Strong uniform value in gambling houses and partially observable Markov decision processes

Xavier Venel
(PSE, CES, University Paris 1 Panthéon-Sorbonne)

with Bruno Ziliotto (CNRS, Paris-Dauphine)

PGMO days (13-14 November)
1. Introduction
   - The model
   - Evaluation of the game

2. Results (old and new)

3. Outline of the proof
1. **Introduction**
   - The model
   - Evaluation of the game

2. **Results (old and new)**

3. **Outline of the proof**
Introduction

The model

Outline

1. Introduction
   - The model
     - Evaluation of the game

2. Results (old and new)

3. Outline of the proof
We consider $\Gamma = (K, A, S, q, g)$ a MDP with partial observation (POMDP):

- a finite state space $K$,
- a finite set of actions $A$,
- a finite set of signals $S$,
- a transition $q : K \times A \rightarrow \Delta(K \times S)$,
- a stage payoff $g : K \rightarrow [0, 1]$.
Given $p \in \Delta(K)$, $\Gamma(p)$ is played as following:

- **Stage 0**: a state $k_1$ is chosen along $p$.
- **Stage 1**:
  - the decision maker chooses an action $a_1$,
  - he receives the (unobserved) payoff $g(k_1)$,
  - a couple $(k_2, s_1)$ is chosen according to $q(k_1, a_1)$.
  - $s_1$ is announced to the decision maker.
- **Stage 2**: the decision maker chooses etc ...
An example: $K = \{0^*, 0_1, 0_2, 1^*\}$, $A = \{Blue, Red\}$, $S = \{s_1, s_2\}$
Definition of strategies

Definition

- **A behavior strategy** for the decision-maker is a function
  \[ \sigma : \bigcup_{t \geq 1} (A \times S)^{t-1} \rightarrow \Delta(A). \]
  The set of such strategies is denoted \( \Sigma \).

- **A pure strategy** for the decision-maker is a function
  \[ \sigma : \bigcup_{t \geq 1} (A \times S)^{t-1} \rightarrow A. \]

A pair \((p, \sigma)\) induces a probability measure \( \mathbb{P}_\sigma^p \) on \((K \times A \times S)^{\mathbb{N}^*}\). 


A behavior strategy for the decision-maker is a function
\[ \sigma : \bigcup_{t \geq 1} (A \times S)^{t-1} \rightarrow \Delta(A). \]
The set of such strategies is denoted \( \Sigma \).

A pure strategy for the decision-maker is a function
\[ \sigma : \bigcup_{t \geq 1} (A \times S)^{t-1} \rightarrow A. \]

A pair \((p, \sigma)\) induces a probability measure \( P^\sigma_p \) on \((K \times A \times S)^{\mathbb{N}^*}\).
Introduction
Evaluation of the game

Outline

1. Introduction
   - The model
   - Evaluation of the game

2. Results (old and new)

3. Outline of the proof
Asymptotic approach

Let \( n \in \mathbb{N}^* \). \textit{n-stage decision problem}: \( \Gamma_n^p = (\Sigma, \gamma_n^p) \) is the problem where

\[
\gamma_n^p(\sigma) := \mathbb{E}_\sigma^p \left( \frac{1}{n} \sum_{t=1}^{n} g(k_t, a_t) \right).
\]

We denote by

\[
v_n(p) = \max_{\sigma \in \Sigma} \gamma_n^p(\sigma) = \max_{\sigma \in \Sigma_{\text{pure}}} \gamma_n^p(\sigma).
\]

**Definition**

\( \Gamma \) has an \textbf{asymptotic value} if \( (v_n) \) converges pointwise to some function \( v_\infty : \Delta(K) \rightarrow \mathbb{R} \).
Uniform approach

Definition

The decision problem $\Gamma(p)$ has a uniform value if it has an asymptotic value $v_\infty(p)$, and

$$\sup_{\sigma \in \Sigma} \left( \liminf_{n \to +\infty} E_p^\sigma \left( \frac{1}{n} \sum_{t=1}^{n} g(k_t, a_t) \right) \right) = v_\infty(p).$$  \hfill (1)
Pathwise approach

A third approach is to consider the infinitely repeated POMDP where the payoff of the strategy $\sigma$ is given by

$$\gamma^p_\infty(\sigma) = \mathbb{E}_\sigma^p \left( \lim_{n \to +\infty} \inf \frac{1}{n} \sum_{t=1}^{n} g(k_t, a_t) \right)$$

We denote by

$$w_\infty(p) = \max_{\sigma \in \Sigma} \gamma^p_\infty(\sigma) = \max_{\sigma \in \Sigma_{pure}} \gamma^p_\infty(\sigma)$$

Definition

The decision problem $\Gamma(p)$ has a **strong uniform value** if it has an asymptotic value and $w_\infty(p) = v_\infty(p)$. 
Introduction
Evaluation of the game

Relation between the three notions (1)

Proposition

\[ w_{\infty}(p) \leq \sup_{\sigma \in \Sigma_{\text{pure}}} \left( \liminf_{n \to +\infty} E_{\sigma}^p \left( \frac{1}{n} \sum_{t=1}^{n} g(k_t, a_t) \right) \right), \]

\[ \leq \sup_{\sigma \in \Sigma} \left( \liminf_{n \to +\infty} E_{\sigma}^p \left( \frac{1}{n} \sum_{t=1}^{n} g(k_t, a_t) \right) \right), \]

\[ \leq \liminf_{n \to +\infty} v_n(p). \]

Consequently, if \( \Gamma(p) \) has a strong uniform value, the above inequalities are equalities, and \( \Gamma(p) \) has a uniform value in pure strategies.
Relation between the three notions (2)
Results (old and new)

Outline

1 Introduction
   • The model
   • Evaluation of the game

2 Results (old and new)

3 Outline of the proof
Results (old and new)

Perfect Information

Theorem (Blackwell 1962)
A finite POMDP where the decision maker observes the state has a uniform value $\nu_\infty$. Moreover it can be guarantee by pure strategies that only depends on the current state.

Corollary
Under these assumptions, there exists a strong uniform value.
General case (Old)

Theorem (Rosenberg Solan Vieille, 2002)
Any POMDP has a uniform value in behavior strategies.

Renault (2011) and Renault and Venel (2012) provide alternative proofs but again with behavior strategies.

Two questions:
- Do we need behavior strategies?
- What can we say on the stronger property of strong uniform value?
These two questions have been answered positively in several model: for example

- perfect information, compact metric actions space
  Feinberg (1978).

Rosenberg et al. showed that pure strategies are sufficient if $S$ is a singleton.
General case (New)

Theorem (Venel and Ziliotto)

The POMDP $\Gamma(p_1)$ has a strong uniform value in behavior strategies:
for all $\epsilon > 0$, there exists $\sigma^*$ a behavior strategy such that

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

Corollary

The POMDP $\Gamma(p)$ has a strong uniform value in pure strategies.
Outline

1. Introduction
   - The model
   - Evaluation of the game

2. Results (old and new)

3. Outline of the proof
Auxilliary MDP

Natural state variable: \( p_t = \mathbb{P}(k_t | \mathcal{H}_t) \)

Let \( \tilde{\Gamma} = (X, A, \tilde{q}, \tilde{g}) \) be defined as

- a set of states: \( X = \Delta(K) \)
- a payoff function: \( \tilde{g} : X \times A \rightarrow [0, 1] \)
  \[
  \tilde{g}(p, a) = \sum_k p^k g(k, a).
  \]
- a transition function: \( \tilde{q} : X \times A \rightarrow \Delta_f(X) \)
  \[
  \tilde{q}(p, a) = \sum_{s \in S} q(p, a)(s) \delta_{\hat{q}(p, a|s)},
  \]
  where \( \hat{q}(p, a|s) = \left( \frac{q(p, a)(k, s)}{q(p, a)(s)} \right)_{k \in K} \).
We present here the outline of the proof of a weaker result (intermediate between strong uniform value and uniform value):\

\[ \tilde{\Gamma}(\rho) \text{ has a strong uniform value} \]

It is weaker than the previous theorem since plays in \( \tilde{\Gamma} \) and in \( \Gamma \) are different.
Fix an initial state $p \in \Delta(K)$.

- Define a special distribution $\mu^*$ over $\Delta(K)$: “invariant measure” from occupation measures.
- Prove that “from this distribution”, the pathwise uniform value exist.
- Show that from $p$, one can generate a distribution $\mu_n$ close to $\mu^*$.
- Deduce a regularity property of the payoff on play starting from states in the support of $\mu_n$. 
Fix an initial state $p \in \Delta(K)$.

- Define a special distribution $\mu^*$ over $\Delta(K)$: “invariant measure” from occupation measures.
- Prove that “from this distribution”, the pathwise uniform value exist.
- Show that from $p$, one can generate a distribution $\mu_n$ close to $\mu^*$.
- Deduce a regularity property of the payoff on play starting from states in the support of $\mu_n$. 
Fix an initial state $p \in \Delta(K)$.

- Define a special distribution $\mu^*$ over $\Delta(K)$: “invariant measure” from occupation measures.
- Prove that “from this distribution”, the pathwise uniform value exist.
- Show that from $p$, one can generate a distribution $\mu_n$ close to $\mu^*$.
- Deduce a regularity property of the payoff on play starting from states in the support of $\mu_n$. 
Fix an initial state $p \in \Delta(K)$.

- Define a special distribution $\mu^*$ over $\Delta(K)$: “invariant measure” from occupation measures.
- Prove that “from this distribution”, the pathwise uniform value exist.
- Show that from $p$, one can generate a distribution $\mu_n$ close to $\mu^*$.
- Deduce a regularity property of the payoff on play starting from states in the support of $\mu_n$. 
Lemma 1

Let \( p_1 \in \Delta(K) \). There exists a distribution \( \mu^* \in \Delta(\Delta(K)) \) and a stationary strategy \( \sigma^* : \Delta(K) \rightarrow \Delta(A) \) such that

- \( \mu^* \) is invariant if playing \( \sigma^* \),
- For every \( \varepsilon > 0 \), there exists a strategy \( \sigma \) and \( n \) such that
  \[
  d_{KR} \left( \frac{1}{n} \sum_{t=1}^{n} z_t(p_1, \sigma), \mu^* \right) \leq \varepsilon,
  \]
  where \( z_t(p_1, \sigma) \) is the distribution over belief at step \( t \).
- \( g(\mu^*) = v_\infty(\mu^*) = v_\infty(p_1) \).
Lemma 2

There exists $B \subset \Delta(K)$ such that

- $\mu^*(B) = 1$,
- for all $p \in B$,

$$\mathbb{E}_{\sigma^*}^p \left( \lim \inf \frac{1}{n} \sum_{t=1}^{n} g(k_t, a_t) \right) = v_\infty(p). \quad P_{\sigma^*}^p - a.s..$$

- Define a Markov chain $\mathcal{M}$ on $K \times A \times \Delta(K)$ (state, action, belief).
- Apply Birkhoff’s ergodic theorem.
Outline of the proof

Proof of Lemma 2

- There exists \( \nu^* \) an invariant probability distribution for \( \mathcal{M} \) (defined from \( \mu^* \)).
- There exists \( B_0 \subset K \times A \times \Delta(K) \) and a function \( w \) such that
  - \( \nu^*(B_0) = 1 \),
  - for all \((k, a, p) \in B_0\), we have
    \[
    \frac{1}{n} \sum_{t=1}^{n} g(k_t, a_t) \xrightarrow{n \to +\infty} w(k, a, p) \quad P_{\sigma^*}^{k,a,p} - \text{almost surely}
    \]
  - \( w(\nu^*) = g(\nu^*) = v_\infty(\mu^*) \).

For almost all \( p \in B \), \( w(p) = E_p(w) = v_\infty(p) \).
Lemma 3

Let \( p, p' \in \Delta(K) \). For all \( \sigma \in \Sigma \), there exists \( \sigma' \in \Sigma \) such that

\[
E_{\sigma'}^{p'} \left( \liminf_{n \to +\infty} \frac{1}{n} \sum_{t=1}^{n} g(k_t, a_t) \right) \geq \\
E_{\sigma}^{p} \left( \liminf_{n \to +\infty} \frac{1}{n} \sum_{t=1}^{n} g(k_t, a_t) \right) - 2\|p - p'\|_1.
\]

The result is also true if one considers the \( n \)-stage payoff.
Outline of the proof

Conclusion of the proof

- Starting from $p$, by Lemma 1, there exists a strategy $\sigma$ and $t_0 \in \mathbb{N}^\ast$ such that with high probability, $(k_{t_0}, a_{t_0}, p_{t_0})$ is close to $B_0$ and $\mathbb{E}_\sigma(v_\infty(p_{t_0}))$ is close to $v_\infty(p_1)$.

- Lemma 2 says that for all $p \in B$, the average-payoff along each play in $\tilde{\Gamma}(p)$ converges.

- Lemma 2 and Lemma 3 prove that the average-payoff along each play in $\tilde{\Gamma}(p_{t_0})$ almost-converges.

Conclusion: $\tilde{\Gamma}(p_1)$ has a pathwise uniform value.
Conclusions:
- No need for randomization when 1 player.
- More general framework in the paper: gambling house with uniformly equicontinuous value functions that we apply/adapt then to
  - 1-Lipschitz gambling houses,
  - MDP with compact state space,
  - POMDP with finite sets.

Further research:
- What can we say in a two-player zero-sum game with one controller and the players playing one after the other?
- What level of generality?
- Can we say something more on the $\varepsilon$-optimal strategies?
Thanks