Constant payoff in zero-sum stochastic games

Olivier Catoni, Miquel Oliu-Barton and Bruno Ziliotto

December 1, 2020

Abstract

In a zero-sum stochastic game, at each stage, two adversary players take decisions and receive a stage payoff determined by them and by a controlled random variable representing the state of nature. The total payoff is the normalized discounted sum of the stage payoffs. In this paper we solve the "constant payoff" conjecture formulated by Sorin, Vigeral and Venel (2010): if both players use optimal strategies, then for any $\alpha > 0$, the expected average payoff between stage 1 and stage α/λ tends to the limit discounted value of the game, as the discount rate goes to 0.

1 Introduction

Stochastic games were introduced by Shapley [13] in order to model a repeated interaction between two opponent players in a changing environment. At each stage $m \in \mathbb{N}$ of the game, players play a zero-sum game that depends on a state variable. Formally, knowing the current state k_m, Player 1 chooses an action i_m and Player 2 chooses an action j_m. Their choices occur independently and simultaneously and have two consequences: first, they produce a stage payoff $g_m := g(k_m, i_m, j_m)$ and second, they determine the law $q(k_m, i_m, j_m)$ of the next period's state k_{m+1}. Thus, the sequence of states follows a Markov chain controlled by the actions of both players. To any discount rate $\lambda \in (0, 1]$ and any initial state k corresponds a λ-discounted game, denoted by $\Gamma_\lambda(k)$, in which Player 1 maximizes the expectation of $\sum_{m\geq 1} \lambda(1-\lambda)^{m-1} g_m$ given that $k_1 = k$, while Player 2 minimizes this same amount. A crucial aspect in this model is that *the current state is commonly observed* by the players at every stage. Another one is stationarity: the transition function and stage payoff function do not change over time. We assume like in Shapley's seminal work, that the set of states and the action sets for both players are *finite*.

Shapley [13] proved that for any initial state k and any discount rate λ, the λ-discounted stochastic game has a value $v_\lambda(k)$, which is the unique fixed point of a contracting map. Furthermore, both players have optimal strategies that depend on the past only through the current state. A wide area of research is to investigate the properties of the discounted game, when λ tends to 0. Intuitively, this corresponds to a game played either between very patient players, or between players who are very likely to interact a great number of times. Building on Shapley's results, Bewley and Kohlberg [2] proved that $v_\lambda(k)$ converges as λ tends to 0. An alternative proof of this result was recently obtained by Oliu-Barton [9], using probabilistic and linear programming techniques. Mertens and Neyman [8] proved then the existence of the so-called *uniform value* $v^*(k)$, and its equality with $\lim_{\lambda\to 0} v_\lambda(k)$. A new characterization for $v_\lambda(k)$, and a formula for $v^*(k)$ were recently obtained by Attia and Oliu-Barton [1]. Efficient algorithms to compute these values were then deduced by Oliu-Barton [11]. The finiteness of the state space plays a crucial role in these results, as highlighted by the counterexamples of Vigeral [18] and Ziliotto [19] who considered, respectively, the case of compact action sets and compact state space.

A remarkable property, referred to as the *constant payoff property* was proved by Sorin, Venel and Vigeral [16] in the framework of single decision-maker problems: for any sufficiently small λ there exists an optimal strategy so that the expectation of the cumulated payoff $\sum_{m=1}^{M} \lambda(1-\lambda)^{m-1} g_m$ given $k_1 = k$ is approximately equal to $(\sum_{m=1}^{M} \lambda(1-\lambda)^{m-1}) v^*(k)$. Note that the positive weights $\lambda(1-\lambda)^{m-1}$ add up to 1, so that $\sum_{m=1}^{M} \lambda(1-\lambda)^{m-1}$ represents the fraction of the game that has already been played at stage M. The constant payoff property holds as soon as the discounted value converges as the discount rate goes to 0, and that the convergence is uniform in the state space. Further, it was conjectured that under similar conditions the constant payoff property should hold for any class of two-player zero-sum stochastic games that satisfy the same assumptions. As the discounted value of finite stochastic games converges, and the convergence is uniform (by finiteness), the conjecture directly applies to this class of games.

The constant payoff property is not straightforward. Lehrer and Sorin [7] provided a simple example of a Markov decision problem over a countable set of states where this property fails: when the decision-maker plays an optimal strategy, he gets a high payoff during the first half of the game, and a low payoff during the second half. However, it was known to hold for finite *absorbing games*, a subset of stochastic games in which all states except one are absorbing. Beyond the finite framework, Sorin and Vigeral [17] established the constant payoff property for absorbing games with compact action sets and jointly continuous payoff and transition functions. Oliu-Barton [10] established the same property for the splitting game, a stochastic game with compact action sets and jointly continuous payoff and transition functions introduced by Sorin [14] to capture the information transmission in repeated games with incomplete information. Let us note that, in spite of their differences, absorbing games and games with incomplete information have in common that the dynamics of the game has an irreversible property, which is not present in stochastic games.

The main contribution of this paper is to establish that finite stochastic games have the constant payoff property, and thus to solve the conjecture in Sorin, Venel and Vigeral [16]. Moreover, a property more general than the conjecture is established (*strong constant payoff property*): for any sufficiently small λ, *for any pair of optimal strategies*, the expectation of the cumulated payoff $\sum_{m=1}^{M}\lambda(1-\lambda)^{m-1}g_m$ given $k_1=k$ is approximately equal to $(\sum_{m=1}^{M}\lambda(1-\lambda)^{m-1})v^*(k)$.
The proof relies heavily on the semi-algebraic approach developed by Bewley and Kohlberg [2], namely that the value function $v_\lambda(k)$ and a family of optimal stationary strategies $(x_\lambda^1, x_\lambda^2)$ admit a Puiseux series expansion in a neighborhood of 0 (*optimal Puiseux strategy profiles*). It is decomposed in two parts. First, we establish that the constant payoff property holds for optimal Puiseux strategy profiles (*weak constant payoff property*). This readily proves the constant payoff conjecture for finite stochastic games. In a second part, we generalize this property to any family of optimal strategies, to obtain the strong constant payoff property.

The remainder of the paper is divided as follows. Section 2 presents the model and main result. Section 3 is devoted to the proof of the weak constant payoff property. Section 4 proves the strong constant payoff property. Section 5 gives some examples and remarks.

2 Model and main results

2.1 Stochastic games

We consider throughout this paper a standard two-player zero-sum stochastic game, as introduced by Shapley [13]. Such games are described by a 5-tuple $\Gamma=(K,I,J,g,q)$, where K is the set of states, I and J are the action sets of Player 1 and 2 respectively, $g: K\times I\times J\to\mathbb{R}$ is the payoff function and $q: K\times I\times J\to\Delta(K)$ is the transition function, where for each finite or countable set X, we denote by $\Delta(X)$ the set of probability distributions over X. **We assume that K, I and J are finite sets**.

Outline of the game. The game Γ proceeds as follows: at every stage $m\geq 1$, knowing the current state k_m, the players choose actions i_m and j_m independently; Player 1 receives the stage payoff $g(k_m,i_m,j_m)$, and Player 2 receives $-g(k_m,i_m,j_m)$. A new state k_{m+1} is drawn according to the probability $q(\cdot\,|\,k_m,i_m,j_m)$. The players observe the pair of actions (i_m,j_m) and the new state k_{m+1}, and the game proceeds to stage $m+1$.

Discounted stochastic games. For any discount rate $\lambda\in(0,1]$, we denote by $\Gamma_\lambda(k)$ the stochastic game Γ with initial state k and where Player 1 maximizes, in expectation, the normalized λ-discounted sum of stage payoffs

$$\sum\nolimits_{m\geq 1}\lambda(1-\lambda)^{m-1}g(k_m,i_m,j_m),$$

while Player 2 minimizes this amount. More precisely, we consider the case where the strategies of the two players form the saddle point of a min-max problem, as explained below.

Strategies. The sequence $(k_1,i_1,j_1,...,k_m,i_m,j_m,...)$ generated along the game is called *a play*. The set of plays is $(K\times I\times J)^{\mathbb{N}}$.

(*i*) A *strategy* for a player specifies a mixed action to each possible set of past observations: formally, a strategy for Player 1 is a collection of maps $\sigma^1=(\sigma^1)_{m\geq 1}$, where $\sigma^1_m: (K\times I\times$

$J)^{m-1} \times K \to \Delta(I)$. Similarly, a strategy for Player 2 is a collection of maps $\sigma^2 = (\sigma^2)_{m\geq 1}$, where $\sigma^2_m : (K \times I \times J)^{m-1} \times K \to \Delta(J)$.

(*ii*) A *stationary strategy* plays according to the current state only. Formally, a stationary strategy for Player 1 is a mapping $x^1 : K \to \Delta(I)$. Similarly, a stationary strategy for Player 2 is a mapping $x^2 : K \to \Delta(J)$.

(*iii*) A *strategy profile* is a pair of strategies (σ^1, σ^2).

Notation. The sets of strategies for Player 1 and 2 are denoted by Σ^1 and Σ^2, respectively, and the sets of stationary strategies by $\Delta(I)^K$ and $\Delta(J)^K$.

We denote by $\mathsf{P}^k_{\sigma^1,\sigma^2}$ the unique probability measure on the set of plays $(K \times I \times J)^{\mathbb{N}}$ such that, for any finite play $h^n = (k_1, i_1, j_1, \dots, k_{n-1}, i_{n-1}, j_{n-1}, k_n)$ one has

$$\mathsf{P}^k_{\sigma^1,\sigma^2}(h_n) = \prod_{m=1}^{n-1} \sigma^1_m[h^n_m](i_m)\sigma^2_m[h^n_m](j_m)q(k_{m+1}|k_m, i_m, j_m)\,.$$

where h^n_m is the restriction of h^n to the first m stages, i.e. $h^n_1 := k_1$ and for all $2 \leq m \leq n$:

$$h^n_m := (k_1, i_1, j_1, \dots, k_{m-1}, i_{m-1}, j_{m-1}, k_m)\,.$$

The extension to infinite plays follows from the Kolmogorov extension theorem. Thus, $\mathsf{P}^k_{\sigma^1,\sigma^2}$ is the unique probability measure on plays induced by the pair (σ^1, σ^2) in the stochastic game starting from state k (note that the dependence on the transition function q is omitted). The expectation with respect to the probability $\mathsf{P}^k_{\sigma^1,\sigma^2}$ is denoted by $\mathbb{E}^k_{\sigma^1,\sigma^2}$. For any $\lambda \in (0, 1]$ and any $k \in K$, we denote by $\gamma_\lambda(k, \cdot, \cdot) : \Sigma^1 \times \Sigma^2 \to \mathbb{R}$ the payoff function corresponding to the game $\Gamma_\lambda(k)$:

$$\gamma_\lambda(k, \sigma^1, \sigma^2) := \mathbb{E}^k_{\sigma^1,\sigma^2}\left[\sum\nolimits_{m\geq 1} \lambda(1-\lambda)^{m-1} g(k_m, i_m, j_m)\right]. \tag{1}$$

Shapley's results. For any discount rate $\lambda \in (0, 1]$ and any initial state $k \in K$, Shapley [13] proved that the discounted stochastic game $\Gamma_\lambda(k)$ has a value, so that the following equalities hold:

$$v_\lambda(k) = \max_{\sigma^1\in\Sigma^1} \min_{\sigma^2\in\Sigma^2} \gamma_\lambda(k, \sigma^1, \sigma^2) = \min_{\sigma^2\in\Sigma^2} \max_{\sigma^1\in\Sigma^1} \gamma_\lambda(k, \sigma^1, \sigma^2)\,. \tag{2}$$

Furthermore, both players have optimal stationary strategies in Γ_λ, where a strategy $\sigma^1 \in \Sigma^1$ is optimal for Player 1 if for any $k \in K$, it realizes the maximum in the left-hand side of (2), and a strategy $\sigma^2 \in \Sigma^2$ is optimal for Player 2 if for any $k \in K$, it realizes the minimum in the right-hand side of (2).

Notation. The set of optimal strategies for Player 1 and 2 in the game Γ_λ are denoted by Σ^1_λ and Σ^2_λ, respectively.

2.1.1 Puiseux strategies

A map $f : (a, b) \to \mathbb{R}$ is a *Puiseux series* on $(a_0, b_0) \subset (a, b)$ if there exists $m_0 \in \mathbb{Z}$, $N \in \mathbb{N}$ and a real sequence $(c_m)_{m\geq 0}$ so that

$$f(\lambda) = \sum_{m\geq m_0} c_m \lambda^{m/N} \qquad \forall \lambda \in (a_0, b_0).$$

A function $f : (0, 1] \to \mathbb{R}$ admits a Puiseux expansion at 0 if there exists λ_0 so that f is a Puiseux series on $(0, \lambda_0)$. Clearly, if f is bounded then one can take $m_0 = 0$.

Definition 2.1. A *Puiseux strategy profile* is a family of stationary strategy profiles $(x^1_\lambda, x^2_\lambda)_{\lambda\in(0,1]}$ so that, for some $\lambda_0 \in (0, 1]$, the mappings $\lambda \mapsto x^1_\lambda(k, i)$ and $\lambda \mapsto x^2_\lambda(k, j)$ are bounded real Puiseux series on $(0, \lambda_0)$, for all $(k, i, j) \in K \times I \times J$.

Definition 2.2. An *optimal Puiseux strategy profile* is a Puiseux strategy profile $(x^1_\lambda, x^2_\lambda)_{\lambda\in(0,1]}$ so that for all $\lambda \in (0, \lambda_0]$ and $k \in K$, the stationary strategies x^1_λ and x^2_λ are optimal in Γ_λ.

2.1.2 The semi-algebraic approach

Fix $\lambda \in (0,1]$ and let $v_\lambda \in \mathbb{R}^K$ be the vector of values. Building on Shapley's results [13], Bewley and Kohlberg [2] defined a subset $S \subset \mathbb{R} \times \mathbb{R}^K \times \mathbb{R}^{K\times I} \times \mathbb{R}^{K\times J}$ by setting

$$(\lambda, v, x^1, x^2) \in S \iff \begin{cases} \lambda \in \mathbb{R} \text{ is a discount rate} \\ v \in \mathbb{R}^K \text{ is the vector of values of } \Gamma_\lambda \\ (x^1, x^2) \in \mathbb{R}^{K\times I} \times \mathbb{R}^{K\times J} \text{ is a pair of optimal stationary strategies in } \Gamma_\lambda \,. \end{cases}$$

The set S is semi-algebraic, as it can be described by the following finite set of polynomial equalities and inequalities:

$$\begin{array}{rcl} 0 < \lambda & \leq & 1 \\ \forall (k,i),\ x^1(k,i) \geq 0, \text{ and } \forall k, \quad \sum_{i\in I} x^1(k,i) & = & 1 \\ \forall (k,j),\ x^2(k,j) \geq 0, \text{ and } \forall k, \quad \sum_{j\in J} x^2(k,j) & = & 1 \\ \forall (k,j), \quad \sum_{i\in I} x^1(k,i) \Big(\lambda g(k,i,j) + (1-\lambda) \sum_{\ell\in K} q(\ell|k,i,j) v(\ell)\Big) & \geq & v(k) \\ \forall (k,i), \quad \sum_{j\in J} x^2(k,j) \Big(\lambda g(k,i,j) + (1-\lambda) \sum_{\ell\in K} q(\ell|k,i,j) v(\ell)\Big) & \leq & v(k)\,. \end{array}$$

By the Tarski-Seidenberg elimination theorem, the functions $\lambda \mapsto v_\lambda(k)$ are real, semi-algebraic functions, for each initial state $k \in K$. Similarly, there exist a selection of optimal strategies $(x^1_\lambda, x^2_\lambda)$ such that the maps $\lambda \mapsto x^1_\lambda(k,i)$ and $\lambda \mapsto x^2_\lambda(k,j)$ are real semi-algebraic functions as well, for all (k,i,j). By the Puiseux theorem, any real semi-algebraic function $f : (0,1] \to \mathbb{R}$ admits a Puiseux expansion in some neighborhood of 0. Hence,

- For each $k \in K$, the map $\lambda \mapsto v_\lambda(k)$ admits a Puiseux expansion at 0, so that the limit $v^*(k) := \lim_{\lambda\to 0} v_\lambda(k)$ exists.
- There exists an optimal Puiseux strategy profile $(x^1_\lambda, x^2_\lambda)_{\lambda\in(0,1]}$.

2.1.3 The game on $[0,1]$

Let $\lambda \in (0,1]$. For any $M \in \mathbb{N}$, define the following map:

$$\eta(\lambda, \cdot\,) : \mathbb{N} \to [0,1], \qquad \eta(\lambda, M) := \sum_{m=1}^{M} \lambda(1-\lambda)^{m-1}\,.$$

It can be interpreted as a clock that indicates the *fraction of the game* that has already been played after any given number of stages. Conversely, to any fraction of the game $t \in [0,1]$ corresponds a stage where the sum of weights of the previous stages is approximately equal to t. Formally, we introduce the *inverse-clock* map by

$$\varphi(\lambda, \cdot\,) : [0,1] \to \mathbb{N} \cup \{+\infty\}\,, \qquad \varphi(\lambda, t) := \inf\{M \geq 1,\ \eta(\lambda, M) \geq t\} \ = \left\lceil \frac{\ln(1-t)}{\ln(1-\lambda)} \right\rceil\,,$$

where $\lceil x \rceil$ denotes the upper integer part of x. The notion of clock and inverse-clock, which are now standard, were initiated by Sorin [15], and allow to consider the discrete-time game $\Gamma_\lambda(k)$ as a game played on the time interval $[0,1]$.

Cumulated payoffs. For any fraction $t \in [0,1]$ we extend the definition of the payoff function to the map $\gamma_\lambda(k, \cdot, \cdot, \cdot) : \Sigma^1 \times \Sigma^2 \times [0,1] \to \mathbb{R}$ by setting

$$\gamma_\lambda(k, \sigma^1, \sigma^2; t) \quad := \quad \mathbb{E}^k_{\sigma^1,\sigma^2} \left[\sum_{m=1}^{\varphi(\lambda,t)} \lambda(1-\lambda)^{m-1} g(k_m, i_m, j_m) \right]. \tag{3}$$

For convenience, for any pair of strategies (σ^1, σ^2) we set $\gamma_\lambda(\sigma^1, \sigma^2; t) \in \mathbb{R}^K$ to be the vector of payoffs $\gamma_\lambda(k, \sigma^1, \sigma^2; t)$, $k \in K$. Note also that $\gamma_\lambda(\sigma^1, \sigma^2) = \gamma_\lambda(\sigma^1, \sigma^2; 1)$ by definition.

2.2 Main result

Our main result is a precise characterisation of the cumulated payoff at time t, when both players use optimal strategies in the game $\Gamma_\lambda(k)$, for sufficiently small $\lambda \in (0,1]$.

Theorem 2.3 (Strong constant payoff property)**.** *For any $\varepsilon > 0$, there exists $\lambda_0 \in (0,1)$ so that for all $\lambda \in (0,\lambda_0)$, $t \in [0,1]$, $k \in K$, and $(\sigma^1,\sigma^2) \in \Sigma^1_\lambda \times \Sigma^2_\lambda$ one has:*

$$\left|\gamma_\lambda(k,\sigma^1,\sigma^2;t) - tv^*(k)\right| \leq \varepsilon\,. \tag{4}$$

This result solves the conjecture raised by Sorin, Venel and Vigeral [16] in a *strong sense*. That is, where [16] conjectured the existence of a strategy profile (σ^1,σ^2) so that (4) holds, we prove that the constant payoff property holds for *every* optimal strategy profile.

3 Weak constant payoff property

Theorem 3.1 (Weak constant payoff property)**.** *For any optimal Puiseux strategy profile $(x^1_\lambda, x^2_\lambda)_{\lambda\in(0,1]}$,*

$$\lim_{\lambda\to 0} \gamma_\lambda(x^1_\lambda, x^2_\lambda; t) = tv^* \qquad \forall t \in [0,1]\,.$$

Remark. Theorem 3.1 establishes the constant payoff conjecture of Sorin, Venel and Vigeral [16] for a specific family of strategy profiles.

Remark. For $\varepsilon > 0$, let $(\sigma_\varepsilon,\tau_\varepsilon)$ be a pair of ε-optimal *uniform* strategies, that is, satisfying $\gamma_\lambda(k,\sigma_\varepsilon,\tau) \geq v^*(k) - \varepsilon$ and $\gamma_\lambda(k,\sigma,\tau_\varepsilon) \leq v^*(k) + \varepsilon$ for all $k \in K$, for all pair of strategies (σ,τ) and all λ small enough. Such a pair exists by Mertens and Neyman [8]. Then, for all $k \in K$, for any sequence of strategies $(\sigma_\lambda,\tau_\lambda)$ and any $t \in [0,1]$ one has $\liminf_{\lambda\to 0} \gamma_\lambda(k,\sigma_\varepsilon,\tau_\lambda;t) \geq tv^*(k) - \varepsilon$ and $\limsup_{\lambda\to 0} \gamma_\lambda(k,\sigma_\lambda,\tau_\varepsilon;t) \leq tv^*(k) + \varepsilon$. In particular, $(\sigma_\varepsilon,\tau_\varepsilon)$ satisfies the weak constant payoff property, up to an error term ε. Nonetheless, these strategies are in general not stationary.

Proof. In the sequel, $(x^1_\lambda, x^2_\lambda)$ denotes an optimal Puiseux strategy profile. Let $\lambda_0 > 0$ be such that all the coordinates of $\lambda \mapsto x^1_\lambda$ and $\lambda \mapsto x^2_\lambda$ are Puiseux series on $(0,\lambda_0)$. The result is clear for $t = 0$ and $t = 1$ so we fix in the sequel some $t \in (0,1)$.

Step 0: *Introduction of tools.*
For any stationary strategy profile (x^1,x^2), define the matrix $\Pi(\lambda,x^1,x^2) \in \mathbb{R}^{K\times K}$ for all $\lambda \in (0,1]$ and the vector $g(x^1,x^2) \in \mathbb{R}^K$ by setting

$$\begin{aligned}
\Pi^{k,\ell}(\lambda,x^1,x^2) &:= \mathbb{E}^k_{x^1,x^2}\left[\sum\nolimits_{m\geq 1} \lambda(1-\lambda)^{m-1}\mathbb{1}_{\{k_m=\ell\}}\right] \qquad \forall (k,\ell) \in K^2,\\
g^k(x^1,x^2) &:= \sum_{(i,j)\in I\times J} x^1(k,i)x^2(k,j)g(k,i,j) \qquad \forall k \in K\,.
\end{aligned}$$

The real $\Pi^{k,\ell}(\lambda,x^1,x^2)$ represents the expected (discounted) fraction of the game spent in state ℓ, given that players play stationary strategies x^1 and x^2, and the initial state is k. The real $g^k(x^1,x^2)$ represents the expected stage payoff, given that players play x^1 and x^2 and the initial state is k. We claim that $\lambda \mapsto \Pi(\lambda,x^1_\lambda,x^2_\lambda)$ is a bounded real Puiseux series, so that the limit $\Pi := \lim_{\lambda\to 0}\Pi(\lambda,x^1_\lambda,x^2_\lambda) \in \mathbb{R}^{K\times K}$ exists. Indeed, define a stochastic matrix $Q_\lambda \in \mathbb{R}^{K\times K}$

$$Q_\lambda(k,\ell) := \sum_{(i,j)\in I\times J} x^1_\lambda(k,i)x^2_\lambda(k,j)q(\ell\,|\,k,i,j) \qquad \forall (k,\ell) \in K^2, \tag{5}$$

so that

$$\Pi(\lambda,x^1_\lambda,x^2_\lambda) = \sum_{m\geq 0} \lambda(1-\lambda)^m Q^m_\lambda.$$

Consider the Markov chain M_λ on $K \cup \{*\}$ defined as follows:

$$M_\lambda(k,\ell) = \begin{cases} (1-\lambda)Q_\lambda(k,\ell) & \text{if } k,\ell \in K\\ \lambda & \text{if } k \in K,\ \ell = *\\ (1+|K|)^{-1} & \text{if } k = *,\ \ell \in K\,. \end{cases}$$

Remark that, if X_n is the Markov chain with transitions M_λ, then

$$\begin{aligned}\sum\nolimits_{m\geq 0}(1-\lambda)^m Q_\lambda^m(k,\ell) &= \mathbb{E}\left[\sum\nolimits_{m=1}^{\tau(K)} \mathbb{1}_{\{X_m=\ell\}} \,|\, X_0=k\right], \\ &= \frac{\sum_{\pi\in G_{k,\ell}(K\setminus\{\ell\})}\prod_{(k',\ell')\in\pi} M_\lambda(k',\ell')}{\sum_{\pi\in G(K)}\prod_{(k',\ell')\in\pi} M_\lambda(k',\ell')}.\end{aligned}$$

where for any set A, $\tau(A) := \inf\{m\geq 0, X_m\notin A\}$, $G(A)$ is the set of acyclic graphs such that exactly one arrow starts from any point of A and no arrow starts outside of A, and $G_{k,\ell}(A)$ is the set of graphs of $G(A)$ such that k leads to ℓ, where $k\in A$ and $\ell\notin A$ (see [4][Lemma 3.1] for a proof). We conclude that $\Pi(\lambda, x_\lambda^1, x_\lambda^2)$ is a Puiseux series since it is the ratio of two finite sums of Puiseux series.

Step 1: *The equality* $\Pi v^* = v^*$.
Define the map $f:(0,\lambda_0)^3\to\mathbb{R}^K$ by setting

$$f(\lambda,\lambda^1,\lambda^2) = \Pi(\lambda, x^1_{\lambda^1}, x^2_{\lambda^2}) g(x^1_{\lambda^1}, x^2_{\lambda^2}) \qquad \forall(\lambda,\lambda^1,\lambda^2)\in(0,\lambda_0)^3.$$

Note that f is differentiable on $(0,\lambda_0)^3$, because it is a power series in the variables λ, $(\lambda^1)^{1/N}$ and $(\lambda^2)^{1/N}$, for some $N\in\mathbb{N}$. For each $k\in K$, x_λ^1 and x_λ^2 are optimal strategies in $\Gamma_\lambda(k)$, so that the map $(\lambda^1,\lambda^2)\to f^k(\lambda,\lambda^1,\lambda^2)$ has a saddle point at (λ,λ), for each $\lambda\in(0,\lambda_0)$. Hence, its partial derivatives satisfy

$$\frac{\partial f}{\partial\lambda^1}(\lambda,\lambda,\lambda) = \frac{\partial f}{\partial\lambda^2}(\lambda,\lambda,\lambda) = 0. \tag{6}$$

For any $\lambda\in(0,1]$, set $h(\lambda) := f(\lambda,\lambda,\lambda)\in\mathbb{R}^K$. By the choice of $(x_\lambda^1, x_\lambda^2)$, $h(\lambda) = v_\lambda$. Define a stochastic matrix $Q_\lambda\in\mathbb{R}^{K\times K}$ by (5) and a payoff vector $g_\lambda\in\mathbb{R}^K$ by:

$$g_\lambda(k) := g^k(x_\lambda^1, x_\lambda^2) = \sum_{(i,j)\in I\times J} x_\lambda^1(k,i) x_\lambda^2(k,j) g(k,i,j) \qquad \forall k\in K.$$

The real $Q_\lambda(k,\ell)$ represents the probability that tomorrow's state is ℓ, given that the state is k today and players play $(x_\lambda^1, x_\lambda^2)$. The relation (6) implies that the derivative of h satisfies

$$\begin{aligned} h'(\lambda) &= \left(\frac{\partial}{\partial\lambda} + \frac{\partial}{\partial\lambda^1} + \frac{\partial}{\partial\lambda^2}\right) f(\lambda,\lambda^1,\lambda^2)_{|\lambda^1=\lambda^2=\lambda} \\ &= \frac{\partial}{\partial\lambda} f(\lambda,\lambda^1,\lambda^2)_{|\lambda^1=\lambda^2=\lambda} \\ &= \sum_{m\geq 0}(1-\lambda)^m Q_\lambda^m g_\lambda - \sum_{m\geq 0} m\lambda(1-\lambda)^{m-1} Q_\lambda^m g_\lambda. \end{aligned}$$

As $\Pi(\lambda, x_\lambda^1, x_\lambda^2)\, g_\lambda = \sum_{m\geq 0}\lambda(1-\lambda)^m Q_\lambda^m g_\lambda = v_\lambda$, it follows that

$$\begin{aligned} \Pi(\lambda, x_\lambda^1, x_\lambda^2)\, v_\lambda &= \Pi(\lambda, x_\lambda^1, x_\lambda^2)\Pi(\lambda, x_\lambda^1, x_\lambda^2)\, g_\lambda \\ &= \sum_{m\geq 0, n\geq 0} \lambda^2(1-\lambda)^{n+m} Q_\lambda^{m+n} g_\lambda \\ &= \sum_{m\geq 0}(m+1)\lambda^2(1-\lambda)^m Q_\lambda^m g_\lambda \\ &= \lambda v_\lambda + \lambda(1-\lambda)\sum_{m\geq 0} m\lambda(1-\lambda)^{m-1} Q_\lambda^m g_\lambda, \end{aligned}$$

where λ^2 stands for "λ square" in the two previous equations. Consequently, replacing the expression of $h'(\lambda)$ one obtains

$$\begin{aligned} \Pi(\lambda, x_\lambda^1, x_\lambda^2) v_\lambda &= \lambda v_\lambda + \lambda(1-\lambda)\left(\lambda^{-1} v_\lambda - h'(\lambda)\right) \\ &= v_\lambda - \lambda(1-\lambda) h'(\lambda). \end{aligned}$$

Since each coordinate of $h(\lambda)$ is a bounded real Puiseux series, it follows that $\lim_{\lambda\to 0}\lambda h'(\lambda) = 0$. Taking λ to 0 in the previous expression thus gives $\Pi v^* = v^*$.

Step 2: *Relation between* Π *and the occupation measure at time* t.
For every $\lambda \in (0,1]$, by the Markov property,

$$\Pi(\lambda, x_\lambda^1, x_\lambda^2) = \sum_{m=1}^{\varphi(\lambda,t)} \lambda(1-\lambda)^{m-1} Q_\lambda^{m-1} + (1-\lambda)^{\varphi(\lambda,t)} Q_\lambda^{\varphi(\lambda,t)} \Pi(\lambda, x_\lambda^1, x_\lambda^2)\,. \tag{7}$$

In order to establish Theorem 3.1, we are going to prove that tv^* is the only accumulation point of $(\gamma_\lambda(x_\lambda^1, x_\lambda^2; t))$, as λ vanishes. Let $\gamma^* \in \mathbb{R}^K$ be such an accumulation point, and let $g^* := \lim_{\lambda\to 0} g_\lambda \in \mathbb{R}^K$. By consecutive extractions, one can find a vanishing sequence (λ_r) such that $(\gamma_{\lambda_r}(x_{\lambda_r}^1, x_{\lambda_r}^2; t))$ converges to γ^*, $Q_{\lambda_r}^{\varphi(\lambda_r,t)}$ converges to some $\pi_t \in \mathbb{R}^{K\times K}$, and $\sum_{m=1}^{\varphi(\lambda_r,t)} \lambda_r(1-\lambda_r)^{m-1} Q_{\lambda_r}^{m-1}$ converges to some $\Pi_t \in \mathbb{R}^{K\times K}$. In particular, $\gamma^* = \Pi_t g^*$, and thus our aim is to prove that $\Pi_t g^* = tv^*$.

Setting $\lambda = \lambda_r$ and having r going to infinity in 7, we obtain

$$\Pi = \Pi_t + (1-t)\pi_t \Pi\,. \tag{8}$$

Iterating this equation, one gets

$$\Pi = \sum_{m\geq 0} (1-t)^m \pi_t^m \Pi_t\,. \tag{9}$$

Set $P_t := \frac{1}{t}\Pi_t$, which is a stochastic matrix on the state space K, and note that the previous relation can be expressed as

$$\Pi = \mathbb{E}(\pi_t^X) P_t, \tag{10}$$

where $X+1$ is a geometric random variable with parameter t, i.e. $\mathbb{P}(X=m) = t(1-t)^m$ for all $m \geq 0$.

Step 3: *A lemma on stochastic matrices.*
Let X be the random variable defined in Step 2. Since $\mathbb{P}(X=0) > 0$, for any stochastic matrix M, the stochastic matrix $N := \mathbb{E}(M^X)$ is aperiodic. Therefore N^n has a limit when n goes to infinity, that we call N^∞. We claim that

$$MN^\infty = N^\infty\,. \tag{11}$$

Indeed, since $\mathbb{P}(X=1) > 0$, $\sup_{n\in\mathbb{N}} M^n(i,j) > 0 \Leftrightarrow \sup_{n\in\mathbb{N}} N^n(i,j) > 0$, so that the recurrent communicating classes of M and N are the same. The number of such classes is equal to the dimension of $\ker(M-I)$, that is therefore also the dimension of $\ker(N-I)$. Moreover, as $Nf = \mathbb{E}(M^X f)$, $Mf = f \Rightarrow Nf = f$, so that

$$\ker(M-I) \subset \ker(N-I).$$

Consequently, these two eigenspaces are equal, since they have the same dimension. Since $(N-I)N^\infty = 0$, the columns of N^∞ belong to $\ker(N-I)$, and therefore also to $\ker(M-I)$, so that $(M-I)N^\infty = 0$ as claimed.

Step 4: *The equality* $\pi_t v^* = v^*$ *for all* $t \in (0,1)$.
Let P_t^∞ be an accumulation point of the sequence $(P_t^n)_n$. The matrices $\mathbb{E}(\pi_t^X)$ and P_t commute because they are limits of weighted sums of powers of Q_λ, which commute. Hence, using the equality $\Pi v^* = v^*$ established in Step 1, and the relation (10), it follows that for all $n \in \mathbb{N}$,

$$\begin{aligned} v^* &= \Pi^n v^* \\ &= \left(\mathbb{E}(\pi_t^X) P_t\right)^n v^* \\ &= \mathbb{E}(\pi_t^X)^n P_t^n v^*\,. \end{aligned}$$

Thus, as n tends to infinity along a subsequence defining P_t^∞, one has

$$v^* = \mathbb{E}(\pi_t^X)^\infty P_t^\infty v^*\,. \tag{12}$$

Combining the equality (11) of Step 3 with $M = \pi_t$ and $N = \mathbb{E}(\pi_t^X)$, and (12), one obtains

$$\begin{aligned} \pi_t v^* &= \pi_t \mathbb{E}(\pi_t^X)^\infty P_t^\infty v^* \\ &= \mathbb{E}(\pi_t^X)^\infty P_t^\infty v^* \\ &= v^*\,. \end{aligned}$$

Step 5: *Conclusion:* $\Pi_t g^* = tv^*$ *for all* $t \in [0,1]$.
Multiplying the two sides of (8) by g^* yields

$$\Pi g^* = \Pi_t g^* + (1-t)\pi_t \Pi g^* .$$

Yet, for all $\lambda \in (0,1]$ one has $\Pi(\lambda, x^1_\lambda, x^2_\lambda) g_\lambda = v_\lambda$. Taking limits as λ goes to 0, it follows that $\Pi g^* = v^*$. Combined with the equality $\pi_t v^* = v^*$ obtained in Step 4, this gives $v^* = \Pi_t g^* + (1-t)v^*$, so that $\Pi_t g^* = tv^*$. □

Remark. In Step 1 we obtained the following expression for the derivative of v_λ:

$$\frac{\partial}{\partial \lambda} v_\lambda = \frac{1}{\lambda(1-\lambda)} \left(v_\lambda - \Pi(\lambda, x^1_\lambda, x^2_\lambda) v_\lambda \right) .$$

4 Strong constant payoff property

We now prove our main result: Theorem 2.3. Roughly speaking, we want to prove that the constant payoff property, which is true for any optimal Puiseux strategy profile, holds for *any* pair of optimal strategies. The main idea is the following: an equivalence between the strong constant payoff property and the convergence to 0 of the values of a certain class of discounted Markov decision processes. We start with a technical property for real sequences, from which we derive an equivalent formulation of the strong constant payoff property.

For each $k \in K$, define $X^1_\lambda(k) \subset \Delta(I)$ (resp., $X^2_\lambda(k) \subset \Delta(J)$) the set of optimal strategies for Player 1 (resp., 2) in the one-shot zero-sum game with action sets I and J and payoff:

$$R(i,j) := \lambda g(k,i,j) + (1-\lambda) \sum_{\ell \in K} q(\ell | k,i,j) v_\lambda(\ell) . \tag{13}$$

The following lemma is a direct consequence of [12, Corollary 2.6.3]. We state it for Player 1 but, as players have symmetric roles, a similar result holds for Player 2.

Lemma 4.1. *A general strategy* σ^1 *of Player* 1 *is optimal in the discounted stochastic game* Γ_λ *if, and only if, for any* $k_1 \in K$*, for any* $m \geq 1$*, for any strategy* $\sigma^2 \in \Sigma^2$ *of Player* 2 *and any finite history* $h^m \in H_m$ *such that* $\mathbb{P}^{k_1}_{\sigma^1,\sigma^2}(h^m) > 0$*, Player* 1 *plays a mixed action in* $X^1_\lambda(k_m)$.

4.1 Characterisation of the strong constant payoff property

4.1.1 A technical lemma on real sequences

Let $(u^\lambda_m)_{m \geq 1}$ be a fixed family of real sequences so that, for some constant $C \geq 0$, for all $\lambda \in (0,1]$ and all $m \geq 1$,

$$\left| u^\lambda_{m+1} - u^\lambda_m \right| \leq C\lambda \quad \text{and} \quad \left| u^\lambda_m \right| \leq C . \tag{14}$$

For each $\delta > 0$ we set

$$B_\lambda(\delta) := \sum\nolimits_{m \geq 1} \delta\lambda(1-\delta\lambda)^{m-1} u^\lambda_m .$$

Proposition 4.2. *The two following statements are equivalent:*

(*i*) *For all* $t \in (0,1)$, $u^\lambda_{\varphi(\lambda,t)}$ *vanishes as* λ *tends to 0.*

(*ii*) *For all* $\delta > 0$, $B_\lambda(\delta)$ *vanishes as* λ *tends to 0.*

Proof. Consider the functions

$$f_\lambda(x) = u^\lambda_{\lfloor x/\lambda \rfloor + 1} \Big(1 - x/\lambda + \lfloor x/\lambda \rfloor \Big) + u_{\lfloor x/\lambda \rfloor + 2} \Big(x/\lambda - \lfloor x/\lambda \rfloor \Big), \quad x \geq 0, \lambda > 0.$$

For all $m \geq 0$, the function f_λ is linear on each interval $[m\lambda, (m+1)\lambda]$, and satisfies $f_\lambda(m\lambda) = u^\lambda_{m+1}$. Remark that $\sup_{x \geq 0} |f_\lambda(x)| \leq C$ and that

$$\left| f_\lambda(y) - f_\lambda(x) \right| \leq C(y-x), \quad 0 \leq x < y.$$

Since $\varphi(\lambda, t) = \left\lceil \frac{\ln(1-t)}{\ln(1-\lambda)} \right\rceil$, one can easily see that for any $t \in (0,1)$,

$$\lim_{\lambda \to 0+} u^\lambda_{\varphi(\lambda,t)} - f_\lambda\big(-\ln(1-t)\big) = 0. \tag{15}$$

Elementary computations also show that for any $\delta > 0$,

$$\lim_{\lambda \to 0+} B_\lambda(\delta) - \int_0^{+\infty} \delta \exp(-\delta x) f_\lambda(x)\, \mathrm{d}x = 0. \tag{16}$$

The continuity properties of the Laplace transform ensure that $\lim_{\lambda\to 0+} f_\lambda(x) = 0$, $x > 0$ if and only if

$$\lim_{\lambda \to 0+} \int_0^{+\infty} \exp(-\delta x) f_\lambda(x)\, \mathrm{d}x = 0, \quad \delta > 0.$$

To see this, we can for example apply [5, XIII.1 Theorem 2. page 431] to the family of probability distributions

$$Z_{\lambda,\alpha}^{-1} \exp(-\alpha x)\big[f_\lambda(x) + 2C\big]\, \mathrm{d}x, \quad \lambda > 0,\ \alpha > 0,$$

on $[0, +\infty)$, where

$$Z_{\lambda,\alpha} = \int_0^{+\infty} \exp(-\alpha x)\big[f_\lambda(x) + 2C\big]\, \mathrm{d}x.$$

This proves the proposition in view of (15) and (16). □

4.1.2 Application to stochastic games

We now provide several alternative characterisations of the strong constant payoff property which will be used in the proof of Theorem 2.3.

Definition 4.3. A family $(\sigma_\lambda^1, \sigma_\lambda^2)_\lambda$ is a *discounted optimal strategy profile* if for all $\lambda \in (0, 1]$, $(\sigma_\lambda^1, \sigma_\lambda^2)$ is a pair of optimal strategies in Γ_λ.

Proposition 4.4. *Let $(\sigma_\lambda^1, \sigma_\lambda^2)_\lambda$ be a discounted optimal strategy profile. The following conditions are equivalent:*

(i) The family $(\sigma_\lambda^1, \sigma_\lambda^2)_\lambda$ satisfies the constant payoff property for all $k \in K$:

$$\lim_{\lambda \to 0} \gamma_\lambda^k(\sigma_\lambda^1, \sigma_\lambda^2; t) = t v^*(k) \qquad \forall t \in [0, 1]\,.$$

(ii) For all $k \in K$, for all $t \in [0, 1)$, $\mathbb{E}^k_{\sigma_\lambda^1, \sigma_\lambda^2}[v_\lambda(k_{\varphi(\lambda,t)})] - v_\lambda(k)$ converges to 0 as λ vanishes.

(iii) For all $k \in K$, for all $\delta > 0$ one has:

$$\lim_{\lambda \to 0} \mathbb{E}^k_{\sigma_\lambda^1, \sigma_\lambda^2} \left[\sum\nolimits_{m \geq 1} \delta\lambda(1 - \delta\lambda)^{m-1} (v_\lambda(k_m) - v_\lambda(k)) \right] = 0. \tag{17}$$

Proof. We start by proving the equivalence between (i) and (ii). Fix $k \in K$ and $t \in (0, 1)$. For any $\lambda \in (0, 1]$, Shapley's equation yields

$$v_\lambda(k) = \mathbb{E}^k_{\sigma_\lambda^1, \sigma_\lambda^2} \left[\sum\nolimits_{m=1}^{\varphi(\lambda,t)-1} \lambda(1-\lambda)^{m-1} g_m \right] + (1-\lambda)^{\varphi(\lambda,t)-1} \mathbb{E}^k_{\sigma_\lambda^1, \sigma_\lambda^2} \left[v_\lambda(k_{\varphi(\lambda,t)}) \right],$$

where k_m stands for the state at stage m, for any $m \in \mathbb{N}$. Consequently,

$$\begin{aligned} \mathbb{E}^k_{\sigma_\lambda^1, \sigma_\lambda^2} \left[\sum\nolimits_{m=1}^{\varphi(\lambda,t)-1} \lambda(1-\lambda)^{m-1} g_m \right] - t v_\lambda(k) + (1-\lambda)^{\varphi(\lambda,t)-1} \left(\mathbb{E}^k_{\sigma_\lambda^1, \sigma_\lambda^2} \left[v_\lambda(k_{\varphi(\lambda,t)}) \right] - v_\lambda(k) \right) \\ = \left(1 - t - (1-\lambda)^{\varphi(\lambda,t)-1} \right) v_\lambda(k)\,. \end{aligned}$$

The equivalence between (i) and (ii) is obtained by taking λ to 0, and recalling the relation $\lim_{\lambda\to 0}(1-\lambda)^{\varphi(\lambda,t)-1} = 1 - t$.

We now prove the equivalence between (i) and (iii). For each $k \in K$, $\lambda \in (0, 1]$ and $m \geq 1$, define

$$u_m^\lambda := \mathbb{E}^k_{\sigma_\lambda^1, \sigma_\lambda^2}[v_\lambda(k_m)] - v_\lambda(k)\,.$$

Note that the family of sequences (u_m^λ) satisfies (14) with $C = \max_{k,i,j} |g(k, i, j)|$. Therefore, Proposition 4.2 applies, and gives the desired result. □

The following result is now a direct consequence of Proposition 4.4.

Corollary 4.5. *Theorem 2.3 holds if and only if for all $(\sigma_\lambda^1, \sigma_\lambda^2)_\lambda$ discounted optimal strategy profile, for all $k \in K$ and $\delta > 0$,*

$$\lim_{\lambda \to 0} \mathbb{E}^k_{\sigma_\lambda^1, \sigma_\lambda^2} \left[\sum\nolimits_{m \geq 1} \delta\lambda(1 - \delta\lambda)^{m-1} (v_\lambda(k_m) - v_\lambda(k)) \right] = 0\,. \tag{18}$$

4.2 Auxiliary MDP and proof of Theorem 2.3

Let $\delta > 0$ and $k \in K$ be fixed. For each $\lambda \in (0,1]$, consider a Markov Decision Process (one-player stochastic game) $MDP_{k,\delta,\lambda}$ with state space K, action set $A_\lambda(\ell) := X^1_\lambda(\ell) \times X^2_\lambda(\ell)$ for each $\ell \in K$, transition function q, payoff function $\ell \mapsto v_\lambda(\ell) - v_\lambda(k)$ and discount factor $\delta\lambda$.

Remark 4.6. At each state, the decision-maker can only play pairs of optimal mixed strategies of the game given in (13). Hence, the sets of possible actions depend on the state and on the discount factor. Similarly, the payoff function does not depend on the actions but depends on the discount factor.

For any pair of optimal strategies $(\sigma^1, \sigma^2) \in \Sigma^1_\lambda \times \Sigma^2_\lambda$ of the original λ-discounted stochastic game and any initial state $\ell \in K$, define

$$h_\lambda(\ell, \sigma^1, \sigma^2) := \mathbb{E}^\ell_{\sigma^1,\sigma^2}\left[\sum\nolimits_{m\geq 1} \delta\lambda(1-\delta\lambda)^{m-1}(v_\lambda(k_m) - v_\lambda(k))\right]. \tag{19}$$

Let $w^\delta_\lambda(\ell)$ denote the value of this MDP with initial state ℓ, i.e.:

$$w^\delta_\lambda(\ell) = \sup_{(\sigma^1,\sigma^2)\in\Sigma^1_\lambda\times\Sigma^2_\lambda} h_\lambda(\ell, \sigma^1, \sigma^2).$$

Proposition 4.7. *Theorem 2.3 holds if and only if for all $k \in K$ and $\delta > 0$ one has $\lim_{\lambda\to 0} w^\delta_\lambda(k) = 0$.*

Proof. This stems from Corollary 4.5. □

End of the proof of Theorem 2.3. Let $\delta > 0$ and $k \in K$ be fixed. By Proposition 4.7, it is enough to prove that $\lim_{\lambda\to 0} w^\delta_\lambda(k) = 0$. Note that, by Theorem 3.1 and Proposition 4.4 *(iii)*, we have $\liminf_{\lambda\to 0} w^\delta_\lambda(k) \geq 0$. Thus, it is enough to prove that $\limsup_{\lambda\to 0} w^\delta_\lambda(k) = 0$. By contradiction, assume that $\limsup_{\lambda\to 0} w^\delta_\lambda(k) > \varepsilon$ for some $\varepsilon > 0$.

We resort to the semi-algebraic approach. Note that, unlike the classical setup described in Section 2.1.2, where a two-player zero-sum stochastic game Γ was fixed and the discount factor λ was put to 0, here we have a Markov Decision Process $MDP_{k,\delta,\lambda}$ (thus, one player only) which depends on λ through its action set and payoff function. Nonetheless, the semi-algebraic approach still applies. Define a subset $S_\varepsilon \subset \mathbb{R} \times \mathbb{R}^K \times \mathbb{R}^{K\times I} \times \mathbb{R}^{K\times J} \times \mathbb{R}^K$ by setting

$$(\lambda, v, x^1, x^2, h) \in S_\varepsilon \quad \text{if, and only if, the following relations hold:}$$

- $\lambda \in \mathbb{R}$ is some discount factor, that is $0 < \lambda \leq 1$.
- $v \in \mathbb{R}^K$ is the vector of values of the λ-discounted stochastic game Γ_λ.
- $(x^1, x^2) \in \mathbb{R}^{K\times I} \times \mathbb{R}^{K\times J}$ is a pair of optimal stationary strategies in Γ_λ.
- $h \in \mathbb{R}^K$ satisfies $h(\ell) = h_\lambda(\ell, x^1, x^2)$ for all $\ell \in K$ and $h(k) \geq \varepsilon$.

The set S_ε is semi-algebraic, as it can be described by the following finite set of polynomial equalities and inequalities (compare with the system in Section 2.1.2):

$$\begin{aligned}
0 < \lambda &\leq 1\\
\forall(\ell,i),\ x^1(\ell,i) \geq 0,\ \text{and}\ \forall\ell,\quad \sum\nolimits_{i\in I} x^1(\ell,i) &= 1\\
\forall(\ell,j),\ x^2(\ell,j) \geq 0,\ \text{and}\ \forall\ell,\quad \sum\nolimits_{j\in J} x^2(\ell,j) &= 1\\
\forall(\ell,j),\quad \sum\nolimits_{i\in I} x^1(\ell,i)\left(\lambda g(\ell,i,j) + (1-\lambda)\sum\nolimits_{\ell'\in K} q(\ell'|\ell,i,j)v(\ell')\right) &\geq v(\ell)\\
\forall(\ell,i),\quad \sum\nolimits_{j\in J} x^2(\ell,j)\left(\lambda g(\ell,i,j) + (1-\lambda)\sum\nolimits_{\ell'\in K} q(\ell'|\ell,i,j)v(\ell')\right) &\leq v(\ell)\\
\forall\ell,\quad \delta\lambda(v_\lambda(\ell) - v_\lambda(k)) + (1-\delta\lambda)\sum\nolimits_{(i,j)\in I\times J} x^1(\ell,i)x^2(\ell,j)\sum\nolimits_{\ell'\in K} q(\ell'|\ell,i,j)h(\ell') &= h(\ell)\\
h(k) &\geq \varepsilon.
\end{aligned}$$

For $\lambda \in (0,1]$, let $S_\varepsilon(\lambda) = \{a \mid (\lambda, a) \in S_\varepsilon\}$. By assumption, $\limsup_{\lambda\to 0} w^\delta_\lambda(k) > \varepsilon$, thus there exists a vanishing subsequence (λ_n) such that for all n, the set $S_\varepsilon(\lambda_n)$ is non-empty. By semi-algebraicity, there exists $\lambda_0 \in (0,1]$ so that $S_\varepsilon(\lambda)$ is non-empty for all $\lambda \in (0,\lambda_0)$. From the Tarski-Seidenberg elimination theorem, it admits a semi-algebraic selection for $\lambda \in (0,\lambda_0)$. In particular, there exists a selection of stationary strategies $z_\lambda := (x^1_\lambda, x^2_\lambda)$ that is a strategy of the Markov decision process $MDP_{k,\delta,\lambda}$, which can be expressed as a Puiseux series near 0, and so that $h_\lambda(k, x^1_\lambda, x^2_\lambda) \geq \varepsilon$ for all λ small enough. But this contradicts Theorem 3.1 and Proposition 4.4 *(iii)* since, together, they imply that $\lim_{\lambda\to 0} h_\lambda(k, x^1_\lambda, x^2_\lambda) = 0$. □

5 Examples and a remark

5.1 An example

Let us illustrate the constant payoff property by an example, studied by Bewley and Kohlberg [3]. The state space is the set $K = \{1^*, k, \ell, 0^*\}$. For all $(i,j) \in I \times J$, one has $q(1^* \,|\, 1^*, i, j) = q(0^* \,|\, 0^*, i, j) = 1$, $g(1^*, i, j) = 1$ and $g(0^*, i, j) = 0$ so that the states 1^* and 0^* are absorbing with payoff 1 and 0 respectively. The transition from states k and ℓ are deterministic and represented by the two following matrices:

	L	R
T	k	ℓ
B	ℓ	1^*

k

	L	R
T	ℓ	k
B	k	0^*

ℓ

The set of actions are $I = \{T, B\}$ and $J = \{L, R\}$. Finally, the payoff function is given by

$$\forall (i,j) \in I \times J, \quad g(k,i,j) = 1 \quad \text{and} \quad g(\ell,i,j) = 0\,.$$

Optimal stationary strategies satisfy $x^1(T) = x^2(L) \to_{\lambda\to 0} 1$ and $x^1_\lambda(B) = x^2_\lambda(R) \sim_{\lambda\to 0} \sqrt{\lambda}$, so that the induced Markov chain satisfies

$$Q_\lambda \sim_{\lambda\to 0} \begin{pmatrix} 1 & 0 & 0 & 0 \\ \lambda & 1 & 2\sqrt{\lambda} & 0 \\ 0 & 2\sqrt{\lambda} & 1 & \lambda \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The limit payoff vector is given by $g^* = (1,1,0,0) \in \mathbb{R}^K$ and, by the symmetry of the game, the vector of limit values is $v^* = (1, 1/2, 1/2, 0) \in \mathbb{R}^K$. Let $t \in (0,1)$. A direct calculation yields that $Q_\lambda^{\varphi(\lambda,t)}$ converges to some π_t, and $\sum_{m=1}^{\varphi(\lambda,t)} \lambda(1-\lambda)^{m-1} Q_\lambda^{m-1}$ converges to some Π_t, such that

$$\Pi_t = \begin{pmatrix} t & 0 & 0 & 0 \\ \frac{t^2}{4} & \frac{2t-t^2}{4} & \frac{2-t^2}{4} & \frac{t^2}{4} \\ \frac{t^2}{4} & \frac{2t-t^2}{4} & \frac{2t-t^2}{4} & \frac{t^2}{4} \\ 0 & 0 & 0 & t \end{pmatrix}, \quad \pi_t = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{t}{2} & \frac{1-t}{2} & \frac{1-t}{2} & \frac{t}{2} \\ \frac{t}{2} & \frac{1-t}{2} & \frac{1-t}{2} & \frac{t}{2} \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

In particular,

$$\Pi = \Pi_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

One can thus easily check the equality $\Pi v^* = v^*$, and that $\Pi_t g^* = t v^*$ and $\pi_t v^* = v^*$ hold for all $t \in (0,1)$.

5.2 The constant payoff is a joint property

Contrary to non-zero sum games, where the notion of Nash equilibrium is a joint property of the players' strategies, the notion of optimality is unilateral in zero-sum games. Indeed, by playing an optimal strategy, Player 1 ensures that his payoff is greater than or equal to the value regardless of the strategy used by his opponent (and similarly for Player 2). Naturally, one would like to know whether the constant payoff property is an unilateral property as well. That is, can Player 1 ensure that the average payoff at time t is greater than or equal to the value at all times $t \in [0,1]$?

The following example gives a negative answer to this question: playing an optimal strategy does not ensure that the average payoffs are greater than or equal to the value at all times. Rather, the constant payoff property requires *both players* to play optimally. The example is "as bad as it can be", since the unique optimal strategy of Player 1 guarantees strictly less than the value at any time $t \in (0,1)$, where the fact that this property holds for $t = 1$ follows from the optimality of his strategy.

Consider the "Big Match", introduced by Gillette [6], a stochastic game with set of states $K = \{k, 0^*, 1^*\}$, action sets $I = \{T, B\}$ and $J = \{L, R\}$, and where states 0^* and 1^* are absorbing with payoff 0 and 1 respectively, i.e. for all $(i, j) \in I \times J$,

$$q(\cdot\,|0^*, i, j) = \delta_{0^*}, \quad q(\cdot\,|1^*, i, j) = \delta_{1^*}, \quad g(0^*, i, j) = 0 \quad \text{and} \quad g(1^*, i, j) = 1.$$

The game with initial state k, the non-absorbing state, can be represented as follows:

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

As far as Player 1 plays action B, he receives stage payoffs 0 or 1, depending on whether Player 2 plays L or R, and the state does not change, i.e.

$$g(k, B, L) = 0, \quad g(k, B, R) = 1 \quad \text{and} \quad q(k\,|\,k, B, L) = q(k\,|\,k, B, R) = 1.$$

When Player 1 plays T, the state moves to an absorbing state, indicated by a $*$ in the picture above, depending on the action of his opponent. For all $\lambda \in (0, 1]$, the value and the unique optimal stationary strategy profile are given by

$$v_\lambda(k) = \frac{1}{2}, \qquad x^1_\lambda(k, T) = \frac{\lambda}{1+\lambda} \quad \text{and} \quad x^2_\lambda(k, L) = \frac{1}{2}.$$

Let x^2 be the strategy that plays L at every stage. Though not optimal, x^2 is a best reply to x^1_λ since $\gamma_\lambda(k, x^1_\lambda, x^2) = v_\lambda(k)$ for all λ. Computations show that

$$\lim_{\lambda \to 0} \gamma_\lambda(k, x^1_\lambda, x^2; t) = \frac{t^2}{2}\,.$$

Since $\frac{t^2}{2} < tv^*(k)$ for all $t \in (0, 1)$, (x^1_λ, x^2) does not satisfy the constant payoff property. In fact, under these strategies, Player 2 obtains strictly less than the value (and this is favorable to him) at all times except for $t = 1$.

Acknowledgments

We are greatly indebted to Sylvain Sorin, whose comments have led to significant improvements in the presentation of the paper. We are also very thankful to Abraham Neyman for his careful reading and numerous remarks, and also to Cyril Labbé, Rida Laraki, Eran Shmaya and Guillaume Vigeral for helpful discussions.

References

[1] L. Attia and M. Oliu-Barton, *A formula for the value of a stochastic game*, Proceedings of the National Academy of Sciences of the USA **116** (2019), no. 52, 26435–26443.

[2] T. Bewley and E. Kohlberg, *The asymptotic theory of stochastic games*, Mathematics of Operation Research **1** (1976), no. 3, 197–208.

[3] ———, *On stochastic games with stationary optimal strategies*, Mathematics of Operation Research **3** (1978), no. 2, 104–125.

[4] O. Catoni, *Simulated annealing algorithms and markov chains with rare transitions*, Séminaire de probabilités XXXIII, Springer, 1999, pp. 69–119.

[5] W. Feller, *An introduction to probability theory and its applications vol. ii*, John Wiley & Sons, 1971.

[6] D. Gillette, *Stochastic games with zero stop probabilities*, Contributions to the Theory of Games, III, *M. Dresher, A.W. Tucker and P. Wolfe (eds.)*, Annals of the Mathematical Studies, 39, Princeton University Press (1957), 179–187.

[7] E. Lehrer and S. Sorin, *A uniform tauberian theorem in dynamic programming*, Mathematics of Operation Research **17** (1992), no. 2, 303–307.

[8] J.-F. Mertens and A. Neyman, *Stochastic games*, International Journal of Game Theory **10** (1981), no. 2, 53–66.

[9] M. Oliu-Barton, *The asymptotic value in stochastic games*, Mathematics of Operations Research **39** (2014), no. 3, 712–721.

[10] ———, *The splitting game: Value and optimal strategies*, Dynamic Games and Applications **8** (2018), no. 1, 157–179.

[11] ———, *New algorithms for solving zero-sum stochastic games*, Mathematics of Operations Research (2020), https://doi.org/10.1287/moor.2020.1055.

[12] J. Renault, *Basics of game theory (class notes)*, (2017).

[13] L.S. Shapley, *Stochastic games*, Proceedings of the National Academy of Sciences of the USA **39** (1953), no. 10, 1095–1100.

[14] S. Sorin, *A First Course on Zero-Sum Repeated Games*, Springer, 2002.

[15] ———, *The operator approach to zero-sum stochastic games*, Stochastic Games and Applications, NATO Science Series C, Mathematical and Physical Sciences **570** (2003), 417–426.

[16] S. Sorin, X. Venel, and G. Vigeral, *Asymptotic properties of optimal trajectories in dynamic programming*, Sankhya **72** (2010), no. 1, 237–245.

[17] S. Sorin and G. Vigeral, *Limit optimal trajectories in zero-sum stochastic games*, Dynamic Games and Applications (2019), 1–18.

[18] G. Vigeral, *A zero-sum stochastic game with compact action sets and no asymptotic value*, Dynamic Games and Applications **3** (2013), no. 2, 172–186.

[19] B. Ziliotto, *Zero-sum repeated games: counterexamples to the existence of the asymptotic value and the conjecture maxmin= lim v (n)*, The Annals of Probability **44** (2016), no. 2, 1107–1133.