 to Γ, we put the action played in the state variable. Indeed, we consider an auxiliary gambling house $\widetilde{\Gamma}$, with state space $K \times I \times K$. At each stage $m \geqslant 1$, the state x_m in the gambling house corresponds to the state (k_m, i_m, k_{m+1}) in the MDP. Formally, $\widetilde{\Gamma}$ is defined as follows:

- The state space is $X := K \times I \times K$, equipped with the distance d defined by

$$\forall (k,i,l), (k',i',l') \in X,\ d((k,i,l),(k',i',l')) = \max(d_K(k,k'), d_I(i,i'), d_K(l,l')).$$

- The payoff function $r : X \to [0,1]$ is defined by: for all $(k,i,k') \in X$, $r(k,i,k') := g(k,i)$.
- The correspondence $F : X \to \Delta(X)$ is defined by:

$$\forall (k,i,k') \in K \times I \times K,\ F(k,i,k') := \{\delta_{k',i'} \otimes q(k',i') : i' \in I\},$$

where $\delta_{k',i'}$ is the Dirac measure at (k',i'), and the symbol $\otimes$ stands for product measure.

Fix some arbitrary state $k_0 \in K$ and some arbitrary action $i_0 \in I$. Given an initial state k_1 in the MDP Γ, the corresponding initial state x_0 in the gambling house $\widetilde{\Gamma}$ is (k_0, i_0, k_1). By construction, the payoff at stage m in $\widetilde{\Gamma}(x_0)$ corresponds to the payoff at stage m in $\Gamma(k_1)$.

Now let us check the assumptions of Theorem 1. The state space X is compact metric. Because g is continuous, r is continuous, and the following lemma holds:

Lemma 6. *The correspondence F has a closed graph.*

Proof. Let $(x_n, u_n)_{n\in\mathbb{N}} \in (Graph\ F)^{\mathbb{N}}$ be a convergent sequence. By definition of F, for every $n \geqslant 1$, there exist $(k_n, i_n, k'_n) \in K \times I \times K$ and $i'_n \in I$ such that

$$x_n = (k_n, i_n, k'_n),$$

and

$$u_n = \delta_{k'_n, i'_n} \otimes q(k'_n, i'_n).$$

Moreover, the sequence $(k_n, i_n, k'_n, i'_n)_{n\geqslant 1}$ converges to some $(k,i,k',i') \in K\times I\times K\times I$. Because the transition q is jointly continuous, we obtain that (u_n) converges to $\delta_{(k',i')} \otimes q(k',i')$, which is indeed in $F(k,i,k')$. □

We now prove that for all $n \in \mathbb{N}^*$, v_n is 1-Lipschitz, and that w_∞ is 1-Lipschitz. It is more convenient to prove this result in the MDP Γ, rather than in the gambling house $\widetilde{\Gamma}$. Thus, in the next proposition, $H_\infty = (K \times I)^\infty$ is the infinite history in Γ, a strategy σ is a map from $\cup_{m \geqslant 1} K \times (I \times K)^{m-1}$ to $\Delta(I)$, and $\mathbb{P}_\sigma^{k_1}$ denotes the probability over H_∞ generated by the pair (k_1, σ). This proposition is similar to Proposition 6.

Proposition 7. *Let $k_1, k_1' \in K$ and σ be a strategy. Then there exist a probability measure $\mathbb{P}_\sigma^{k_1,k_1'}$ on $H_\infty \times H_\infty$, and a strategy τ such that:*

- *$\mathbb{P}_\sigma^{k_1,k_1'}$ has first marginal $\mathbb{P}_\sigma^{k_1}$,*
- *$\mathbb{P}_\sigma^{k_1,k_1'}$ has second marginal $\mathbb{P}_\tau^{k_1'}$,*
- *The following inequalities hold: for every $n \geqslant 1$,*

$$\mathbb{E}_\sigma^{k_1,k_1'}\left(\frac{1}{n}\sum_{m=1}^{n}|g(K_m,I_m)-g(K_m',I_m')|\right) \leqslant d_K(k_1,k_1'),$$

and

$$\mathbb{E}_\sigma^{k_1,k_1'}\left(\limsup_{n\to+\infty}\frac{1}{n}\sum_{m=1}^{n}|g(K_m,I_m)-g(K_m',I_m')|\right) \leqslant d_K(k_1,k_1'),$$

where K_m, I_m (resp. K_m', I_m') is the m-th coordinate of the first (resp. second) infinite history.

- *Under $\mathbb{P}_\sigma^{k_1,k_1'}$, for all $m \geqslant 1$, $I_m = I_m'$.*

Proof. Exactly as in Lemma 5, one can construct a measurable mapping $\psi : K \times K \times I \to \Delta(K \times K)$ such that for all $(k, k', i) \in K \times K \times I$, $\psi(k, k', i) \in \Delta(K \times K)$ is an optimal coupling between $q(k, i)$ and $q(k', i)$ for the KR distance.

We define a stochastic process on $I \times K \times I \times K$, in the following way: given an arbitrary action i_0, we set $I_0 = I_0' = i_0$, $K_1 = k_1$, $K_1' = k_1'$. Then, for all $m \geqslant 2$, given $(I_{m-1}, K_m, I_{m-1}', K_m')$, we construct $(I_m, K_{m+1}, I_m', K_{m+1}')$ as follows:

- I_m is drawn from $\sigma(K_1, I_1, ..., K_m)$,
- (K_{m+1}, K_{m+1}') is drawn from $\psi(K_m, K_m', I_m)$,
- we set $I_m' := I_m$.

By construction, $\mathbb{P}_\sigma^{k_1,k_1'}$ has first marginal $\mathbb{P}_\sigma^{k_1}$. For $m \geqslant 1$ and $h_m = (k_1', i_1', ..., k_m') \in H_m$, define $\tau(h_m) \in \Delta(I)$ as being the law of I_m', conditional to $K_1' = k_1', I_1' = i_1', ..., K_m' = k_m'$. This defines a strategy. Moreover, for all $m \geqslant 1$, we have

$$\mathbb{E}_\sigma^{k_1,k_1'}(d_K(K_{m+1}, K_{m+1}')|K_m, K_m') \leqslant d_K(K_m, K_m').$$

The process $(d_K(K_m, K_m'))_{m \geqslant 1}$ is a positive supermartingale, thus it converges almost surely. We conclude exactly as in the proof of Proposition 6. □

The previous proposition implies that the value functions v_n and w_∞ are 1-Lipschitz. Therefore, the family $\{v_n, n \geqslant 1\}$ is equicontinuous, and w_∞ is continuous. By Theorem 1, the gambling house $\widetilde{\Gamma}$ has a pathwise uniform value in pure strategies. It follows that the MDP Γ has a pathwise uniform value in pure strategies, and Theorem 3 holds.

Remark 4. *Renault and Venel [16] define slightly differently the auxiliary gambling house associated to a MDP. Instead of taking $K \times I \times K$ as the auxiliary state space, they take $[0, 1] \times K$, where the first component represents the stage payoff. In our framework, applying this method would lead to a measurability problem, when trying to transform a strategy in the auxiliary gambling house into a strategy in the MDP.*

4.3 Proof of Theorem 4 for POMDPs

In this subsection, we consider a POMDP $\Gamma = (K, I, S, g, q)$, as described in Subsection 2.3: the state space K and the signal space S are finite, the action set (I, d_I) is compact metric, and the transition function q and the payoff function g are continuous.

A standard way to analyze Γ is to consider the belief $p_m \in \Delta(K)$ at stage m about the state as a new state variable, and thus consider an auxiliary problem in which the state is perfectly observed and lies in $\Delta(K)$ (see [17], [19], [20]). The function g is linearly extended to $\Delta(K) \times \Delta(I)$, in the following way: for all $(p, u) \in \Delta(K) \times \Delta(I)$,

$$g(p, u) := \sum_{k \in K} \int_I g(k, i) u(di).$$

Let $\widetilde{q} : \Delta(K) \times I \to \Delta(\Delta(K))$ be the transition on the beliefs about the state, induced by q: if at some stage of the game, the belief of the decision-maker is p, and he plays the action i, then his belief about the next state will be distributed according to $\widetilde{q}(p, i)$. We extend linearly the transition $\widetilde{q}$ on $\Delta(K) \times \Delta(I)$, in the following way: for all $f \in \mathcal{C}(\Delta(K), [0, 1])$,

$$\int_{\Delta(K)} f(p)\ [\widetilde{q}(p, u)](dp) = \int_I \int_{\Delta(K)} f(p)\ [\widetilde{q}(p, i)](dp) u(di).$$

We can also define an auxiliary gambling house $\widetilde{\Gamma}$, with state space $[0, 1] \times I \times \Delta(K)$: at stage m, the auxiliary state x_m corresponds to the triple $(g(p_m, i_m), i_m, p_{m+1})$. Formally, the gambling house $\widetilde{\Gamma}$ is defined as follows:

- State space $X := [0, 1] \times I \times \Delta(K)$: the set $\Delta(K)$ is equipped with the norm 1 $\|.\|_K$, and the distance d on X is $d := \max(|.|, d_I, \|.\|_K)$.
- Payoff function $r : X \to [0, 1]$ such that for all $x = (a, i, p) \in X$, $r(x) := a$.
- Correspondence $F : X \to \Delta(X)$ defined for all $x = (a, i, p) \in X$ by $F(x) := \{g(p, i') \otimes \delta_{i'} \otimes \widetilde{q}(p, i') : i' \in I\}$.

Fix some arbitrary $a_0 \in [0, 1]$ and $i_0 \in I$. To each initial belief $p_1 \in \Delta(K)$ in Γ, we associate an initial state $x_0(p)$ in $\widetilde{\Gamma}$ by:

$$x_0(p_1) := (a_0, i_0, p_1).$$

By construction, the payoff at stage m in the auxiliary gambling house $\widetilde{\Gamma}(x_0(p_1))$ corresponds to the payoff $g(p_m, i_m)$ in the POMDP $\Gamma(p_1)$. In particular, for all $n \in \mathbb{N}^*$, the value of the n-stage gambling house $\widetilde{\Gamma}(x(p_1))$ coincides with the value of the n-stage POMDP $\Gamma(p_1)$, which is denoted by $v_n(p_1)$.

One could check that $\widetilde{\Gamma}$ satisfies the assumptions of Theorem 1 and therefore has a pathwise uniform value. This would especially imply that $\widetilde{\Gamma}$ has a uniform value in pure strategies, and it would prove that Γ has a uniform value in pure strategies. Indeed, let $p_1 \in \Delta(K)$ and $\widetilde{\sigma}$ be a strategy in $\widetilde{\Gamma}(x_0(p_1))$. Let σ be the associated strategy in the POMDP $\Gamma(p_1)$. For all $n \geqslant 1$, we have

$$\begin{aligned} \mathbb{E}_{\widetilde{\sigma}}^{x_0}\left(\frac{1}{n}\sum_{m=1}^{n} r(x_m)\right) &= \mathbb{E}_{\sigma}^{p_1}\left(\frac{1}{n}\sum_{m=1}^{n} g(p_m, i_m)\right) \\ &= \mathbb{E}_{\sigma}^{p_1}\left(\frac{1}{n}\sum_{m=1}^{n} g(k_m, i_m)\right). \end{aligned}$$

Consequently, the fact that $\widetilde{\Gamma}(x_0(p_1))$ has a uniform value in pure strategies implies that $\Gamma(p_1)$ also has a uniform value in pure strategies.

Unfortunately, this approach does not prove Theorem 4, i.e. the existence of the pathwise uniform value in Γ, due to the following problem:

Problem It may happen that

$$\mathbb{E}_{\widetilde{\sigma}}^{x_0(p_1)}\left(\liminf_{n\to+\infty}\frac{1}{n}\sum_{m=1}^{n}r(x_m)\right) > \mathbb{E}_{\sigma}^{p_1}\left(\liminf_{n\to+\infty}\frac{1}{n}\sum_{m=1}^{n}g(k_m,i_m)\right).$$

Indeed, $r(x_m)$ is not equal to $g(k_m, i_m)$: it is the expectation of $g(k_m, i_m)$ with respect to p_m. Consequently, the fact that $\widetilde{\sigma}$ is a pathwise ϵ-optimal strategy in $\widetilde{\Gamma}(x_0(p_1))$ does not imply that σ is a pathwise ϵ-optimal strategy in $\Gamma(p_1)$.

To prove Theorem 4, we adapt the proof of Theorem 1 to the framework of POMDPs. Recall that the proof of Theorem 1 was decomposed into three lemmas (Lemmas 1, 2 and 3) and a conclusion (Subsection 3.5). We adapt the three lemmas, and the conclusion is similar.

In order to obtain the first lemma, we check that F has a closed graph.

Proposition 8. *The correspondence F has a closed graph.*

Proof. Let $(x_n, u_n)_{n\in\mathbb{N}} \in (Graph\ F)^{\mathbb{N}}$ be a sequence that converges to $(x, u) \in X \times \Delta(X)$. By definition of F, for every $n \geqslant 1$ there exists $(a_n, i_n, p_n, i'_n) \in ([0,1] \times I \times \Delta(K) \times I)$ such that

$$x_n = (a_n, i_n, p_n),$$

and

$$u_n = g(p_n, i'_n) \otimes \delta_{i'_n} \otimes \widetilde{q}(p_n, i'_n).$$

It follows that the sequence $(a_n, i_n, p_n, i'_n)_{n\geqslant 1}$ converges to some $(a, i, p, i') \in [0,1]\times I\times\Delta(K)\times I$ and $x = (a, i, p)$.

By Feinberg [12, Theorem 3.2], the function $\widetilde{q}$ is jointly continuous. Because the payoff function g is also continuous, we obtain that u_n converges to $u = q(p, i') \otimes \delta_{i'} \otimes \widetilde{q}(p', i')$ which is indeed in $F(x)$. □

Now we can apply Lemma 1 to the gambling house $\widetilde{\Gamma}$. For $p \in \Delta(K)$, define $v(p) := \limsup_{n\to+\infty} v_n(p)$. Note that for all $x = (a, i, p) \in X$, the set $F(x)$ depends only on the third component p. Thus, Lemma 1 implies the following lemma for the POMDP Γ:

Lemma 7. *Let $p_1 \in \Delta(K)$. There exists a distribution $\mu^* \in \Delta(\Delta(K))$ and a stationary strategy $\sigma^* : \Delta(K) \to \Delta(I)$ such that*

- *μ^* is σ^*-invariant: for all $f \in \mathcal{C}(\Delta(K), [0,1])$,*

$$\int_{\Delta(K)} \hat{f}(\widetilde{q}(p, \sigma^*(p))\mu^*(dp) = \hat{f}(\mu^*)$$

- *For every $\varepsilon > 0$ and $N \geqslant 1$, there exists a (pure) strategy σ in Γ and $n \geqslant N$ such that σ is 0-optimal in $\Gamma_n(p_1)$ and*

$$d_{KR}\left(\frac{1}{n}\sum_{m=1}^{n} z_m(p_1, \sigma), \mu^*\right) \leqslant \varepsilon,$$

where $z_m(p_1, \sigma)$ is the distribution over $\Delta(K)$ at stage m, starting from p_1,

- $\int_{\Delta(K)} g(p,\sigma^*(p))\mu^*(dp) = \hat{v}(\mu^*) = v(p_1)$.

We can now state a new lemma about pathwise convergence in Γ. This replaces Lemma 2.

Lemma 8. *Let $p_1 \in \Delta(K)$ and μ^* be the corresponding measure in the previous lemma. There exists a measurable set $B \subset \Delta(K)$ such that $\mu^*(B) = 1$ and for all $p \in B$,*

$$\mathbb{E}^p_{\sigma^*}\left(\liminf_{n\to+\infty} \frac{1}{n}\sum_{m=1}^{n} g(k_m,i_m)\right) = v(p) \quad \mathbb{P}^p_{\sigma^*} - a.s.$$

Proof. It is not enough to apply Birkhoff's theorem to the Markov chain $(p_m)_{m\geqslant 1}$, due to the problem mentioned previously. Instead, we consider the random process $(y_m)_{m\geqslant 1}$ on $Y := K \times I \times \Delta(K)$, defined for all $m \geqslant 1$ by $y_m := (k_m, i_m, p_m)$: (current state, action played, belief about the current state). Under $\mathbb{P}^{\mu^*}_{\sigma^*}$, this is a Markov chain. Indeed, given $m \geqslant 1$ and $(y_1, y_2, ..., y_m) \in Y^m$, the next state y_{m+1} is generated in the following way:

- a pair (k_{m+1}, s_m) is drawn from $q(k_m, i_m)$,
- the decision-maker computes the new belief p_{m+1} according to p_m and s_m,
- the decision-maker draws an action i_{m+1} from $\sigma^*(p_{m+1})$.

By construction, the law of y_{m+1} depends only on y_m, and $(y_m)_{m\geqslant 1}$ is a Markov chain. Define $\nu^* \in \Delta(Y)$ such that the third marginal of ν^* is μ^*, and for all $p \in \Delta(K)$, the conditional law $\nu^*(.|p) \in \Delta(K \times I)$ is $p \otimes \sigma(p)$. Under $\mathbb{P}^{\mu^*}_{\sigma^*}$, for all $m \geqslant 1$, the third marginal of y_m is distributed according to μ^*. Moreover, conditional on p_m, the random variables k_m and i_m are independent, the conditional distribution of k_m knowing p_m is p_m, and the conditional distribution of i_m knowing p_m is $\sigma^*(p_m)$. Thus, ν^* is an invariant measure for the Markov chain $(y_m)_{m\geqslant 1}$. Define a measurable map $f : Y \to [0,1]$ by: for all $(k,i,p) \in Y$, $f(k,i,p) = g(k,i)$. Now we can apply Theorem 5 to $(y_m)_{m\geqslant 1}$, and deduce that there exist $B_0 \subset \Delta(K)$ and $w : Y \to [0,1]$ such that for all $p \in B_0$,

$$\frac{1}{n}\sum_{m=1}^{n} f(y_m) \underset{n\to+\infty}{\to} w(k_1,i_1,p) \quad \mathbb{P}^p_{\sigma^*} - \text{almost surely}, \tag{5}$$

and

$$\hat{w}(\nu^*) = \hat{f}(\nu^*).$$

By definition of f, for all $m \geqslant 1$, $f(y_m) = g(k_m, i_m)$. Moreover, by definition of ν^*, we have

$$\hat{f}(\nu^*) = \int_{\Delta(K)} g(p,\sigma^*(p))\mu^*(dp),$$

and by Lemma 7, we deduce that $\hat{f}(\nu^*) = \hat{v}(\mu^*)$. Consequently, $\hat{w}(\nu^*) = \hat{v}(\mu^*)$. Given $p \in B_0$, denote by $w_0(p)$ the expectation of $w(.,p)$ with respect to $\mathbb{P}^p_{\sigma^*}$. By Equation (5), we have

$$\mathbb{E}^p_{\sigma^*}\left(\lim_{n\to+\infty} \frac{1}{n}\sum_{m=1}^{n} g(k_m,i_m)\right) = w_0(p).$$

Let us prove that $w_0 = v \quad \mathbb{P}^{\mu^*}_{\sigma^*}$-almost surely. Note that $\hat{w_0}(\mu^*) = \hat{w}(\nu^*) = \hat{v}(\mu^*)$. Consequently, it is enough to show that $w_0 \leqslant v \quad \mathbb{P}^{\mu^*}_{\sigma^*}$-almost surely. By the dominated convergence theorem and the definition of v, we have

$$\begin{aligned}\mathbb{E}^p_{\sigma^*}\left(\lim_{n\to+\infty} \frac{1}{n}\sum_{m=1}^{n} g(k_m,i_m)\right) &= \lim_{n\to+\infty} \mathbb{E}^p_{\sigma^*}\left(\frac{1}{n}\sum_{m=1}^{n} g(k_m,i_m)\right) \\ &\leqslant v(p),\end{aligned}$$

and the lemma is proved. ∎

For every $n \geqslant 1$, the value function v_n is 1-Lipschitz, as a consequence of the following proposition.

Proposition 9. *Let $p, p' \in \Delta(K)$ and σ be a strategy in Γ. Then, for every $n \geqslant 1$,*

$$\left| \mathbb{E}_\sigma^p \left(\frac{1}{n} \sum_{m=1}^{n} g(k_m, i_m) \right) - \mathbb{E}_\sigma^{p'} \left(\frac{1}{n} \sum_{m=1}^{n} g(k_m, i_m) \right) \right| \leqslant \|p - p'\|_1.$$

This proposition is proved in Rosenberg, Solan and Vieille [18, Proposition 1]. In their framework, I is finite, but the fact that I is compact does not change the proof at all.

Last, we establish the junction lemma, which replaces Lemma 3.

Lemma 9. *Let $p, p' \in \Delta(K)$ and σ be a strategy such that*

$$\mathbb{E}_\sigma^p \left(\lim_{n \to +\infty} \frac{1}{n} \sum_{m=1}^{n} g(k_m, i_m) \right) = v(p).$$

Then, the following inequality holds:

$$\mathbb{E}_\sigma^{p'} \left(\liminf_{n \to +\infty} \frac{1}{n} \sum_{m=1}^{n} g(k_m, i_m) \right) \geqslant v(p') - 2 \|p - p'\|_1 .$$

Proof. Let $k \in K$ and $p_1 \in \Delta(K)$. Denote by $\mathbb{P}_\sigma^{p_1}(h_\infty | k)$ the law of the infinite history $h_\infty \in (K \times I \times S)^{\mathbb{N}^*}$ in the POMDP $\Gamma(p_1)$, under the strategy σ, and conditional to $k_1 = k$. Then $\mathbb{P}_\sigma^{p}(h_\infty | k) = \mathbb{P}_\sigma^{p'}(h_\infty | k)$ and

$$\mathbb{E}_\sigma^{p'} \left(\liminf_{n \to +\infty} \frac{1}{n} \sum_{m=1}^{n} g(k_m, i_m) \right) \geqslant v(p) - \|p - p'\|_1 .$$

For every $n \geqslant 1$, v_n is 1-Lipschitz, thus the function v is also 1-Lipschitz, and the lemma is proved. □

The conclusion of the proof is similar to Section 3.5. Note that apart from the three main lemmas, the only additional property used in Section 3.5 was that the family $(v_n)_{n \geqslant 1}$ is uniformly equicontinuous. For every $n \geqslant 1$, v_n is 1-Lipschitz, thus the family $(v_n)_{n \geqslant 1}$ is indeed uniformly equicontinuous.

Acknowledgments

Both authors gratefully acknowledge the support of the Agence Nationale de la Recherche, under grant ANR JEUDY, ANR-10-BLAN 0112, and thank Eugene A. Feinberg, Fabien Gensbittel, Jérôme Renault and Eilon Solan for fruitful discussions.

References

[1] C Aliprantis and K. Border. Infinite dimensional analysis. 2006.

[2] E. Altman. Denumerable constrained markov decision processes and finite approximations. *Mathematics of operations research*, 19(1):169–191, 1994.

[3] A. Arapostathis, V. Borkar, E. Fernández-Gaucherand, M. Ghosh, and S. Marcus. Discrete-time controlled markov processes with average cost criterion: a survey. *SIAM Journal on Control and Optimization*, 31(2):282–344, 1993.

[4] R. Bellman. A markovian decision process. Technical report, DTIC Document, 1957.

[5] D. Bertsekas and S. Shreve. *Stochastic optimal control: The discrete time case*. Athena Scientific, 1996.

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

[7] V.S. Borkar. A convex analytic approach to markov decision processes. *Probability Theory and Related Fields*, 78(4):583–602, 1988.

[8] V.S. Borkar. Average cost dynamic programming equations for controlled markov chains with partial observations. *SIAM Journal on Control and Optimization*, 39:673, 2000.

[9] C. Dellacherie and P-A Meyer. *Probabilities and Potential, C: Potential Theory for Discrete and Continuous Semigroups*. Elsevier, 2011.

[10] L.E. Dubins and L.J. Savage. *How to gamble if you must: Inequalities for stochastic processes*. McGraw-Hill New York, 1965.

[11] E. Feinberg. On measurability and representation of strategic measures in markov decision processes. *Lecture Notes-Monograph Series*, pages 29–43, 1996.

[12] E. Feinberg, P. Kasyanov, and M. Zgurovsky. Partially observable total-cost markov decision processes with weakly continuous transition probabilities. *arXiv preprint arXiv:1401.2168*, 2014.

[13] O. Hernández-Lerma and J. Lasserre. *Markov chains and invariant probabilities*, volume 211. Springer Science & Business Media, 2003.

[14] A. Maitra and W. Sudderth. *Discrete gambling and stochastic games*, volume 32. Springer Verlag, 1996.

[15] J. Renault. Uniform value in dynamic programming. *Journal of the European Mathematical Society*, 13(2):309–330, 2011.

[16] J. Renault and X. Venel. A distance for probability spaces, and long-term values in markov decision processes and repeated games. *Arxiv preprint arXiv:1202.6259*, 2012.

[17] D. Rhenius. Incomplete information in markovian decision models. *The Annals of Statistics*, pages 1327–1334, 1974.

[18] D. Rosenberg, E. Solan, and N. Vieille. Blackwell optimality in markov decision processes with partial observation. *Annals of Statistics*, 30:1178–1193, 2002.

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

[20] A. Yushkevich. Reduction of a controlled markov model with incomplete data to a problem with complete information in the case of borel state and control space. *Theory of Probability and Its Applications*, 21(1):153–158, 1976.