Playing in stochastic environment: from multi-armed bandits to two-player games

Wiesław Zielonka[1]

1 **LIAFA, Université Paris 7 Denis Diderot, Paris, France**
zielonka@liafa.jussieu.fr

Abstract

Given a zero-sum infinite game we examine the question if players have optimal memoryless deterministic strategies. It turns out that under some general conditions the problem for two-player games can be reduced to the same problem for one-player games which in turn can be reduced to a simpler related problem for multi-armed bandits.

Digital Object Identifier 10.4230/LIPIcs.FSTTCS.2010.65

1 Introduction

Activities of a computer system interacting with the environment are often modeled as two-player games with one player representing the system and the other player impersonating the environment. In the worst case analysis we assume that the environment is hostile and then we deal with two-player zero-sum games. Traditionally in verification and in automata theory we use some variants of parity games [7] while the traditional game theory focuses on mean-payoff games, discounted games and total payoff games [1].

Parity games capture qualitative system properties, but sometimes this is not enough and we are interested in finer quantitative analysis. Mean-payoff and total payoff games capture quantitative properties but do not seem really pertinent in the context of computer systems. For these reasons there were recently several attempts to define new quantitative measures or payoffs better suited to the analysis of computer systems. This is an ongoing activity, each such attempt gives rise to a new game (a new payoff mapping).

And the recurrent basic question arising when new games (payoffs) are introduced is if players have optimal strategies. However for a computer scientist the existence of optimal or nearly optimal strategies is not sufficient, we want to be able to implement effectively such strategies and strategies requiring an unbounded memory are unfeasible from the practical point of view. A finite memory can often be incorporated directly into the game and then it is sufficient to answer the simpler question if the players have optimal memoryless strategies. Instead of examining various games one by one with some ad hoc methods it is much more interesting to look for general sufficient conditions guaranteeing the existence of optimal memoryless strategies. Such conditions are useful only if they are robust and can be applied to a sufficiently large class of games.

Our aim is to present such general conditions, we do it first for one-player games (Markov decision processes), next for two-player games.

2 Perfect information stochastic games – basic definitions

2.1 Notation.

For each finite set X, $\mathcal{D}(X)$ is the set of probability measures over X, i.e. it is the set of mappings $p : X \to [0,1]$ such that $\sum_{x \in X} p(x) = 1$. The support of $p \in \mathcal{D}(X)$ is the set

IARCS Int'l Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2010).
Editors: Kamal Lodaya, Meena Mahajan; pp. 65–72
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

$\{x \in X \mid p(x) > 0\}$.

X^i will denote the set of all finite sequences (words) of length i composed of elements of X, $X^* = \cup_{i=0}^{\infty} X^i$ is the set of all finite words over X, X^ω will stand for the set of all infinite words over X. We endow X^ω with the structure of a topological space with open sets of the form $\bigcup_{u \in L} uX^\omega$ for $L \subseteq X^*$. By $\mathcal{B}(X^\omega)$ we denote the smallest σ-algebra containing all open sets (the Borel σ-algebra). Thus $(X^\omega, \mathcal{B}(X^\omega))$ is a measurable space.

By $\pi_i : X^\omega \to X$, $i \in \mathbb{N}$, we denote the mapping defined as $\pi_i(x_1x_2x_3\ldots) = x_i$, i.e. the mapping giving the ith element of an infinite word.

If we equip X with the σ-algebra $\mathcal{P}(X)$ of all subsets of X then π_i are measurable mappings from $(X^\omega, \mathcal{B}(X^\omega))$ to $(X, \mathcal{P}(X))$ and, given any probability measure P on $(X^\omega, \mathcal{B}(X^\omega))$, $\{\pi_i; i \in \mathbb{N}\}$ becomes a discrete process with values in X. Note that $\mathcal{B}(X^\omega)$ is in fact the smallest σ-algebra such that all π_i are measurable.

2.2 Games and Arenas

Two players, Min and Max, play an infinite game on an arena

$$\mathcal{A} = (\mathbf{S}, \mathbf{A}, \text{source}, \delta, \text{player}) \tag{1}$$

where

- $\mathbf{S}$ is a finite set of states,
- $\mathbf{A}$ is a finite set of actions,
- source : $\mathbf{A} \to \mathbf{S}$ provides for each action $a \in \mathbf{A}$ a state source$(a) \in \mathbf{S}$ called the source of a. Action a can be executed only if the current state is $s = \text{source}(a)$ and then we say that a is available at s. We write $\mathbf{A}(s)$ for the set of actions available at s and we assume that $\mathbf{A}(s) \neq \emptyset$ for each state s.
- The dynamic aspect of $\mathcal{A}$ is described by δ, for each action $a \in \mathbf{A}$ and each state $s \in \mathbf{S}$ $\delta(a, s)$ is the probability of going to a state s if a is executed. It is tacitly assumed that a can be executed only if $\mathcal{A}$ is at the state source(a). For each action $a \in \mathbf{A}$, $\delta(a, \cdot)$ is a probability distribution over $\mathbf{S}$, i.e. $\delta(a, \cdot) \in \mathcal{D}(\mathbf{S})$.
- Finally, player : $\mathbf{S} \to \{\text{Min}, \text{Max}\}$ is a mapping assigning to each state $s \in \mathbf{S}$ the player player(s) controlling s.

The game is played by stages. If at stage $i \in \mathbb{N}$ the game is at state $s_i \in \mathbf{S}$ then player player(s_i) chooses an available action $a_i \in \mathbf{A}(s_i)$ and the game enters a new state s_{i+1} with probability $\delta(a_i, s_{i+1})$.

Let $\mathbf{S}_{\text{Max}} = \{s \in \mathbf{S} \mid \text{player}(s) = \text{Max}\}$ be the set of states controlled by player Max. A strategy σ for player Max is a mapping

$$\sigma : \mathbf{A}^* \times \mathbf{S}_{\text{Max}} \to \mathcal{D}(\mathbf{A})$$

such that $\sigma(h, s) \in \mathcal{D}(\mathbf{A}(s))$ for $h \in \mathbf{A}^*$ and $s \in \mathbf{S}_{\text{Max}}$. Intuitively, if the game is at state $s \in \mathbf{S}_{\text{Max}}$, h is the sequence of executed actions and player Max plays using strategy σ then Max will play action $a \in \mathbf{A}(s)$ with probability $\sigma(h, s)(a)$.

The strategy σ is *memoryless* (or stationary) if the past history is not taken into account, i.e. if $\sigma(h, s) = \sigma(\mathbf{1}, s)$ for each finite history $h \in \mathbf{A}^*$, where $\mathbf{1}$ is the empty history.

The strategy σ is *deterministic* (or pure) if for each finite history h and each state $s \in \mathbf{S}_{\text{Max}}$ the support of $\sigma(h, s)$ consists of one action.

Thus a memoryless deterministic strategy σ for player Max is just a mapping $\sigma : \mathbf{S}_{\text{Max}} \to \mathbf{A}$ such that, for each state $s \in \mathbf{S}_{\text{Max}}$, $\sigma(s) \in \mathbf{A}(s)$. Intuitively, $\sigma(s)$ is the action that player Max plays each time s is visited.

Strategies of player Min (general, memoryless, deterministic) are defined mutatis mutandis.

Our basic probability space associated with a given arena is the space $(\mathbf{A}^\omega, \mathcal{B}(\mathbf{A}^\omega))$ of infinite histories (infinite action sequences) equipped with the Borel σ-algebra. The basic stochastic process associated with each game is the process $\{A_i; i \in \mathbb{N}\}$ with values in $\mathbf{A}$, where A_i is the action taken at stage i. Another process of interest is the auxiliary stochastic process $\{S_i; i \in \mathbb{N}\}$ with values in $\mathbf{S}$ defined as $S_i = \text{source} \circ A_i$, i.e. S_i gives the source of the ith action (or equivalently the state at stage i).

Fixing strategies σ and τ of players Max and Min and an initial probability distribution over states $\mu \in \mathcal{D}(\mathbf{S})$ there exists a unique probability measure $\mathrm{P}_\mu^{\sigma,\tau}$ on $(\mathbf{A}^\omega, \mathcal{B}(\mathbf{A}^\omega))$ satisfying the following conditions:

$$\mathrm{P}_\mu^{\sigma,\tau}\{S_1 = s\} = \mu(s), \tag{2}$$

i.e. the initial state probability is given by μ,

$$\mathrm{P}_\mu^{\sigma,\tau}\{A_{n+1} = a_{n+1} | A_1 = a_1, \ldots, A_n = a_n, S_{n+1} = s_{n+1}\} = \\ \begin{cases} \sigma(a_1 \ldots a_n, s_{n+1})(a_{n+1}) & \text{if player}(s_{n+1}) = \text{Max}, \\ \tau(a_1 \ldots a_n, s_{n+1})(a_{n+1}) & \text{if player}(s_{n+1}) = \text{Min}, \end{cases} \tag{3}$$

i.e. if $a_1 \ldots a_n$ is the current history and s_{n+1} the current state then the probability distribution over actions taken on stage $n+1$ is dictated by the strategy of the player controlling s_{n+1},

$$\mathrm{P}_\mu^{\sigma,\tau}\{S_{n+1} = s_{n+1} | A_1 = a_1, \ldots, A_n = a_n\} = \delta(a_n, s_{n+1}), \tag{4}$$

i.e. the state on stage $n+1$ depends only on the action executed at stage n.

2.3 Payoff mappings

After an infinite play player Max receives a payoff from player Min. The players have opposite goals, Max wishes to maximize the payoff while player Min wants to minimize the payoff.

A payoff function is a Borel measurable mapping

$$u : \mathbf{A}^\omega \to (-\infty, \infty]$$

from infinite histories to real numbers extended with plus infinity. To avoid integrability problems we assume that u is bounded from below, i.e. there exists $K \in \mathbb{R}$ such that $u(h) \geq K$ for all $h \in \mathbf{A}^\omega$, and we note by $\mathcal{M}_b$ the class of such payoff functions.

A game is a couple $\Gamma = (\mathcal{A}, u)$ made of an arena and a payoff function $u \in \mathcal{M}_b$.

Let us recall that the *tail σ-algebra* relative to the sequence $\{A_i; i \in \mathbb{N}\}$ of r.v. is the σ-algebra $\bigcap_{n=1}^\infty \sigma(A_i; i \geq n)$, where $\sigma(A_i; i \geq n)$ is the σ-algebra generated by random variables $\{A_i; i \geq n\}$. Thus a payoff function u is measurable relative to the tail σ-algebra if and only if u is measurable relative to $(\mathbf{A}^\omega, \mathcal{B}(\mathbf{A}^\omega))$ and u does not depend on initial finite histories, i.e. $u(a_1a_2\ldots) = u(a_2\ldots)$ for each history $h = a_1a_2\ldots \in \mathbf{A}^\omega$. We note by $\mathcal{T}_b$ the class of all tail measurable mappings belonging to $\mathcal{M}_b$.

2.3.1 Mean-payoff games

A reward function is a function $r : \mathbf{A} \to \mathbb{R}$. Given a reward function r the payoff of a mean-payoff game is calculated as follows:

$$u(a_1a_2a_3\ldots) = \limsup \frac{1}{n} \sum_{i=1}^n r(a_i).$$

Since for a given arena the set $\mathbf{A}$ of actions is finite the payoff of mean-payoff games belongs to $\mathcal{T}_b$.

2.3.2 Parity games

In parity games we assume that there is a priority mapping $\alpha : \mathbf{A} \to \mathbb{N}$ and the payoff is calculated as

$$u(a_1 a_2 a_3 \dots) = (\liminf \alpha(a_i)) \mod 2.$$

Again this payoff mapping belongs to $\mathcal{T}_b$.

2.4 Priority mean-payoff games

In priority mean-payoff games we combine priorities and rewards. There are several forms of such games [5, 4, 3] but the most general one is defined as follows. We have three mappings $\alpha : \mathbf{A} \to \mathbb{N}$, $w : \mathbf{A} \to \mathbb{R}_+$ and $r : \mathbf{A} \to \mathbb{R}$ assigning to each state a non-negative integer priority, a positive real weight and a real reward respectively.

The payoff is calculated in the following way. For each infinite sequence $x = a_1 a_2 a_3 \dots$ of actions we extract the subsequence $a_{i_1} a_{i_2} a_{i_3} \dots$ composed of all actions with priority c where c is the minimal priority such that the set $\{i \mid \alpha(a_i) = c\}$ is infinite. Then the payoff for x is calculated as

$$u(x) = \limsup \frac{\sum_{k=1}^{n} w(a_{i_k}) r(a_{i_k})}{\sum_{k=1}^{n} w(a_{i_k})}$$

i.e. this is a weighted mean-payoff but calculated only over actions with the minimal priority visited infinitely often.

The games with such payoff contain as special cases parity games as well as mean-payoff games.

2.5 Optimal strategies

Given an initial state distribution μ and strategies σ and τ of Max and Min the expected value of the payoff u under $\mathrm{P}_{\mu}^{\sigma,\tau}$ is denoted $\mathrm{E}_{\mu}^{\sigma,\tau}[u]$.

If μ is such that $\mu(s) = 1$ for some state s then to simplify the notation the corresponding probability and expectation are noted $\mathrm{P}_s^{\sigma,\tau}$ and $\mathrm{E}_s^{\sigma,\tau}$.

Given a game $(\mathcal{A}, u)$ strategies $\sigma^\sharp$ and $\tau^\sharp$ of players Max and Min are said to be *optimal* if for each state s there exists a value $\mathrm{val}(s) \in \mathbb{R}$ (the value of s) such that

$$\mathrm{E}_s^{\sigma^\sharp,\tau}[u] \geq \mathrm{val}(s) \geq \mathrm{E}_s^{\sigma,\tau^\sharp}[u]$$

for all strategies σ and τ of players Max and Min.

Martin's theorem [9] guarantees that every state has a value. However it does not guarantee the existence of optimal strategies.

3 Playing without players – 0-player games

Let $\mathcal{A}$ be an arena such that each state has only one available action. Then each player has only one possible trivial strategy consisting in choosing at each state the unique available action. Since the players have no decision to take we can as well forget them, once the game starts the actions can be executed automatically.

The resulting process $\{S_i; i \in \mathbb{N}\}$ is a (homogeneous) Markov process with states $\mathbf{S}$ and with one step transition probabilities $p : \mathbf{S} \times \mathbf{S} \to [0, 1]$ such that, for all states $x, x' \in \mathbf{S}$, $p(x, x') = \delta(a_x, x')$, where a_x is the unique action available at x. Then we have $\mathrm{P}(S_n = x_n | S_1 = x_1, \dots, S_{n-1} = x_{n-1}) = p(x_{n-1}, x_n)$ for all n.

Since we have a natural one to one correspondence between actions and states not only the process $\{A_i; i \in \mathbb{N}\}$ determines $\{S_i; i \in \mathbb{N}\}$ but also, conversely, $\{S_i; i \in \mathbb{N}\}$ determines $\{A_i; i \in \mathbb{N}\}$.

Let us recall some basic notions from the theory of Markov chains [8].

A state s of a Markov chain is said to be *transient* if the probability to return to s (for the chain starting at s) is strictly smaller than 1.

A set C of states of a Markov chain is *closed* if, for each $s \in C$ and each $t \notin C$, $p(s, t) = 0$.

A set C of states of a Markov chain is *irreducible* if for any states $s, t \in C$ there is a positive probability to enter t for a chain starting at s and vice versa.

The set of states of each Markov chain can be decomposed as $T \cup C_1 \cup \ldots \cup C_k$, where T are transient states and C_i are closed irreducible sets.

A Markov chain containing no transient states and one closed irreducible set is called irreducible.

Each finite state Markov chain enters almost surely, after a finite number steps, some closed irreducible component. Thus if the payoff is tail measurable then it is determined by its value in each such component.

As a rather straightforward consequence of the Kolmogorov $0 - 1$ law we obtain

▶ **Theorem 1.** *Let $u \in \mathcal{T}_b$ be a tail measurable payoff.*

Then for each irreducible Markov process with a finite state set $\mathbf{S}$ and action set $\mathbf{A}$ there exists a constant c such that, $\mathrm{P}_\mu\{w \in \mathbf{A}^\omega \mid u(w) = c\} = \mathrm{P}_\mu\{u = c\} = 1$, where P_μ is the measure on $(\mathbf{A}^\omega, \mathcal{B}(\mathbf{A}^\omega))$ induced by the Markov process with the initial state distribution μ.

Clearly Theorem 1 implies that a tail measurable payoff is almost surely constant in each closed irreducible component of a finite Markov chain.

4 Multi-armed bandits

A multi-armed bandit is just a finite sequence of Markov chains (or equivalently 0-player games) $\mathbb{B} = (\mathcal{B}_1, \dots, \mathcal{B}_n)$. Each $\mathcal{B}_i$ is an arm of $\mathbb{B}$. We assume that each arm $\mathcal{B}_i$ is in some state s_i, thus the state of $\mathbb{B}$ is the vector $(s_1, \dots, s_n)$, where s_i is the state of the ith arm.

Player Max plays an infinite game on $\mathbb{B}$. Let $(s_1, \dots, s_n)$ be the state of $\mathbb{B}$. Player Max chooses one of the arms i, the nature executes the unique action available at s_i, $\mathcal{B}_i$ enters a new state s_i', and $(s_1, \dots, s_i', \dots, s_n)$ becomes the new global state of $\mathbb{B}$.

A payoff mapping is defined on the set of infinite sequence of actions of $\mathbb{B}$ and, as usually, player Max wants to maximize the expected payoff. A multi-armed bandit game is a pair $(\mathbb{B}, u)$ consisting of a multi-armed bandit and a payoff mapping. Thus a multi-armed bandit game is just a special type of a one-player stochastic game.

We say the multi-armed bandit is irreducible if each $\mathcal{B}_i$ is an irreducible Markov chain.

▶ **Definition 2.** A strategy of player Max in an irreducible multi-armed bandit is said to be trivial if at each step Max chooses the same arm i.

Note that each trivial strategy is deterministic and memoryless but the triviality condition is stronger, if $\mathbb{B}$ is composed of more than one Markov chain then there are many deterministic memoryless strategies that are not trivial in the sense of Definition 2.

It is easy to see that a multi-armed irreducible bandit with the mean payoff or with the parity payoff has optimal trivial strategies. The same holds for priority mean-payoff.

In general the question if there exists a trivial optimal strategy for a multi-armed bandit game is easier to handle than the question if there exists an optimal memoryless deterministic strategy for the corresponding one-player stochastic game thus it is interesting to note that the last problem can be reduced to the former one.

5 Optimal strategies for one-player games

In this section we consider general one-player games with a tail measurable payoff. We assume that all states of $\mathcal{A} = (\mathbf{S}, \mathbf{A}, \mathrm{source}, \delta, \mathrm{player})$ are controlled by the same player, without a loss of generality we assume that this is player Max.

We call such an arena a one-player arena. A one-player game (or a Markov decision process) is a game on a one-player arena.

We examine the question when player Max has an optimal deterministic memoryless strategy for a given one-player game $(\mathcal{A}, u)$ with $u \in \mathcal{T}_b$.

It turns out that this question can be reduced to the problem of the existence of trivial optimal strategies for some related irreducible multi-armed bandit games.

An arena $\mathcal{A}_\sharp = (\mathbf{S}_\sharp, \mathbf{A}_\sharp, \mathrm{source}_\sharp, \delta_\sharp, \mathrm{player}_\sharp)$ is a subarena of $\mathcal{A} = (\mathbf{S}, \mathbf{A}, \mathrm{source}, \delta, \mathrm{player})$ if $\mathcal{A}_\sharp$ is an arena obtained from $\mathcal{A}$ by removing some states and actions. Note that the requirement that $\mathcal{A}_\sharp$ be an arena means that each state of $\mathcal{A}_\sharp$ retains at least one available action.

We say that a multi-armed bandit $\mathbb{B} = (\mathcal{B}_1, \ldots, \mathcal{B}_n)$ is embeddable into an arena $\mathcal{A}$ if each $\mathcal{B}_i$ is a subarena of $\mathcal{A}$. Note that we allow the same chain to be used several times in $\mathbb{B}$, i.e. the chains $\mathcal{B}_i$ and $\mathcal{B}_j$ can be equal even if $i \neq j$. This implies that each finite arena $\mathcal{A}$ has an infinite number of embeddable multi-armed bandits.

The following theorem reduces the question about the existence of optimal memoryless deterministic strategies in one-player games to a question about optimal trivial strategies in related multi-armed bandit games:

▶ **Theorem 3.** *Let $(\mathcal{A}, u)$ be a one-player game with $u \in \mathcal{T}_b$.*

If for each irreducible multi-armed bandit $\mathbb{B}$ embeddable in $\mathcal{A}$ there exists an optimal trivial strategy in the game $(\mathbb{B}, u)$ then player Max *has an optimal memoryless deterministic strategy in $(\mathcal{A}, u)$.*

In particular we can immediately deduce that one-player parity games, mean-payoff games and priority mean-payoff games have optimal memoryless deterministic strategies.

However the real value of Theorem 3 is not in recovering old results, I hope that it will prove to be useful for establishing the existence of optimal memoryless deterministic strategies for new games.

6 From one-player games to two-player games

Let $\mathcal{A} = (\mathbf{S}, \mathbf{A}, \mathrm{source}, \delta, \mathrm{player})$ and $\mathcal{A}_\sharp = (\mathbf{S}_\sharp, \mathbf{A}_\sharp, \mathrm{source}_\sharp, \delta_\sharp, \mathrm{player}_\sharp)$ be two arenas. A morphism from $\mathcal{A}_\sharp$ to $\mathcal{A}$ is a pair (f, g) of mappings $f : \mathbf{S}_\sharp \to \mathbf{S}$ and $g : \mathbf{A}_\sharp \to \mathbf{A}$ such that

- for each $s_\sharp \in \mathbf{S}_\sharp$, $\mathrm{player}_\sharp(s_\sharp) = \mathrm{player}(f(s_\sharp))$, i.e. f preserves the controlling player,
- for each $a_\sharp \in \mathbf{A}_\sharp$, $f(\mathrm{source}_\sharp(a_\sharp)) = \mathrm{source}(g(a_\sharp))$, i.e. the source of each action is preserved,
- for each $s_\sharp \in \mathbf{S}_\sharp$, for $a_\sharp, b_\sharp \in \mathbf{S}_\sharp(s_\sharp)$, if $a_\sharp \neq b_\sharp$ then $g(a_\sharp) \neq g(b_\sharp)$, i.e. g is locally surjective (but actions with different sources can be mapped to the same action),

- (f, g) preserves (positive) transition probabilities, for all $s_\sharp \in \mathbf{S}_\sharp$ and $a_\sharp \in \mathbf{A}_\sharp$, if $\delta_\sharp(a_\sharp, s_\sharp) > 0$ then $\delta(g(a_\sharp), f(s_\sharp)) = \delta_\sharp(a_\sharp, s_\sharp)$.

The degree of the morphism (f, g) is defined as $\max_{s \in \mathbf{S}} |f^{-1}(s)|$.

Let $(\mathcal{A}, u)$ be a game and (f, g) a morphism from an arena $\mathcal{A}_\sharp$ to $\mathcal{A}$. The lifting of u is the payoff mapping $u_\sharp : (\mathbf{A}_\sharp)^\omega \to (-\infty, \infty]$ such that, for $w = a_1 a_2 a_3 \ldots \in (\mathbf{A}_\sharp)^\omega$, $u_\sharp(w) = u(g(w))$, where $g(w) = g(a_1)g(a_2)g(a_3)\ldots$. The game $(\mathcal{A}_\sharp, u_\sharp)$ will be called the lifting of $(\mathcal{A}, u)$ through the morphism (f, g).

In this section we adopt a slightly extended notion of a one-player arena. We say that $\mathcal{A}$ is a one-player arena controlled by player Max if for each state s controlled by player Min the set $\mathbf{A}(s)$ of available actions contains only one element.

Since for states s with one available action it does not matter who controls s this modified notion of a one-player arena is essentially equivalent to the one used in the previous section.

The following is an enhanced version of the main result of [6]:

▶ **Theorem 4.** *Let $\Gamma = (\mathcal{A}, u)$ be a two-player game. Suppose that for each morphism (f, g) of degree at most 2 from a one-player arena $\mathcal{A}_\sharp$ to $\mathcal{A}$ the player controlling $\mathcal{A}_\sharp$ has an optimal deterministic memoryless strategy in the corresponding lifted one-player game $\Gamma_\sharp = (\mathcal{A}_\sharp, u_\sharp)$. Then both players have optimal deterministic memoryless strategies in Γ.*

Note that for each arena $\mathcal{A}$ there is only a finite number of morphisms of degree at most 2 into $\mathcal{A}$. Thus Theorem 4 states that to establish the existence of optimal memoryless deterministic strategies in a two-player game it suffices to examine a finite number of one-player games.

Note also that a lifting of a mean-payoff game is again a mean-payoff game, similarly a lifting of a parity game is a parity game, and a lifting of a priority mean-payoff game is a priority mean-payoff game thus Theorem 4 allows to deduce that two-player versions of these games have optimal deterministic memoryless strategies for both players. Again these results are not new and the true value of Theorem 4 resides rather in potential applications to new games.

7 Final remarks

There is a large body of literature devoted to multi-armed bandits but it concerns mainly bandits with discounted payoff. The result announced in Theorem 3 relating Markov decision processes to multi-armed bandits seems to be new. Another sufficient condition for the existence of optimal memoryless deterministic strategies in one-player games with a tail measurable payoff is due to H. Gimbert [2]:

▶ **Theorem 5** (H. Gimbert)**.** *Let u be a tail-measurable payoff. Suppose that for all infinite words (histories) $w, w_1, w_2 \in \mathbf{A}^\omega$ such that w is a shuffle of w_1 and w_2, u satisfies the following inequality*

$$u(w) \leq \max\{u(w_1), u(w_2)\}.$$

Then finite state Markov decision processes (one-player games controlled by Max*) with payoff u have optimal deterministic memoryless strategies.*

In practice it is easier to verify the condition of Theorem 5 than the one stated in Theorem 3. However the condition of Theorem 3 seems to be more robust since there exist one-player games where we can prove the existence of optimal memoryless deterministic strategies by means of Theorem 3 but not by Theorem 5 (at least not directly).

It is an open problem how to extend Theorem 3 to payoffs which are measurable but not tail measurable.

References

1 J. Filar and K. Vrieze. *Competitive Markov Decision Processes.* Springer, 1997.

2 H. Gimbert. Pure stationary optimal strategies in Markov decision processes. In *STACS 2007, 24th Annual Symposium on Theoretical Aspects of Computer Science*, volume 4393 of *LNCS*, pages 200–211. Springer, 2007.

3 H. Gimbert and W. Zielonka. Limits of multi-discounted Markov decision processes. In *22th Annual IEEE Symposium on Logics in Computer Science, LICS 2007*, pages 89–98. IEEE Computer Society, 2007.

4 H. Gimbert and W. Zielonka. Perfect information stochastic priority games. In *ICALP 2007*, volume 4596 of *LNCS*, pages 850–861. Springer, 2007.

5 H. Gimbert and W. Zielonka. Blackwell-optimal strategies in priority mean-payoff games. In *GandALF 2010, First International Symposium on Games, Automata, Logics and Formal Verification*, volume 25 of *Electronic Proceedings in Theoretical Computer Science*, pages 7–21, 2010.

6 H. Gimbert and W. Zielonka. Pure nad stationary optimal strategies in perfect-information stochastic games. Technical report, HAL 00438359, 2010. Available as `http://hal.archives-ouvertes.fr/hal-00438359/en`.

7 E. Grädel, W. Thomas, and T. Wilke, editors. *Automata, Logics, and Infinite Games*, volume 2500 of *LNCS*. Springer, 2002.

8 G. Grimmett and D. Stirzaker. *Probability and Random Processes.* Oxford University Press, 2001.

9 D.A. Martin. The determinacy of Blackwell games. *Journal of Symbolic Logic*, 63(4):1565–1581, 1998.