Long-term Values in Partial Observation Markov Decision Processes.

Xavier Venel
(LUISS Guido Carli)

Current Trends in Graph and Stochastic Games
(7-8 April 2022)

Outline

Model

We consider $\Gamma = (K, A, S, q, r)$ a Partial Observation Markov Decision Problem:

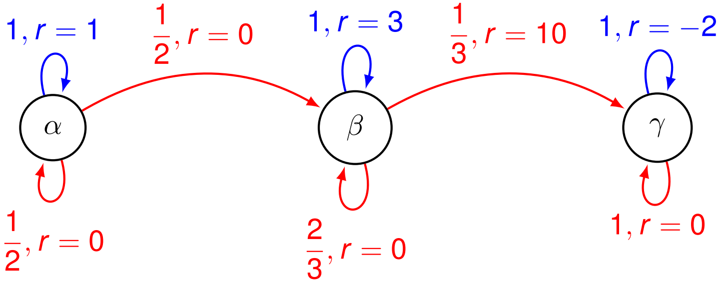
- a finite state space K,
- a finite set of actions A,
- a finite set of signals S,
- a transition $q : K \times A \to \Delta(K \times S)$,
- a stage payoff $r : K \times A \to [0, 1]$.

How is the POMDP played?

Given $p \in \Delta(K)$, $\Gamma(p)$ is played as following:

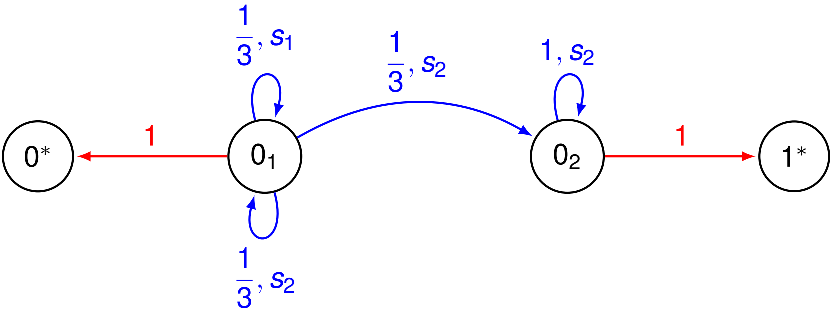
- Stage 0: a state k_1 is chosen along p and nothing is told to the Decision Maker (DM).
- Stage 1:
  - DM chooses an action a_1,
  - He receives the (unobserved) payoff $r(k_1, a_1)$,
  - (k_2, s_1) is chosen according to $q(k_1, a_1)$,
  - s_1 is announced to the DM.
- Stage 2: the DM chooses etc ...

An example: $K = \{\alpha, \beta, \gamma\}$, $A = \{Blue, Red\}$, $S = K$



We assume that with probability one, the state is equal to the signal. Hence, the decision maker knows the state.

An example: $K = \{0^*, 0_1, 0_2, 1^*\}$, $A = \{Blue, Red\}$, $S = \{s_1, s_2\}$



Payoff only depends on the current state and his equal to the "name" of the state.

Definition of strategies

Definition

- A behavioral strategy for the decision-maker is a function $\sigma : \bigcup_{t\geq 1}(A\times S)^{t-1} \to \Delta(A)$.
  The set of such strategies is denoted Σ.
- A pure strategy for the decision-maker is a function $\sigma : \bigcup_{t\geq 1}(A\times S)^{t-1} \to A$.

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

Definition of strategies

Definition

- A behavioral strategy for the decision-maker is a function $\sigma : \bigcup_{t \geq 1} (A \times S)^{t-1} \to \Delta(A)$.
  The set of such strategies is denoted Σ.
- A pure strategy for the decision-maker is a function $\sigma : \bigcup_{t \geq 1} (A \times S)^{t-1} \to A$.

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

Outline

How to aggregate this stage payoffs?

There are many possibilities that differ in several ways:

- event happening in finite time have a positive weight or not,
- the relative weight of each stage is independent of the play or not,
- averaging or not,

Not covered in this talk: Parity games, Buchi game...

Different criteria

- Finite game payoff:

$$\gamma_n(p,\sigma) = \mathbb{E}_{p,\sigma}\left(\frac{1}{n}\sum_{t=1}^{n} r(k_t, a_t)\right)$$

- Discounted payoff:

$$\gamma_\lambda(p,\sigma) = \mathbb{E}_{p,\sigma}\left(\lambda \sum_{t=1}^{+\infty} (1-\lambda)^{t-1} r(k_t, a_t)\right)$$

- A constant weighted θ-evaluation for $\theta \in \Delta(\mathbb{N}^*)$

$$\gamma_\theta(p,\sigma) = \mathbb{E}_{p,\sigma}\left(\sum_{t=1}^{+\infty} \theta_t r(k_t, a_t)\right)$$

Different criteria

- Finite game payoff:

$$\gamma_n(p,\sigma) = \mathbb{E}_{p,\sigma}\left(\frac{1}{n}\sum_{t=1}^{n} r(k_t, a_t)\right)$$

- Discounted payoff:

$$\gamma_\lambda(p,\sigma) = \mathbb{E}_{p,\sigma}\left(\lambda\sum_{t=1}^{+\infty}(1-\lambda)^{t-1} r(k_t, a_t)\right)$$

- A constant weighted θ-evaluation for $\theta \in \Delta(\mathbb{N}^*)$

$$\gamma_\theta(p,\sigma) = \mathbb{E}_{p,\sigma}\left(\sum_{t=1}^{+\infty}\theta_t r(k_t, a_t)\right)$$

- Put a strictly positive weight on what happens in finite time.
- Independent of the play.

Different criteria

- Uniform approach-payoff:

$$\gamma_u(p,\sigma) = \liminf_{n\to+\infty} \mathbb{E}_{p,\sigma}\left(\frac{1}{n}\sum_{t=1}^{n} r(k_t, a_t)\right)$$

- lim inf-payoff:

$$\underline{\gamma}(p,\sigma) = \mathbb{E}_{p,\sigma}\left(\liminf_{n\to+\infty}\left(\frac{1}{n}\sum_{t=1}^{n} r(k_t, a_t)\right)\right)$$

- lim sup-payoff:

$$\overline{\gamma}(p,\sigma) = \mathbb{E}_{p,\sigma}\left(\limsup_{n\to+\infty}\left(\frac{1}{n}\sum_{t=1}^{n} r(k_t, a_t)\right)\right)$$

Different criteria

- Uniform approach-payoff:

$$\gamma_u(p,\sigma) = \liminf_{n\to+\infty} \mathbb{E}_{p,\sigma}\left(\frac{1}{n}\sum_{t=1}^{n} r(k_t,a_t)\right)$$

- $\liminf$-payoff:

$$\underline{\underline{\gamma}}(p,\sigma) = \mathbb{E}_{p,\sigma}\left(\liminf_{n\to+\infty}\left(\frac{1}{n}\sum_{t=1}^{n} r(k_t,a_t)\right)\right)$$

- $\limsup$-payoff:

$$\overline{\overline{\gamma}}(p,\sigma) = \mathbb{E}_{p,\sigma}\left(\limsup_{n\to+\infty}\left(\frac{1}{n}\sum_{t=1}^{n} r(k_t,a_t)\right)\right)$$

- put a null weight on what happens in finite time.
- relative weights depend on the play.

Different criteria

Definition

An evaluation is a sequence of functions $\theta = (\theta_t)_{t\geq 1}$ from $(K \times S \times A)^{\infty}$ to [0,1]. It is

- history-dependent if θ_m is measurable with respect to the observed past before stage m,
- normalized if for every infinite history, the weights sum to 1.

One defines the θ-evaluation for θ an evaluation by

$$\gamma_\theta(p,\sigma) = \mathbb{E}_{p,\sigma}\left(\sum_{t=1}^{+\infty} \theta_t r(k_t, a_t)\right)$$

Different criteria

Definition

An evaluation is a sequence of functions $\theta = (\theta_t)_{t \geq 1}$ from $(K \times S \times A)^\infty$ to [0,1]. It is

- history-dependent if θ_m is measurable with respect to the observed past before stage m,
- normalized if for every infinite history, the weights sum to 1.

One defines the θ-evaluation for θ an evaluation by

$$\gamma_\theta(p, \sigma) = \mathbb{E}_{p,\sigma}\left(\sum_{t=1}^{+\infty} \theta_t r(k_t, a_t)\right)$$

- put a positive weight on what happens in finite time.
- relative weights depend on the play.

Value

Definition

For every evaluation γ and initial probability distribution, we denote by

$$v(p) = \max_{\sigma \in \Sigma} \gamma(p, \sigma).$$

What are the links between all these values ?

Outline

Inequalities 101

- For every infinite play,

$$\liminf_{n\to+\infty}\left(\frac{1}{n}\sum_{t=1}^{n} r(k_t,a_t)\right) \le \limsup_{n\to+\infty}\left(\frac{1}{n}\sum_{t=1}^{n} r(k_t,a_t)\right).$$

So

$$\underline{\underline{v}} \le \overline{\overline{v}}$$

- By Fatou's lemma for a given strategy

$$\mathbb{E}_{p,\sigma}\left(\liminf_{n\to+\infty}\left(\frac{1}{n}\sum_{t=1}^{n} r(k_t,a_t)\right)\right) \le \liminf_{n\to+\infty}\mathbb{E}_{p,\sigma}\left(\frac{1}{n}\sum_{t=1}^{n} r(k_t,a_t)\right)$$

so

$$\underline{\underline{v}} \le v_u.$$

Inequalities 101

- For every infinite play,

$$\liminf_{n\to+\infty}\left(\frac{1}{n}\sum_{t=1}^{n} r(k_t, a_t)\right) \leq \limsup_{n\to+\infty}\left(\frac{1}{n}\sum_{t=1}^{n} r(k_t, a_t)\right).$$

So

$$\underline{\underline{v}} \leq \overline{\overline{v}}$$

- By Fatou's lemma for a given strategy

$$\mathbb{E}_{p,\sigma}\left(\liminf_{n\to+\infty}\left(\frac{1}{n}\sum_{t=1}^{n} r(k_t, a_t)\right)\right) \leq \liminf_{n\to+\infty}\mathbb{E}_{p,\sigma}\left(\frac{1}{n}\sum_{t=1}^{n} r(k_t, a_t)\right)$$

so

$$\underline{\underline{v}} \leq v_u.$$

Inequalities 102

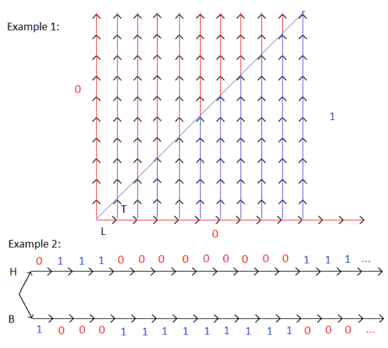
$$v_u \leq \liminf_{n \to +\infty} v_n.$$

Intuition:

- It is easier to guarantee a payoff if the DM can adapt to the length of the game.
- The decision maker maximizes the payoff.
- $\liminf$ is hiding an infimum over the stages.
- $\text{maxmin} \leq \text{minmax}$

Relation between the three notions (countable case)

With a countable set of states, these inequalities can be strict.



What happens when the state space is finite?

Outline

Outline

Limit of finite values and Uniform value

Theorem (Rosenberg-Solan-Vieille 2002)

The POMDP has a uniform value:

- $(v_n)_{n\geq 1}$ converges uniformly to some function v_∞.
- $v_u = v_\infty$.

Remark

- Extended by Renault (2011) to infinite action and signal spaces (with continuity assumptions).
- The proof involves behavorial strategies

The decision maker can play well in long games without knowing the length of the game.

Outline

Liminf value and Uniform value

Theorem (Venel and Ziliotto 2016)

The POMDP has a strong uniform value:

$$\underline{\underline{v}} = v_u = v_\infty.$$

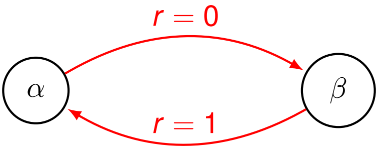
Corollary

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

Even when very pessimistic, the decision maker can still guarantee this value (and without randomizing).

Outline

Uniform value and weighted evaluation: a counterexample



- $K = \{\alpha, \beta\}$, A={Red}.
- Consider the following evaluation

$$\theta^{odd,n} = \left(\frac{1}{2n}, 0, \frac{1}{2n}, 0, \cdots, ..., \frac{1}{2n}, 0, 0, 0, \cdot \right)$$

- Then the value under $\theta^{odd,n}$ is equal to 1 starting from β and 0 from α.

Different from the uniform value (equal to $1/2$).

Uniform value and weighted evaluation: play-independent

Given a constant weighted evaluation $\theta \in \Delta(\mathbb{N}^*)$, we define

$$I(\theta) = |\theta_1| + \sum_{t \geq 1} |\theta_{t+1} - \theta_t|.$$

Theorem (Renault and Venel 2017)

When $I(\theta)$ goes to 0, v_θ also converges to v_u.

Remarks

- If the sequence of weight is non-increasing,

$$I(\theta) = 2\theta_1.$$

- Stronger results in the article: uniform θ-value.

Uniform value and weighted evaluation: history-dependent (1/2)

Given a (not constant) evaluation θ, we define

$$I(p,\theta,\sigma) = \mathbb{E}_{p,\sigma}\left(|\theta_1| + \sum_{t\geq 1} |\theta_{t+1} - \theta_t|\right)$$

and $I(\theta, p)$ as the supremum over all possible strategies.

Definition

The POMDP has a weighted value if for all $\varepsilon > 0$, there exists $\alpha > 0$ and σ^* a strategy such that for all normalized history-dependent evaluation θ,

$$I(\theta,p) \leq \alpha \Rightarrow \gamma_\theta(p,\sigma^*) \geq v_u(p) - \varepsilon.$$

Uniform value and weighted evaluation: history-dependent (2/2)

Theorem (Venel and Ziliotto 2020)

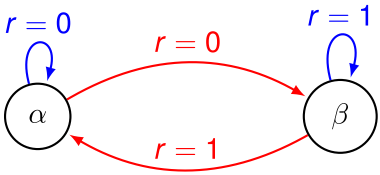
- Any finite POMDP has a weighted value.
- Moreover, it can be guarantee with a pure strategy with finite memory.

Outline of the proof

Outline

Optimistic decision maker: a counterexample

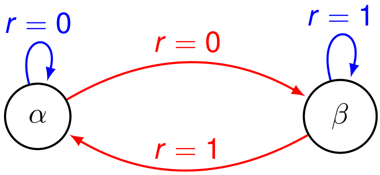
What happens if the decision maker is optimistic?



- POMDP:$K = \{\alpha, \beta\}$, A={Red,Blue}; No signal.
- The decision maker can guarantee ? .

Optimistic decision maker: a counterexample

What happens if the decision maker is optimistic?



- POMDP:$K = \{\alpha, \beta\}$, A={Red,Blue}; No signal.
- The decision maker can guarantee 1.

Intution

Value of POMDP = Value of MDP

if

the implicit weight in the evaluation only depends on what the player observes.

Not the case for the limsup

An intermediate limsup (1/2): Auxiliary MDP

It is classical to associate to a POMDP an auxiliary MDP on the belief space.

In the previous example:

- $X = [0, 1]$ (the probability to be in α),
- The transition is deterministic:

$$\widetilde{q}(p, red) = 1 - p \text{ and } \widetilde{q}(p, blue) = p.$$

- The payoff is the linear extension of r:

$$g(p, red) = g(p, blue) = p.$$

An intermediate limsup (2/2): limsup-belief evaluation

Define the limsup-belief evaluation where we aggregate the payoffs for the same belief

$$\overline{\gamma}(p,\sigma) = \mathbb{E}_{p,\sigma}\left(\limsup_{n\to+\infty}\left(\frac{1}{n}\sum_{t=1}^{n} g(p_t, a_t)\right)\right)$$

Theorem (Venel and Ziliotto 2020)

$$v_u = \overline{v}.$$

Assymetry between $\limsup$ and $\liminf$

- Playing non stationary may lower the payoff for the $\liminf$ since then
$$\mathbb{E}_{p,\sigma} \liminf < \liminf \mathbb{E}_{p,\sigma} .$$
Therefore, the optimality of strategies with finite memory yields equality.

- On the contrary playing non stationary may increase the payoff for the $\limsup$ hence a strictly higher payoff.

- To summarize
$$\underline{\underline{v}} = \underline{v} = \overline{v} < \overline{\overline{v}}.$$

Outline of the proof

Outline

Conclusions :

- Equality between many different notions of values
- Proof highlights links between the weighted average approach and the $\limsup$.

Current research:

- Weighted evaluation can be reinterpreted in terms of a terminal payoff with a stopping clock.
- Investigate what happens with different type of clocks.

Further research:

- What happens for two-player zero-sum game with one controller?
- What can we say in other class of stochastic games?

Thanks

Outline of the proof:lower bound

DM can guarantee v_u (up to ε)

- Chatterjee et al. (2020): $\exists$ a pure ε-optimal strategy with finite memory for the uniform value.
- It reduces the problem to the case without player (Homogeneous Finite Markov chain)
- True for Markov chain: ergodic structure+periodicity of the process+computation.

Outline of the proof:upper bound

DM can not do better

- Consider $(\theta^l)_{l\geq 1}$ such that $I(p,\theta^l) \to 0$.
- Associate to the sequence $v_{\theta^l}(p)$, an invariant distribution μ^* of the POMDP summarizing the payoff.
- Payoff at μ^* is smaller than uniform value at μ^*,
- Since uniform value decreases along play (in a martingale sense), smaller than the uniforrm value at p.

$$\lim_{\ell\to+\infty} v_{\theta^\ell} = g(\mu^*) \leq v_u(\mu^*) \leq v_u(p).$$

Outline of the proof: Lower bound

Play for the liminf evaluation.

Outline of the proof: Upper bound

Can not do more

- Fix (ε, p, σ) and $l \geq 1$. One can define a r.v. η^l such that

$$\mathbb{E}_{p,\sigma}\left(\frac{1}{\eta^l}\sum_{t=1}^{\eta^l} g(x_t, a_t)\right) \geq \mathbb{E}_{p,\sigma}\left(\limsup_{n\to+\infty}\left(\frac{1}{n}\sum_{t=1}^{n} g(x_t, a_t)\right)\right) - \frac{1}{l}$$

- **1st Problem**: Not measurable w.r.t the past **but** one can replace by

$$\hat{\theta}^l_n = \mathbb{E}_\sigma\left(\frac{1}{\eta^l}\mathbb{1}_{n\leq\eta^l}|\mathcal{F}_n\right).$$

  then

$$\gamma_{\hat{\theta}^l}(p,\sigma) \geq \overline{\gamma}(p,\sigma) - \frac{1}{l}.$$

- **2nd Problem**: Not normalized **but** almost.
- One can apply the upper bound for weighted evaluation.