Randomized Strategies are Useless in Markov Decision Processes
Hugo Gimbert

To cite this version:
Hugo Gimbert. Randomized Strategies are Useless in Markov Decision Processes. 2009. hal-00403463v4

HAL Id: hal-00403463
https://hal.archives-ouvertes.fr/hal-00403463v4
Submitted on 3 Dec 2009

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.
Randomized Strategies are Useless in Markov Decision Processes

Hugo Gimbert

December 3, 2009

We show that in a Markov decision process with arbitrary payoff mapping, restricting the set of behavioral strategies from randomized to deterministic does not influence the value of the game nor the existence of almost-surely or positively winning strategies. As a corollary, we get similar results for Markov decision processes with partial observation.

1 Definitions

We use the following notations throughout the paper. Let $S$ be a countable set. The set of finite (resp. infinite) sequences on $S$ is denoted $S^*$ (resp. $S^\omega$). and $S^\omega$ denotes the set of infinite sequences $u \in S^\mathbb{N}$. A probability distribution on $S$ is a function $\delta : S \rightarrow \mathbb{R}$ such that $\forall s \in S$, $0 \leq \delta(s) \leq 1$ and $\sum_{s \in S} \delta(s) = 1$. The set of probability distributions on $S$ is denoted $\mathcal{D}(S)$.

**Definition 1** (Markov Decision Processes). A Markov decision process $M = (S, A, (A(s))_{s \in S}, p)$ is composed of a countable set of states $S$, a countable set of actions $A$, for each state $s \in S$, a set $A(s) \subseteq A$ of actions available in $s$, and transition probabilities $p : S \times A \rightarrow \mathcal{D}(S)$.

In the sequel, we only consider Markov decision processes with finitely many states and actions.

An infinite history in $M$ is an infinite sequence in $(SA)^\omega$. A finite history in $M$ is a finite sequence in $S(A^S)^*$. The first state of an history is called its source, the last state of a finite history is called its target. A strategy in $A$ is a function $\sigma : S(A^S)^* \rightarrow \mathcal{D}(A)$ such that for any finite history $s_0a_1\cdots s_n$, and every action $a \in A$, $(\sigma(s_0a_1\cdots s_n)(a) > 0) \implies (a \in A(s_n))$. 

1
We are especially interested in strategies of the following kind.

**Definition 2 (Deterministic strategies).** A strategy \( \sigma \) is deterministic if for every finite history \( h \) and action \( a \), \( (\sigma(h)(a) > 0) \iff (\sigma(h)(a) = 1) \).

Given a strategy \( \sigma \) and an initial state \( s \in S \), the set of infinite histories with source \( s \) is naturally equipped with a \( \sigma \)-field and a probability measure denoted \( \mathbb{P}^\sigma_s \). Given a finite history \( h \) and an action \( a \), the set of infinite histories in \( h(AS)^\omega \) and \( ha(SA)^\omega \) are cylinders that we abusively denote \( h \) and \( ha \). The \( \sigma \)-field is the one generated by cylinders and \( \mathbb{P}^\sigma_s \) is the unique probability measure on the set of infinite histories with source \( s \) such that for every finite history \( h \) with target \( t \), for every action \( a \in A \) and for every state \( r \),

\[
\begin{align*}
\mathbb{P}^\sigma_s(ha | h) &= \sigma(h)(a) \quad \text{(1)} \\
\mathbb{P}^\sigma_s(har | ha) &= p(r | t, a) \quad \text{(2)}
\end{align*}
\]

For \( n \in \mathbb{N} \), we denote \( S_n \) and \( A_n \) the random variables \( S_n(s_0a_1s_1\cdots) = s_n \) and \( A_n(s_0a_1s_1\cdots) = a_n \).

Some strategies are better than other ones, this is measured by mean of a payoff function. Every Markov decision process comes with a bounded and measurable function \( f : (SA)^\omega \to \mathbb{R} \), called the payoff function, which associates with each infinite history \( h \) a payoff \( f(h) \).

**Definition 3 (Values and guaranteed values).** Let \( M \) be a Markov decision process with a bounded measurable payoff function \( f : (SA)^\omega \to \mathbb{R} \). The expected payoff associated with an initial state \( s \) and a strategy \( \sigma \) is the expected value of \( f \) under \( \mathbb{P}^\sigma_s \), denoted \( \mathbb{E}^\sigma_s[f] \).

## 2 Randomized strategies are useless

Randomizing his own behaviour is useless when there is no adversary to fool. This is the intuitive interpretation of the following theorem:

**Theorem 4.** Let \( M \) be a Markov decision process with a bounded measurable payoff function \( f : (SA)^\omega \to \mathbb{R} \), \( x \in \mathbb{R} \) and \( s \) a state of \( M \). Suppose that for every deterministic strategy \( \sigma \), \( \mathbb{E}^\sigma_s[f] \leq x \). Then the same holds for every randomized strategy \( \sigma \).
Proof. For simplifying the notations, suppose that for every state $s$ there are only two available actions 0, 1 and for every action $a \in \{0, 1\}$ there are only two successor states $L(s, a)$ and $R(s, a)$ distinct and chosen with equal probability $\frac{1}{2}$.

Let $\sigma$ be a strategy and $s$ an initial state. We define a mapping

$$f_{s, \sigma} : \{L, R\}^\omega \times [0, 1]^\omega \to (SA)^\omega$$

that will be used for proving that $P^\sigma_s$ is a product measure. With every infinite word $u \in \{L, R\}^\omega$ and every sequence of real numbers $x = (x_n)_{n \in \mathbb{N}} \in [0, 1]^\omega$ between 0 and 1 we associate the unique infinite play $f_{s, \sigma}(u, x) \in (SA)^\omega = s_0a_1s_1 \cdots$ such that $s_0 = s$, for every $n \in \mathbb{N}$ if $u_n = L$ then $s_{n+1} = L(s_n, a_{n+1})$ otherwise $s_{n+1} = R(s_n, a_{n+1})$ and for every $n \in \mathbb{N}$, if $\sigma(s_0a_1 \cdots s_n)(0) \geq x_n$ then $a_{n+1} = 0$ otherwise $a_{n+1} = 1$.

We equip $\{L, R\}^\omega$ with the $\sigma$-field generated by cylinders and the natural head/tail probability measure denoted $\mu_1$. We equip $[0, 1]^\omega$ with the $\sigma$-field generated by cylinders $I_0 \times I_1 \times \cdots \times I_n \times [0, 1]^\omega$ where $I_1, I_2, \ldots, I_n$ are intervals of $[0, 1]$, and the associated product of Lebesgue measures denoted $\mu_2$.

Then $P^\sigma_s$ is the image by $f_{s, \sigma}$ of the product of measures $\mu_1$ and $\mu_2$, i.e. for every measurable set of infinite plays $A$,

$$P^\sigma_s(A) = (\mu_1 \times \mu_2)(f_{s, \sigma}^{-1}(A)).$$

(3)

This holds for cylinders hence for every measurable $A$.

Now:

$$E^\sigma_s[f] = \int_{p \in (SA)^\omega} f(p) dP^\sigma_s$$

$$= \int_{(u, x) \in \{L, R\}^\omega \times [0, 1]^\omega} f(f_{s, \sigma}(u, x)) d(\mu_1 \times \mu_2)$$

$$= \int_{x \in [0, 1]^\omega} \left( \int_{u \in \{L, R\}^\omega} f(f_{s, \sigma}(u, x)) d\mu_1 \right) d\mu_2$$

where the first equality is by definition of $E^\sigma_s[f]$, the second equality is a basic property of image measures and the third equality is Fubini’s theorem, that we can apply since $f$ is bounded and the measures are probability measures.

Once $x$ is fixed, the behaviour of strategy $\sigma$ is deterministic. Formally, for every $x \in [0, 1]$ let $\sigma_x$ be the deterministic strategy defined by $\sigma_x(s_0a_1 \cdots s_n) =$
0 if and only if $\sigma(s_0a_1\cdots s_n)(0) \geq x_n$. Then for every $y \in ]0,1[^{\omega}$ and $u \in \{L,R\}^{\omega}$, $f_{\sigma,s}(u,y) = f_{\sigma,s}(u,x)$ hence:

$$E^\sigma_s[f] = \int_{u \in \{L,R\}^{\omega}} f(f_{\sigma,s}(u,x)) d\mu_1,$$

and finally:

$$E^\sigma[f] = \int_{x \in ]0,1[^{\omega}} E^\sigma_s[f] d\mu_2,$$

hence the theorem, since for every $x$, strategy $\sigma_x$ is deterministic.

\[\Box\]

3 Applications

We provide an extension of Theorem 4 to Markov decision processes with partial observation.

A Markov decision process with partial observation is similar to a Markov decision process except every state $s$ is labelled with a color $\text{col}(s)$ and strategies should depend only on the sequence of colors. Formally, a strategy is said to be observational if for every finite plays $s_0\cdots s_n$ and $t_0\cdots t_n$, if $\text{col}(s_0\cdots s_n) = \text{col}(t_0\cdots t_n)$ then $\sigma(s_0\cdots s_n) = \sigma(t_0\cdots t_n)$.

**Corollary 5.** Let $\mathcal{M}$ be a Markov decision process with a bounded measurable payoff function $f : (\mathcal{S}\mathcal{A})^{\omega} \to \mathbb{R}$, $x \in \mathbb{R}$ and $s$ a state of $\mathcal{M}$. Suppose that for every deterministic observational strategy $\sigma$, $E^\sigma_s[f] \leq x$. Then the same holds for every randomized observational strategy $\sigma$.

**Proof.** Fix an initial state $s$. Consider the Markov decision process whose state space is the set of finite sequences $a_0c_0a_1\cdots a_nc_n \in (\mathcal{A}\mathcal{C})^{\omega}$ of colors interleaved with actions. The initial state is the empty sequence. From state $a_0c_0a_1\cdots a_nc_n$, playing action $a$ leads to state $a_0c_0a_1\cdots a_nc_nac$ with probability:

$$P^s_{a}(A_{n+1} = a, \text{col}(S_{n+1}) = c \mid A_0C_0A_1\cdots A_nC_n = a_0c_0a_1\cdots a_nc_n),$$

and the payoff associated with an infinite play is defined by:

$$g(a_0c_0a_1\cdots) = E^\sigma_s[f \mid A_0C_0A_1\cdots = a_0c_0a_1\cdots],$$

4
where in both definitions $\sigma$ is any deterministic strategy such that for every $i \in \mathbb{N}$, $\sigma(c_0 \cdot \cdot \cdot c_i) = a_{i+1}$.

The state space of this new Markov decision process is countable therefore we can apply Theorem 4 to it, which immediately gives us the result.