Rolling Horizon Procedure on Controlled Semi-Markov Models. The Discounted Case.

Eugenio Della Vecchia [*], Silvia Di Marco [*], and Alain Jean-Marie [◇]

[*] CONICET - UNR, Argentina
{eugenio,dimarco}@fceia.unr.edu.ar
[◇] INRIA - LIRMM, France. ajm@lirmm.fr

Abstract. We study the behavior of the rolling horizon procedure for semi-Markov decision processes, with infinite-horizon discounted reward, when the state space is a Borel set and the action spaces are considered compact. We prove the convergence of the rewards produced by the rolling horizon policies to the optimal reward function, when the horizon length tends to infinity, under different assumptions on the instantaneous reward function. The approach is based on extensions of the results obtained in [7] for the discrete-time Markov decision process case and in [3] for the case of discrete-time Markov games. Finally, we also analyse the performance of an approximate rolling horizon procedure.

Keywords: Semi-Markov decision processes - Rolling horizon - Discounted criterion.

1 Introduction

In this work we deal with semi-Markov decision processes with the expected total discounted reward as the performance criterion. Semi-Markov decision processes (**SMDP**) generalize Markov decision processes (**MDP**) by allowing the decision maker to choose actions whenever the system state changes, modeling the system evolution in continuous-time and allowing the time spent in a particular state to follow an arbitrary probability distribution. The system state may change several times between decision epochs but only the state at a decision epoch is relevant to the decision maker.

Semi-Markov decision processes with discounted reward are analyzed in [1], [10], [11]. In particular, continuous-time controlled Markov chains are treated in [4]. Zero sum semi-Markov games with discounted payoff are studied in [9].

Rolling horizon (**RH**) is an usual procedure for making decisions in many infinite stage decision problems. It is based on choosing the best most immediate action based on the knowledge of the information of the problem just for a certain number of periods in the future. One design issue of the controller will be then to determine how many periods in the future must be taken into account, in order to make the optimal immediate decision [12]. **RH** strategies are largely used in several areas: we can mention here production control problems, stabilization of control systems, and macroplanning problems. The study of this and other applications can be found in [8].

In [6] and [7], the accuracy of this procedure for discrete-time **MDP**s with bounded and unbounded rewards functions respectively can be found. Similar results for discrete-time zero-sum Markov games with finite spaces are obtained in [3].

In [5] we show the uniform geometrical convergence of the values obtained by **RH** to the optimal one, in Semi-Markov games with discounted payoff, when the reward function is *bounded*. Similar results to those described there, could be obtained if we deal just with the bounded reward **SMDP**s case.

The objective of this work is to study the accuracy of the **RH** method applied to **SMDP**'s with total discounted reward criterion, when the state space is assumed to be Borel, and the action space compact. As a particular case, all the results obtained apply to continuous-time **MDP**s.

This paper is organized as follows. In Section 2, we present the model, the notations and we state the assumptions on the data of the problem. In Section 3 we present the performance criterion and the dynamic operator for this case, mentioning the results on the optimality equation and the recursion scheme associated. Section 4 contains our contributions about the convergence of **RH** policies rewards to the optimal reward. The approach in this section is based on [7] where the discrete-time case is treated. We include also the study of an approximate rolling horizon procedure. Finally, Section 5 is devoted to the concluding remarks.

2 Preliminaries and Notations

We consider a semi-Markov decision model of the form

$$M := (S, A, \{A_s : s \in S\}, Q, F, r, \alpha)$$

where S is the state space and A is the action space. For every $s \in S$, we define the set A_s as the set of actions available in state s, and, in such a way $A = \bigcup_{s \in S} A_s$. We put $\mathbb{K} = \{(s, a) : s \in S, a \in A_s\}$. The transition law $Q(\cdot|\cdot)$, is a stochastic kernel on S given $\mathbb{K}$, and $F(\cdot|s, a)$ is the distribution function of the holding time in state $s \in S$ when action $a \in A_s$ is chosen. The reward function r is a real-valued function on $\mathbb{K}$ and α is a discount factor.

If at time of the n-th decision epoch, the state of the system is $s_n = s$, and the chosen action is $a_n = a \in A_s$, then the system remains in the state s during a nonnegative random time δ_{n+1} with distribution $F(\cdot|s_n, a_n)$ and an instantaneous reward $r(s_n, a_n)$ is received.

In order to prove some relevant results about this control model, we have to make some extra assumptions on the state and action spaces as well on the reward function.

Assumption 1 *a) The state space S is a Borel subset of a complete and separable metric space.*

b) For each $s \in S$, the set A_s is compact.
c) For each $s \in S$, $r(s,\cdot)$ is upper semicontinuous on A_s.
d) For each $(s,a) \in \mathbb{K}$ and each bounded measurable function v on S, the function $a \mapsto \int v(y)Q(dy|s,a)$ is continuous on A_s.
e) For each $t \geqq 0$, $F(t|\cdot)$ is continuous on $\mathbb{K}$.

We will note, for Borel sets X and Y, with $\mathbb{P}(X)$ to the family of probability measures on X endowed with the weak topology, and with $\mathbb{P}(X|Y)$ to the family of transition probabilities from Y to X.

We define the spaces of admissible histories of the process up to the n-th decision epoch by $H_0 := S$, and $H_n := H_{n-1} \times (\mathbb{K} \times \mathbb{R}^+)$ for $n \in \mathbb{N}$. A typical element of H_n is written as $h_n = (s_0, a_0, \delta_1, ..., s_{n-1}, a_{n-1}\delta_n, s_n)$.

A Markov strategy (or Markov policy) is a sequence $\pi = \{\pi_n\}$ of stochastic kernels $\pi_n \in \mathbb{P}(A|H_n)$ such that $\pi_n(A_{s_n}|h_n) = 1$ for all $h_n \in H_n$ and $n \in \mathbb{N}$. We denote by Π the set of all strategies. A strategy $\pi = \{\pi_n\}$ is called stationary if there exists $f \in \mathbb{P}(A|S)$ such that $f(s) \in \mathbb{P}(A_s)$ and $\pi_n = f$ for all $s \in S$ and $n \in \mathbb{N}$. In this case, we identify π with f, i.e., $\pi = f = \{f, f, ...\}$. We denote by Π_S the set of all stationary strategies.

Observe that the decision epochs are $T_n := T_{n-1} + \delta_n$ for $n \in \mathbb{N}$, and $T_0 = 0$. The random variable $\delta_{n+1} = T_{n+1} - T_n$ is called the sojourn or holding time at stage n.

For each strategy $\pi \in \Pi$, and any initial state s there exist a unique probability measure P_s^π and stochastic processes $\{S_t\}$, $\{A_t\}$ and $\{\delta_t\}$. S_t and A_t represent the state and the action at the n-th decision epoch. $\mathbb{E}_s^\pi$ denotes the expectation operator with respect P_s^π.

We note $\beta(s,a) := \int_0^\infty e^{-\alpha t} F(dt|s,a)$ and $\tau(s,a) = \frac{1-\beta(s,a)}{\alpha}$. Also, for any given function $h : \mathbb{K} \to \mathbb{R}$ and any $\xi \in \mathbb{P}(A_s)$ we will write $h(s,\xi)$ instead of $h(s,\xi(s))$, and it will be

$$h(s,\xi) = \int_{A_s} h(s,a)\xi(da)$$

whenever the integral is well defined. In particular, we apply this notation to the functions $\beta(s,\cdot)$, $\tau(s,\cdot)$, $r(s,\cdot)$, $Q(j|s,\cdot)$.

Remark 1 *This semi-Markov environment covers two important special cases:*

1. *Discrete-time models. In this case $F(\cdot|s,a) = \delta_1(\cdot)$ for all $(s,a) \in \mathbb{K}$. This correspond to the theory of **MDPs**.*
2. *Continuous-time Markov models. This arises if the holding time distributions is exponential: $F(du|s,a) = \beta(s,a)e^{-\beta(s,a)u}du$, where $\beta(s,a)$ is a continuous function from $\mathbb{K}$ into $[0,\infty)$. The process $\{S_t\}$ turns out to be a Markov process when π is a Markov policy, and a time homogeneous Markov process when π is a stationary policy.*

We shall make further assumptions on the distribution probabilities of the holding time and the instantaneous reward function, under which we will work in the future.

Assumption 2 $\rho = \sup_{(s,a)\in\mathbb{K}} \beta(s,a) < 1$.

Proposition 1 *If there exists a pair of positive numbers θ and ϵ such that*

$$F(\theta|s,a) \leqq 1-\epsilon$$

for all $s \in S$ and $a \in A_s$, then Assumption 2 holds with $\rho = 1-\epsilon+\epsilon e^{\alpha\theta}$.

For any nonnegative measurable function v, define a new function Lv by

$$(Lv)(s) = \sup_{a\in A_s} \int v(j)Q(dj|s,a) \tag{1}$$

for $s \in S$. Let $L^n v = L(L^{n-1}v)$ for $n \in \mathbb{N}$, with $L^0 v = v$.

Assumption 3 $R(s) = \sum_{t=0}^{\infty} \rho^t (L^t r_0)(s) < \infty$ *for all* $s \in S$, *where* $r_0(s) = \sup_{a\in A_s} |r(s,a)\tau(s,a)|$.

Assumption 4 *There exist a measurable function $\omega : S \to [1,\infty)$ and a positive constant m such that for all $(s,a) \in \mathbb{K}$,*

1. $|r(s,a)| \leqq m\omega(s)$,
2. $\int \omega(j)Q(dj|s,a) \leqq \omega(s)$.

Remark 2 *If r is a bounded function on $\mathbb{K}$, then by setting $\omega \equiv 1$ and M any bound of r, Assumption 4 is satisfied. In [7], it is shown that Assumption 4 implies Assumption 3.*

3 Performance Criterion and Related Results.

In order to evaluate the performance of policies, we use a total discounted criterion. We assume that the rewards are continuously discounted, with a discount factor α. More precisely let, for $n \geqq 1$, $s \in S$ and $\pi \in \Pi$, the expected n-stage α-discounted reward defined by

$$V_n^\pi(s) := \mathbb{E}_s^\pi \sum_{k=0}^{n-1} e^{-\alpha T_k} r(S_k, A_k) \ ,$$

where $T_0 = 0$ and $T_n = T_{n-1} + \delta_n$. The infinite horizon total expected α-discounted payoff is

$$V^\pi(s) := \mathbb{E}_s^\pi \sum_{k=0}^{\infty} e^{-\alpha T_k} r(S_k, A_k) \ .$$

The objective of the controller is to find (when it exists) a policy that solves, given the current state s: $\pi(s) = \arg\max_\pi V^\pi(s)$. Such a strategy $\pi^* \in \Pi$ is said to be α-optimal, and the function $V^*(s) = \sup_{\pi\in\Pi} V^\pi(s)$ is the optimal value function.

Alternatively, given a policy π, we can write its reward using the variables β and τ and obtain for the finite stage horizon and for the infinite horizon respectively,

$$V^{\pi}(s) = \tau(s,f_0)r(s,f_0) + \mathbb{E}_s \sum_{t=1}^{\infty} \prod_{k=0}^{t-1} \beta(S_k,A_k)\tau(S_k,A_k)r(S_k,A_k) \ , \qquad (2)$$

$$V_n^{\pi}(s) = \tau(s,f_0)r(s,f_0) + \mathbb{E}_s \sum_{t=1}^{n-1} \prod_{k=0}^{t-1} \beta(S_k,A_k)\tau(S_k,A_k)r(S_k,A_k) \ . \qquad (3)$$

Let us note $\mathcal{M}(S)$ the space of measurable functions on S and $\mathcal{M}_+(S)$ the subspace of nonnegative functions of $\mathcal{M}(S)$. Under Assumption 3, define $\mathcal{R} = \{v \in \mathcal{M}(S) : |v(s)| \leqq R(s) \ \text{ for all } s \in S\}$. If $\omega \in \mathcal{M}_+(S)$ is strictly positive, for $v \in \mathcal{M}(S)$ we define the ω-weighted norm $||v||_\omega = \sup_{s\in S} |v(s)|/\omega(s)$ and $\mathcal{M}_\omega(S)$ the linear subspace of $\mathcal{M}(S)$ of the functions with finite ω-weighted norm, which results a Banach space.

Define the operator T on $\mathcal{R}$ or $\mathcal{M}_\omega(S)$ by

$$(Tv)(s) := \sup_{a\in A_s} \left\{ r(s,a)\tau(s,a) + \beta(s,a) \int_S v(z)Q(dz|s,a) \right\} \ . \qquad (4)$$

Observe that, if $v \in \mathcal{R}$, then $|Tv(s)| \leqq r_0(s) + \rho L|v|(s) \leqq r_0(s) + \rho L R(s) \leqq R(s)$, which implies $Tv \in \mathcal{R}$. On the other hand, if $v \in \mathcal{M}_\omega(S)$, for all $(s,a) \in \mathbb{K}$,

$$\int v(j)Q(dj|s,a) = \int \frac{v(j)}{\omega(j)} \omega(j)Q(dj|s,a) \leqq ||v||_\omega \omega(s) \ ,$$

and $||Tv||_\omega \leqq m + \rho||v||_\omega$ which implies that $Tv \in \mathcal{M}_\omega(S)$. The operator T maps therefore $\mathcal{R}$ to itself, and $\mathcal{M}_s$ to itself.

Under Assumptions 1, 2 and 3, the supremum is attained for each $s \in S$. Denoting with $f(s)$ a maximizing action on state s, a well-known result of measurable selections (see for example section 7.5 *Semicontinuous Functions and Borel Measurable Selection* of [2]), let us assure that $f \in \mathcal{M}(S)$.

The next theorem joins results for the finite-stage horizon problem as well as for the infinite-stage horizon problem under the previous assumptions. Particularly for the infinite-stage horizon case, provides a Value Iteration scheme for approximating the optimal value. For its proof we refer to [1].

Theorem 1. *Suppose that Assumptions 1, 2 and 3 hold and that we choose $V_0 \equiv 0$. Then, for $n \in \mathbb{N}$, the function $V_n := TV_{n-1}$ is the optimal value function for the n-stage horizon problem, and the Markovian policy $\{f_0^*, f_1^*, ..., f_n^*\}$ (where the functions f_n^* are the corresponding maximizing functions) is optimal. Besides, for all $s \in S$, $|V^*(s) - V_n(s)| \leqq \rho^n L^n R(s) \leqq \sum_{t=n}^{\infty} \rho^t L^t r_0(s) \to 0$ as $n \to \infty$, and V^* is the unique function in $\mathcal{R}$ satisfying the optimality equation $TV^* = V^*$.*

Moreover, there exists an optimal stationary policy f^* for the infinite horizon case.

In addition, if Assumption 4 holds, T is a contraction mapping on $\mathcal{M}_\omega(S)$ of modulus ρ, and then, for any $V_0 \in \mathcal{M}_\omega(S)$, $||V^* - V_n||_\omega \leqq \rho^n ||V^* - V_0||_\omega \to 0$ as $n \to \infty$. In particular, if $V_0 \equiv 0$, $||V^* - V_n||_\omega \leqq \frac{m\rho^n}{1-\rho}$.

4 Rolling Horizon Procedure

For a wide class of stochastic control problems, obtaining an optimal policy explicitly is a difficult task. This is why practitioners often use instead a heuristic method called the Rolling Horizon procedure (also, Receding Horizon, or Model Predictive Control), which works as follows. To the infinite-stage horizon control problem is associated a finite-stage horizon problem: for a given integer N (the horizon length) and a state s, find:

$$(FHP) \quad \sup_{\pi} \mathbb{E}_s^\pi \sum_{t=0}^{N-1} e^{-\alpha T_k} r(s_t, a_t) \ . \tag{5}$$

Solving this problem results in a sequence of decision rules (i.e. a Markovian policy):

$$\pi_N^* = (f_N, f_{N-1}, \ldots, f_2, f_1) \tag{6}$$

where $f_1(s_{N-1})$ is the best action to be applied at time $t = N - 1$ when only one step remains to reach the horizon, f_2 is the best decision rule to be applied when two steps remain to get the horizon, at time $t = N - 2$, and so on. In particular, $f_N(s)$ is the best decision rule to be applied to the initial state s.

The **RH** method prescribes to repeatedly solve a finite horizon problem, taking the current state as initial state. Then, the procedure offers a control sequence where only the first one of them will be applied.

Specifically, the procedure to construct a **RH** policy is the following one. Fix some integer N.

1. At time t, and for the current state s_t, find the value of $f_N(s_t)$ in the control problem (FHP).
2. Apply $a_t = f_N(s_t)$.
3. Observe the achieved state at time $t + 1$: s_{t+1}.
4. Set $t := t + 1$ and $s_t := s_{t+1}$ and go to step 1.

The **RH** procedure does not specify how to compute the value $f_N(s_t)$. Its efficiency is based on the idea that computing the value $f_N(s_t)$ alone is usually much easier than solving entirely the (FHP), which involves computing the N decision rules in (6). On the other hand, the performance of the resulting policy is not the optimal one, although the intuition is that when N is "large enough", the performance should be close to the optimal. The practical issue is then to choose N so as to obtain a proper compromise between precision and the computational effort needed to obtain $f_N(s_t)$. We address this issue through two formal qualitative and quantitative questions. Let $U_N(s)$ be the performance achieved by the **RH** procedure with horizon length N, starting in state s:

Q1 Under which conditions on the problem is it true that $\lim_{N\to\infty} U_N(s) = V^*(s)$?

Q2 Given a state s and $\epsilon > 0$, is it possible to compute N such that $|U_N(s) - V^*(s)| < \epsilon$?

The next Lemma is a straightforward adaptation of Lemma B1.e of [7].

Lemma 1 $\mathbb{E}_s^\pi v(S_t) \leqq (L^t v)(s)$ *for all* $\pi \in \varPi$, $s \in S$, $v \in \mathcal{M}_+(S)$ *and* $t \in \mathbb{N}$.

With this it is possible to prove the following theorem, which generalizes Theorem 4.2 of [7] to the semi-Markov case.

Theorem 2. *Suppose that Assumptions 1, 2 and 3 hold. Then, for all* $s \in S$,

$$\begin{aligned} 0 \leqq V^*(s) - U_N(s) \ &\leqq\ \rho^N L^N R(s) + \frac{1}{\alpha}\sum_{t=N}^{\infty} \rho^t (L^t r_0)(s) \\ &\leqq \left(1 + \frac{1}{\alpha}\right) \sum_{t=N}^{\infty} \rho^t (L^t r_0)(s)\ . \end{aligned}$$

Proof. Let f_N be the N-th **RH** policy. From Theorem 1, since for all n, the sequence $\{V_n\}$ results from successive application of the operator T,

$$V_N(s) = r(s, f_N)\tau(s, f_N) + \beta(s, f_N) \int_S V_{N-1}(j) Q(dj|s, f_N)\ . \tag{7}$$

Also, by definition, the function V_{N-1} verifies

$$V_{N-1}(s) = \sup_\pi \mathbb{E}_s^\pi \left[r(s, f_0)\tau(s, f_0) + \sum_{t=1}^{N-2} \prod_{k=0}^{t-1} \beta(S_k, A_k)\tau(S_t, A_t) r(S_t, A_t) \right],$$

and if we we add and subtract $\prod_{k=0}^{N-1} \beta(S_k, A_k)\tau(S_{N-1}, A_{N-1}) r(S_{N-1}, A_{N-1})$:

$$\begin{aligned} V_{N-1}(s) \leqq \sup_\pi \mathbb{E}_s^\pi \Bigg[& r(s, f_0)\tau(s, f_0) \\ &+ \sum_{t=1}^{N-1} \prod_{k=0}^{t-1} \beta(S_k, A_k)\tau(S_t, A_t) r(S_t, A_t) \\ &- \prod_{k=0}^{N-1} \beta(S_k, A_k)\tau(S_{N-1}, A_{N-1}) r(S_{N-1}, A_{N-1}) \Bigg]\ . \end{aligned}$$

Since for all $s \in S$ and $a \in A_s$, by Assumption 3, $-r(s,a) \leqq r_0(s)$ we have:

$$V_{N-1}(s) \leqq V_N(s) + \sup_{\pi} \mathbb{E}_s^{\pi} \prod_{k=0}^{N-1} \beta(S_k, A_k)\tau(S_{N-1}, A_{N-1}) r_0(S_{N-1}) \ .$$

By Lemma 1, $\mathbb{E}_s^{\pi}|r_0(S_t)| \leqq (L^t r_0)(s)$, and since $\tau(s,a) \leqq \frac{1}{\alpha}$ for all $s \in S$ and $a \in A_s$:

$$V_{N-1}(s) \leqq V_N(s) + \frac{\rho^{N-1}}{\alpha}(L^{N-1} r_0)(s) \ . \tag{8}$$

If we use this inequality in (7):

$$\begin{aligned} V_N(s) &= r(s,f_N)\tau(s,f_N) \\ &\quad + \beta(s,f_N)\int_S (V_N(j) + \frac{\rho^{N-1}}{\alpha}(L^{N-1} r_0)(j)) Q(dj|s,f_N) \\ &\leqq r(s,f_N)\tau(s,f_N) + \frac{\rho^N}{\alpha}(L^N r_0)(s) + \rho \int_S V_N(j) Q(dj|s,f_N). \end{aligned}$$

Iterations of this last inequality gives

$$\begin{aligned} V_N(s) &\leqq r(s,f_N)\tau(s,f_N) \\ &\quad + \beta(s,f_N)\int_S r(j,f_N)\tau(j,f_N) Q(dj|s,f_N) \\ &\quad + \frac{\rho^N}{\alpha}(L^N r_0)(s) + \frac{\rho^{N+1}}{\alpha} r_0(s) + \rho^2 \int_S V_N(j) Q^2(dj|s,f_N) \ , \end{aligned}$$

or, in general for $n \in \mathbb{N}$,

$$\begin{aligned} V_N(s) &\leqq r(s_0,f_0)\tau(s_0,f_0) \\ &\quad + \mathbb{E} \sum_{t=1}^{n} \prod_{k=0}^{t-1} \beta(S_k,f_N)\tau(S_t,A_t) r(S_t,f_N) \\ &\quad + \frac{1}{\alpha}\sum_{t=N}^{N+n} \rho^t (L^t r_0)(s) + \rho^{n+1}\mathbb{E}_s^{f_N}[V_N(S_{n+1})] \ . \end{aligned} \tag{9}$$

Let us analyse the terms of the r.h.s. of the last inequality, as $n \to \infty$.

The sum of the first and the second ones converges to $U_N(s)$ and the third to $\frac{1}{\alpha}\sum_{t=N}^{\infty} \rho^t L^t r_0(s)$. On the other hand, since for all $s \in S$, $|V_N(s)| \leqq R(s)$ the third term satisfies, by Lemma 1

$$\rho^{n+1}\mathbb{E}_s^{f_N}[V_N(S_{n+1})] \leqq \rho^{n+1}\mathbb{E}_s^{f_N}[R(S_{n+1})] \leqq \rho^{n+1}(L^{n+1}R)(s) \ ,$$

which converges to zero, since

$$\rho^{n+1}(L^{n+1}R)(s) \leqq \sum_{t=n+1}^{\infty} \rho^t (L^t r_0)(s) \tag{10}$$

and, by Assumption 3, the series $\sum_{t=0}^{\infty} \rho^t (L^t r_0)(s)$ is supposed to be convergent for all $s \in S$. Finally (9) implies

$$V_N(s) \leqq U_N(s) + \frac{1}{\alpha} \sum_{t=N}^{\infty} \rho^t (L^t r_0)(s) \ , \tag{11}$$

and

$$\begin{aligned} V^*(s) - U_N(s) &\leqq V^*(s) - V_N(s) + \frac{1}{\alpha} \sum_{t=N}^{\infty} \rho^t (L^t r_0)(s) \\ &\leqq \rho^N L^N R(s) + \frac{1}{\alpha} \sum_{t=N}^{\infty} \rho^t (L^t r_0)(s) \ . \end{aligned}$$

The second inequality stated by the Theorem is justified by (10).

Corollary 1 *If in Theorem 2, $r \geqq 0$, then, for all $s \in S$,*

$$0 \leqq V^*(s) - U_N(s) \leqq \rho^N L^N R(s).$$

Proof. If $r \geqq 0$, for all $s \in S$ and $n \in \mathbb{N}$, $V_n(s) \leqq V_{n+1}(s)$, and we obtain from (17) that $V_N(s) \leqq V_{N+1}(s)$, instead of the Inequality (8), and Inequality (9) is now

$$\begin{aligned} V_N(s) \leqq\ & r(s_0, f_0)\tau(s_0, f_0) \\ & + \mathbb{E}_s^{f_N} \sum_{t=1}^{n} \prod_{k=0}^{t-1} \beta(S_k, f_N)\tau(S_t, A_t) r(S_t, f_N) \\ & + \rho^{n+1} \mathbb{E}_s^{f_N} [V_N(S_{n+1})] \ . \end{aligned}$$

From here, following with the proof of Theorem 2, the new bound follows.

As in the case of Theorem 2, the next theorem generalizes Theorem 5.2 of [7] to the Semi-Markov environment.

Theorem 3. *Suppose that Assumptions 1, 2 and 4 hold. Then, for all $s \in S$,*

$$0 \leqq V^*(s) - U_N(s) \leqq \left(1 + \frac{1}{\alpha}\right) \frac{m\rho^N}{1-\rho} \omega(s) \ ,$$

or, equivalently

$$||V^* - U_N||_\omega \leqq \left(1 + \frac{1}{\alpha}\right) \frac{m\rho^N}{1-\rho} \ .$$

If in addition $r \geqq 0$, then

$$||V^* - U_N||_\omega \leqq \frac{m\rho^N}{1-\rho} \ .$$

Proof. Following the proof of Theorem 2 up to inequality (8) we obtain $V_{N-1}(s) \leqq V_N(s) + \frac{\rho^{N-1}}{\alpha}(L^{N-1}r_0)(s)$. If we assume now that Assumption 4 holds, then, for any $t \in \mathbb{N}$

$$||L^t r_0||_\omega \leqq ||L^{t-1}r_0||_\omega \leqq ... \leqq ||r_0||_\omega \leqq m \ ,$$

and $||R||_\omega \leqq \frac{m}{1-\rho}$, or, for all $s \in S$, $R(s) \leqq \frac{m}{1-\rho}\omega(s)$, and inequality (8) gives

$$V_{N-1}(s) \leqq V_N(s) + \frac{m}{\alpha}\frac{\rho^{N-1}}{1-\rho}\omega(s) \tag{12}$$

and (9) becomes

$$V_N(s) \leqq r(s, f_N)\tau(s, f_N) + \frac{m}{\alpha}\frac{\rho^N}{1-\rho}\omega(s) + \rho\int_S V_N(j)Q(dj|s, f_N) \ .$$

Keeping iterating, we obtain

$$\begin{aligned} V_N(s) \leqq\, & r(s_0, f_0)\tau(s_0, f_0) \\ & + \mathbb{E}_s^{f_N} \sum_{t=1}^{n} \prod_{k=0}^{t-1} \beta(S_k, f_N(S_k)) r(S_t, f_N(S_t))) \\ & + \frac{m}{\alpha}\sum_{t=N}^{N+n} \frac{\rho^t}{1-\rho}\omega(s) + \rho^{n+1}\mathbb{E}_s^{f_N}[V_N(S_{n+1})] \ . \end{aligned}$$

As in the proof of Theorem 2, the sum of the first and second term converges to $U_N(s)$ and the fourth to zero, as $n \to \infty$. The third converges to $\frac{m}{\alpha}\sum_{t=N}^{\infty}\frac{\rho^t}{1-\rho}\omega(s) = \frac{m}{\alpha}\frac{\rho^N}{1-\rho}\omega(s)$, and then the last inequality becomes

$$V_N(s) \leqq U_N(s) + \frac{m}{\alpha}\frac{\rho^N}{1-\rho}\omega(s) \ .$$

Finally, from Theorem 1,

$$V^*(s) - V_N(s) \leqq \frac{m\rho^N}{1-\rho}\omega(s) \ , \tag{13}$$

and then

$$V^*(s) - U_N(s) \leqq \left(1 + \frac{1}{\alpha}\right)\frac{m\rho^N}{1-\rho}\omega(s) \ .$$

Now, if we have $r \geqq 0$, inequality (12) could be put in the tighter form $V_{N-1}(s) \leqq V_N(s)$, and continuing with the proof just to inequality (13), for all $s \in S$, $V_N(s) \leqq U_N(s)$. Again the result follows combining Theorem 1 with this inequality.

Corollary 2 *If r is a bounded function on $\mathbb{K}$, and M is a bound of r, then for all $s \in S$*

$$0 \leqq V^*(s) - U_N(s) \leqq \left(1 + \frac{1}{\alpha}\right)\frac{M\rho^N}{1-\rho},$$

or equivalently

$$||V^* - U_N|| \leqq \left(1 + \frac{1}{\alpha}\right) \frac{M\rho^N}{1-\rho} \ .$$

If in addition $r \geqq 0$*,*

$$||V^* - U_N|| \leqq \frac{M\rho^N}{1-\rho} \ .$$

Proof. It follows immediately form Theorem 3, taking $m = M$ and $\omega \equiv 1$.

An Approximate Rolling Horizon Procedure.

Suppose now that the controller does not have an exact information of the value function of the problem of horizon $N-1$, which he should use to compute the optimal immediate actions, but he know (or is able to compute) an approximation of this value. Then the controller, standing in state s can choose within the available actions, the most favorable.

That is, for a function V, supposed to be close in some sense to V_{N-1}, choose

$$\tilde{f}_N(s) \in \arg\max_{a \in A_s} \left\{ r(s,a)\tau(s,a) + \beta(s,a) \int_S V(j) Q(dj|s,a) \right\} \ . \qquad (14)$$

We will note with $\tilde{U}_N$ the total discounted reward of the policy $\tilde{f}_N$. The next result gives answers to questions **Q1** and **Q2** stated in this section for the sequence of successive rewards $\tilde{U}_N$.

Theorem 4. *Suppose that Assumptions 1, 2 and 4 holds. Given a function* $V \in \mathcal{M}_\omega(S)$ *such that* $TV(s) \geqq V(s)$ *for all* $s \in S$*, and for some* $N \geqq 1$*,* $||V_{N-1} - V||_\omega \leqq \varepsilon$*, consider a policy* $f_N \in \mathit{\Pi}_S$ *such that, for any* $s \in S$*,* $f_N(s)$ *verifies (14). Then*

$$||V^* - \tilde{U}_N||_\omega \leqq \frac{m\rho^N}{1-\rho} + \frac{2\rho\varepsilon}{1-\rho} \ .$$

Proof. Start with the equality

$$V^* - \tilde{U}_N = V^* - TV + TV - \tilde{U}_N \ . \qquad (15)$$

First,

$$\begin{aligned} ||V^* - TV||_\omega &\leqq ||V^* - TV_{N-1}||_\omega + ||TV_{N-1} - TV||_\omega \\ &= ||TV^* - TV_{N-1}||_\omega + ||TV_{N-1} - TV||_\omega \\ &\leqq \rho||V^* - V_{N-1}||_\omega + \rho||V_{N-1} - V||_\omega \leqq \frac{m\rho^N}{1-\rho} + \rho\varepsilon \ . \qquad (16) \end{aligned}$$

On the other hand, for the definition of $\tilde{f}_N$:

$$\begin{aligned} TV(s) &= r(s,\tilde{f}_N)\tau(s,\tilde{f}_N) + \beta(s,\tilde{f}_N) \int_S V(j) Q(j|s,\tilde{f}_N) \\ &\leqq r(s,\tilde{f}_N)\tau(s,\tilde{f}_N) + \beta(s,\tilde{f}_N) \int_S [TV(j) + \varepsilon(1-\rho)\omega(s)] Q(j|s,\tilde{f}_N) \ . \end{aligned}$$

If we iterate this inequality under the integral sign of the operator T, for all $n \in \mathbb{N}$,

$$\begin{aligned} TV(s) &\leqq r(s, \tilde{f}_N)\tau(s, \tilde{f}) \\ &+ \mathbb{E}_s^{\tilde{f}_N} \sum_{t=1}^{n} \prod_{k=0}^{n-1} \beta(S_k, \tilde{f}_N)\tau(S_k, \tilde{f}_N) r(S_k, \tilde{f}_N) \\ &+ \sum_{k=1}^{n+1} \varepsilon(1+\rho)\rho^k \omega(s) + \rho^{n+1}\mathbb{E}_s^{\tilde{f}_N}[TV(S_{n+1})] \ . \end{aligned}$$

The sum of the first and second term of the r.h.s. of the last inequality tends to $\tilde{U}_N$ and the third to $\sum_{n=1}^{\infty} \varepsilon(1+\rho)\rho^k\omega(s) = \frac{\rho(1+\rho)\varepsilon}{1-\rho}\omega(s)$. The fourth term tends to zero. Therefore it follows that

$$TV(s) \leqq \tilde{U}_N(s) + \frac{\rho(1+\rho)\varepsilon}{1-\rho}\omega(s) \ , \tag{17}$$

and, joining (15), (16) and (17):

$$\begin{aligned} V^*(s) - \tilde{U}_N(s) &\leqq \left[\frac{m\rho^N}{1-\rho} + \rho\varepsilon + \frac{\rho(1+\rho)\varepsilon}{1-\rho}\right]\omega(s) \\ &= \left[\frac{m\rho^N}{1-\rho} + \frac{\rho\varepsilon}{1-\rho}\right]\omega(s) \ , \end{aligned}$$

or equivalently

$$||V^* - \tilde{U}_N||_\omega \leqq \frac{m\rho^N}{1-\rho} + \frac{2\rho\varepsilon}{1-\rho} \ .$$

Corollary 3 *If in the previous theorem r is a bounded function on $\mathbb{K}$, and M is an upper bound, then*

$$0 \leqq V^*(s) - \tilde{U}_N(s) \leqq \frac{M\rho^N}{1-\rho} + \frac{2\rho\varepsilon}{1-\rho} \ ,$$

or equivalently

$$||V^* - \tilde{U}_N|| \leqq \frac{M\rho^N}{1-\rho} + \frac{2\rho\varepsilon}{1-\rho} \ .$$

5 Concluding Remarks.

Through this work we have dealed with semi-Markov control models with discounted payoff, analyzing the performance of the rolling horizon procedure and of an approximate rolling horizon procedure.

We have imposed different conditions on the immediate reward function, the strongest one being its uniform boundedness.

We proved the convergence of the values related to the rolling horizon procedure to the optimal reward function under these assumptions. We obtain simple pointwise convergence if Assumption 3 is verified and pointwise geometrical convergence when Assumption 4 holds. As a particular case, we have obtained uniform geometrical convergence for the case of uniformly bounded rewards functions.

Finally we discuss an approximate rolling horizon procedure, based on the possibility of the controller of not to having perfect prediction of the future needed to take the best immediate action, but approximations of it.

References

1. Bhattacharya R., Majumdar, M.; *Controlled semi-Markov models - the discounted case.* Journal of Statistical Planning and Inference, 21, 365–381, 1989.
2. Bertsekas D.P., Shreve S.E.; *Stochastic Optimal Control: The Discrete Time Case.* Academic Press, New York, 1978.
3. Chang H.S., Marcus, S.I., *Two-person zero-sum games: receding horizon approach.* IEEE Transactions on Automatic Control, 48, 11, 2003, pp. 1951–1961.
4. Guo, X., Hernández-Lerma O., *Continuous-time controlled Markov chains.* Ann. Appl. Prob. 13, 2003, pp. 363–388.
5. Della Vecchia E., Di Marco S. and Jean-Marie A., *Rolling horizon and state space truncation approximations for zero-sum semi-Markov games with discounted payoff,* 16th Informs Applied Society Conference in Stockholm, July 6-8, 2011.
6. Hernández-Lerma O, Lasserre J.B., *Error bounds for rolling horizon policies in discrete-time Markov control processes.* IEEE Transactions on Automatic Control, 35, 10, 1990, pp. 1118 - 1124.
7. Hernández-Lerma O, Lasserre J.B., *Value iteration and rolling plans for Markov control processes with unbounded rewards.* J. Math. Anal. Appl, 177, 1993, pp. 38 55.
8. Kwon W, Han S., *Receding Horizon Control. Predictive Control for State Models.* Advanced Textbooks in Control and Signal Processing, Springer, 2005.
9. Luque-Vásquez, F., *Zero-sum semi-Markov game in Borel spaces with discounted payoff.* Universidad de Sonora, Mexico. 2002.
10. Puterman L., *Markov Decision Processes.* Wiley and Sons, 2005.
11. Schweitzer P., Federgruen A., and Tijms, H., *Denumerable undiscounted semi-Markov decision processes with unbounded costs.* Math. Op. Res. 8, 1983, pp. 298–311.
12. Sethi T., Sorger G., *A theory of rolling horizon decision making.* Ann. Op. Res. 29, 1, 1991, pp. 387–415.
13. van Numen J., Jaap W., *A note on dynamic programming with unbounded rewards.* Managements Science 24, 1978, pp. 576–580.