Journées SMAI MODE 2018

# Markov Decision Processes with long duration

**Xavier VENEL**
Université Paris 1 Panthéon-Sorbonne, Paris School of Economics

**Bruno ZILIOTTO**
CNRS, Université de Dauphine

**Résumé.** In a Markov Decision Process (MDP), at each stage, knowing the current state, the decision-maker chooses an action, and receives a reward depending on the current state of the world. Then a new state is randomly drawn from a distribution depending on the action and on the past state. Many optimal payoffs concepts have been introduced to analyze the strategic aspects of MDPs with long duration: asymptotic value, uniform value, liminf average payoff criterion, limsup average payoff criterion... We provide sufficient conditions under which these concepts coincide completing previous results of Renault and Venel [4] and Venel and Zilioto [6].

**Mots-clefs :** MDP, Gambling house, Uniform value, Liminf value, Limsup value

The standard model of Markov Decision Process (or Controlled Markov chain) was introduced by Bellman [2] and has been extensively studied since then. In this model, at the beginning of every stage, a decision-maker perfectly observes the current state, and chooses an action accordingly, possibly randomly. The current state and the selected action determine a stage payoff and the law of the next state.

There are two standard ways to aggregate the stream of payoffs. Given a strictly positive integer $n$, in the $n$-stage MDP, the total payoff is the Cesaro mean $n^{-1}\sum_{m=1}^{n} g_m$, where $g_m$ is the payoff at stage $m$. Given $\lambda \in (0,1]$, in the $\lambda$-discounted MDP, the total payoff is the $\lambda$-discounted sum $\lambda\sum_{m\geq 1}(1-\lambda)^{m-1}g_m$. The maximum expected payoff that the decision-maker can obtain in the $n$-stage problem (resp. $\lambda$-discounted problem) is denoted by $v_n$ (resp. $v_\lambda$).

A huge part of the literature investigates *long-term* MDPs, that is, MDPs which are repeated a large number of times. It can be done following several approaches. The first approach is to determine whether $(v_n)$ and $(v_\lambda)$ converge when $n$ goes to infinity and $\lambda$ goes to 0, and whether the two limits coincide. When this is the case, the MDP is said to have an *asymptotic value*.

A second approach is to define the payoff in the infinite problem as the inferior limit of the expectation of $n^{-1}\sum_{m=1}^{n} g_m$. In the literature, this is referred as the *long-run average payoff criterion* (AP criterion, see Arapostathis et al. [1] for a review of the subject). When the asymptotic value exists and coincides with the value in behavior (resp. pure) strategies of the infinite problem, the MDP is said to have a *uniform value* in behavior (resp. pure) strategies.

A third approach is to define the payoff in the infinite problem as being the expectation of $\liminf_{n\to+\infty} n^{-1}\sum_{m=1}^{n} g_m$ as studied in Gillette [3]. The decision maker is particularly pessimistic when he aggregates the payoff. Opposite to this case, one can look at the game where the payoff in the infinite problem is the expectation of $\limsup_{n\to+\infty} n^{-1}\sum_{m=1}^{n} g_m$ and more generally have evaluation which depends on the strategy of the decision maker.

Renault [4] showed the existence of the uniform value for MDP with compact set of states. Renault and Venel [5] showed that the uniform value in this case has more properties since the same strategy is good for not only Cesaro Means and Abel means but for other mean evaluation. In a previous work, Venel and Ziliotto [6] then showed that the liminf value also coincides with the uniform value when the set of states and the set of actions are compact. In this paper, we extend the result to more general evaluation and in particular to the limsup evaluation proving that all the notions coincides and yield the uniform value. In particular, even if the decision maker is particularly optimist in the way he is aggregating the payoff, he can not guarantee more than the uniform value.

## Références


[1] A. Arapostathis, V. Borkar, E. Fernandez-Gaucherand, M. Gosh and S. Marcus.. *Discrete-time controlled Markov processes with the average cost criterion: a survey*. SIAM Journal on Control and Optimization, 31(2):282-344, 1993.

[2] R. Bellman. *A Markovian decision process*. Technical report, DTIC DOcument, 1957.

[3] D. Gillette. *Stochastic games with zero stop probabilities*. Contributions to the Theory f Games, 3:179-187,1957.

[4] J. Renault. *Uniform value in dynamic programming*. Journal of European Mathematical Society,13(2):309-330, 2011.

[5] J. Renault and X. Venel. *Long-Term Values in Markov Decision Processes and Repeated Games, and a New Distance for Probability Spaces* . Mathematics of Operations Research, 42(2):349-376, 2017.

[6] X. Venel, B. Ziliotto. *Strong Uniform Value in Gambling Houses and Partially Observable Markov Decision Processes*. SIAM Journal on Control and Optimization, 54(4):1983-2008, 2016.