

Illustrated review of convergence conditions of the value iteration algorithm and the rolling horizon procedure for average-cost MDPs

Eugenio Della Vecchia, Silvia C. Di Marco, Alain Jean-Marie

▶ **To cite this version:**

Eugenio Della Vecchia, Silvia C. Di Marco, Alain Jean-Marie. Illustrated review of convergence conditions of the value iteration algorithm and the rolling horizon procedure for average-cost MDPs. [Research Report] RR-7710, LIRMM; INRIA. 2011. inria-00617271

HAL Id: inria-00617271
https://inria.hal.science/inria-00617271

Submitted on 26 Aug 2011

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.

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Illustrated review of convergence conditions of the value iteration algorithm and the rolling horizon procedure for average-cost MDPs

Eugenio Della Vecchia — Silvia Di Marco — Alain Jean-Marie

N° 7710

August 2011

Domaine 3



ISSN 0249-6399 ISRN INRIA/RR--7710--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SOPHIA ANTIPOLIS - MÉDITERRANÉE

Illustrated review of convergence conditions of the value iteration algorithm and the rolling horizon procedure for average-cost MDPs

Eugenio Della Vecchia [*] , Silvia Di Marco [†] , Alain Jean-Marie[‡]

Domaine : Réseaux, systèmes et services, calcul distribué
Équipe-Projet Maestro

Rapport de recherche n° 7710 — August 2011 — 23 pages

Abstract: This paper is concerned with the links between the Value Iteration algorithm and the Rolling Horizon procedure, for solving problems of stochastic optimal control under the long-run average criterion, in Markov Decision Processes with finite state and action spaces. We review conditions of the literature which imply the geometric convergence of Value Iteration to the optimal value. Aperiodicity is an essential prerequisite for convergence. We prove that the convergence of Value Iteration generally implies that of Rolling Horizon. We also present a modified Rolling Horizon procedure that can be applied to models without analyzing periodicity, and discuss the impact of this transformation on convergence. We illustrate with numerous examples the different convergence results.

Key-words: Markov decision problems, Value iteration, Heuristic methods, Rolling horizon.

[*] CONICET - UNR, Argentina
[†] CONICET - UNR, Argentina
[‡] INRIA and LIRMM, CNRS/Université Montpellier 2, 161 Rue Ada, F-34392 Montpellier, ajm@lirmm.fr.

Centre de recherche INRIA Sophia Antipolis – Méditerranée
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

Une revue illustrée des conditions de convergence pour l'algorithme d'itération de valeur et la procédure de l'horizon roulant, pour les processus de décision Markoviens en coût moyen

Résumé : Nous nous intéressons aux relations entre l'algorithme d'itération de valeurs et la procédure de l'horizon roulant, pour résoudre les problèmes de contrôle optimal stochastique Markovien sous le critre du coût moyen, dans le cas d'espaces d'états et d'actions finis. Nous passons en revue des conditions issues de la littérature qui impliquent la convergence géométrique de l'itération de valeurs vers la valeur optimale. L'apériodicité du modèle est un pré-requis essentiel. Nous montrons que la convergence de l'itération de valeurs implique de façon générale celle de l'horizon roulant. Nous présentons également une procédure modifiée d'horizon roulant qui peut être appliquée sans avoir besoin d'analyser l'apériodicité, et nous étudions l'impact de cette transformation sur la convergence. Nous illustrons les différents résultats avec de nombreux exemples.

Mots-clés : Processus de décision Markovien, itération de valeurs, méthodes heuristiques, horizon roulant.

1 Introduction

1.1 Statement of the problem: precision of the rolling horizon procedure

Consider a random dynamical system, observed at discrete times. At each time $t \in \mathbb{N}$, the state s_t is observed and an action a_t is chosen, resulting in an instantaneous gain $r_t(s_t, a_t)$. From here on, we work with time-homogeneous gains: $r_t = r$, independent on the time. The actions a_t chosen at each time t in the respective state s_t determine a policy π whose performance is evaluated through a long-run average criterion. More precisely, let

$$g^\pi(s) := \lim \inf_{n\to\infty} \frac{1}{n} \mathbb{E}_s^\pi \sum_{t=0}^{n-1} r(s_t, a_t). \tag{1}$$

The objective of the controller is to find (when it exists) the policy that solves, given the current state s:

$$\pi^*(s) \;=\; \arg\max_\pi g^\pi(s)\ .$$

However, for a wide class of stochastic control problems in discrete time and infinite horizon, 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, Moving Horizon or Model Predictive Control), which works as follows. To the infinite-horizon control problem is associated a finite-horizon problem (**FHP**): for a given integer n (the horizon length) and a state s, find:

$$\max_\pi \mathbb{E}^\pi \left[\sum_{t=0}^{n-1} r(s_t, a_t) | s_0 = s \right]\ . \tag{2}$$

Solving this problem results in a sequence of decision rules:

$$\pi_n^* = (d_n, d_{n-1}, \ldots, d_2, d_1) \tag{3}$$

where $d_1(s_{n-1})$ is the best action to be applied at time $t = n-1$ when only one step remains to reach the horizon, d_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, $d_n(s_0)$ is the best decision rule to be applied to the initial state s_0.

The Rolling Horizon method (abbreviated as **RH** from here on), prescribes to repeatedly solve a **FHP**, 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 rolling horizon policy is the following one. Fix some integer n.

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

The **RH** procedure does not specify how to compute the value $d_n(x_t)$. Its efficiency is based on the idea that computing the value $d_n(x_t)$ alone is usually much easier than solving entirely the **FHP**, which involves computing the n decision rules in (3). 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 $d_n(x_t)$. 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 $\lim_{n\to\infty} u_n(s) = g^{\pi^*}(s)$?

Q2 Given a state s and $\epsilon > 0$, is it possible to compute n such that $|u_n(s) - g^{\pi^*}(s)| < \epsilon$?

In this paper, we look at these questions in the context of Markov Decision Processes, through the link it has with the *Value Iteration* (**VI**) algorithm and the **RH** procedure. We focus on two objectives:

- make a review of the results about convergence of the **VI** algorithm in the literature, including the multichain model,
- analyze the effects of those properties on the convergence of the **RH** procedure and, in this way, make more practical and wide the use of this method.

We finally propose a modification of the **RH** procedure that makes convergence easier, and discuss its efficiency.

The paper is organized as follows. We complete this introduction with a brief literature review. In Section 2 and 3, we recapitulate the relationship between **RH** and the Value Iteration algorithm concerning their convergence concepts. Then in Section 4 we propose and evaluate a Modified Rolling Horizon procedure. Finally, in Section 5, we discuss about stopping rules for both of algorithms and we conclude in Section 6.

1.2 Literature review

Markov decision problems have been widely studied during the last sixty years, and the advances and applications have been synthesized in well-known books such as for example [3, 13, 2, 18, 12, 9]. The value iteration algorithm is an usual topic in the bibliography, frequently associated to discounted criteria. The analysis of the problem when the performance of the policies is evaluated with the criterion of average rewards over an infinite horizon, in the most general case, presents additional difficulties. This is why this topic has been developed more recently in the literature. For example, it is not present in [3] and few words are devoted to it in [13]. In most of the references, convergence of the value iteration algorithm for average rewards criterion is analyzed only for unichain models. Since the distinction between unichain and multichain turns out to be NP-complete to decide (see [9]), it should be useful to give convergence results for multichain models which work also for unichain models.

We shall consider Markov Decision Processes as described for instance in Puterman [12], of whom we adopt the notation. We also refer to the classification of Markov decision problems according to their deterministic stationary policies, i.e. unichain or multichain, as it appears in [9, 12].

The theoretical analysis of **RH** for MDPs has comparatively received less attention in the literature, where this procedure is often encountered in a heuristic presentation, without precise references to accuracy or convergence. In the *discounted* case, Puterman [12, Theorem 6.3.1, p. 161] proves that both **VI** and **RH** converge at the same time (see the definitions of convergence below). Results for the case of average costs include those of Hernández-Lerma and Lasserre, who present in [7] error bounds for rolling horizon policies in general, stationary and nonstationary, Markov control problems on Borel spaces, with both discounted and average reward criteria. They give a condition (Assumption 5.1 in their work and Condition 5 below in this work) under which the reward of the rolling horizon policy converges geometrically to the optimal reward function, uniformly in the initial state, as the length of the rolling horizon increases. The convergence rate is explicit in their result. Previously, Alden and Smith in [1] provided an error bound, still for nonstationary MDPs, between a rolling horizon policy and an expected-average optimal policy, considering finite states and finite policies under a Doeblin-like condition (see [11]). Guo and Shi, in [6], deal with the limiting average criteria for nonstationary Markov decision processes on Borel state spaces with possibly unbounded rewards . They give conditions under which the existence of both a solution to the optimality equations and the limiting average ε-optimal Markov policies can be derived and also present a rolling horizon algorithm for computing limiting average ε-optimal Markov policies. The proof of the convergence is under a condition similar to those in [7].

2 The Value Iteration algorithm for the average reward criterion

In what follows, the state space, S, and the decision set for each $s \in S$, A_s are both finite. Also, without losing generality, we consider $r(s,a) \geq 0, \forall\ (s,a) \in S \times A_s$.

We shall focus on stationary policies $\pi = (d)^\infty = (d, d, \ldots)$ where d is a decision rule that maps every state s of S to A_s. Every decision rule can be seen as a vector with $|S|$ components. When vectorial notation is possible, for short we write r_d for the vector whose components are $r(s, d(s))$ and P_d for the transition matrix where $P_d(s, s') = p(s'|s, d(s))$. Moreover, making an abuse of notation, we use D indifferently for the set of decision rules or the set of stationary policies.

In this finite state space/finite action space setting, it is well-known that there exists an optimal, pure stationary policy, for the infinite-horizon average reward criterion (1). Let g^* denote the associated optimal gain vector.

Consider now the Value Iteration algorithm:

Value iteration algorithm

1. $n = 0$, $v_0 = 0$.
2. Compute
$$v_{n+1} = \max_{d \in D}\{r_d + P_d v_n\} =: T v_n \tag{4}$$
and some
$$d_{n+1} \in \arg\max_d \{r_d + P_d v_n\}. \tag{5}$$
3. If an adequate stopping rule holds, then go to step 4. Otherwise, set $n := n + 1$ and go to step 2.

4. Return d_{n+1}.

When **VI** stops, for some N, it has computed a sequence of decision rules $(d_N, \ldots, d_1)$ which actually solves the **FHP** in (2). On the other hand, it has been observed in the literature (*e.g.* in [7]) that the **RH** procedure with horizon n generates decisions precisely according to the stationary policy $(d_n)^\infty = (d_n, d_n, \ldots)$.

Convergence concepts. The usual theoretical and practical challenge for **VI** is to determine the "adequate stopping rule" of step 3, so that the algorithm does stop at some iteration n such that the policy computed is "good enough".

It is known that v_n/n always converges to g^*, see [8, Corollary 2.8, p. 49], but this result alone does not help to identify optimal or ϵ-optimal policies. We choose therefore the following, more practical, notion of convergence. The **VI** algorithm is said to converge if the following limit exists, for some vector h^*: [1]

$$\lim_{n\to\infty} v_n \; - \; ng^* \; = \; h^* \; . \tag{6}$$

In addition, the convergence is said to be geometric if there exists $N \in \mathbb{N}$, $C > 0$ and $\delta < 1$ such that, with a suitable norm, $\forall n \geq N$,

$$||v_n - ng^* - h^*|| \; < \; C\delta^n \; .$$

These definitions are motivated by the fact that the (nonstationary but periodic) policy $(d_N, \ldots, d_1, d_N, \ldots)$ has the performance v_N/N for the average criterion (1). On the other hand, most practitioners are likely to use instead the decision rule d_N alone repeatedly (out of simplicity, or the belief that this rule must be the "best" of those computed by **VI**), thereby implementing effectively a **RH** procedure with horizon N. The performance obtained by using this stationary policy is $u_N = g^{(d_N)^\infty}$, and in general, $u_N \neq v_N/N$ and $u_N \neq v_{N+1} - v_N$, nor is there a particular order between these sequences. We illustrate this fact below. As a consequence, any convergence result or stopping rule for **VI** is not guaranteed to provide a performance bound for **RH**. This motivates further the need for results concerning specifically **RH**.

By analogy, the **RH** procedure is said to converge if

$$\lim_{n\to+\infty} g^{(d_n)^\infty} \; = \; g^* \; .$$

The convergence is said to be geometric if there exists $N \in \mathbb{N}$, $C > 0$ and $\delta < 1$ such that, with a suitable norm, $\forall n \geq N$,

$$||g^{(d_n)^\infty} - g^*|| \; < \; C\delta^n \; .$$

This is an abuse of terminology, since the **RH** procedure itself does not "converge". It merely means that the procedure can be made to perform arbitrarily close to optimal through the choice of a suitable horizon length n.

Example 1. *The example detailed in the Appendix serves to illustrate these issues concerning the convergence of the different sequences involved in the previous discussion.*

[1] Observe the discrepancy with the general notion of convergence of algorithms in Computer Science, which requires that an algorithm stops *and* returns the correct result.

We have applied to this model the transformation to be described in Section 4 with $\tau = 0.99$*. In Figure 1, we show the evolutions of the sequences*

$$v_{n+1} - v_n, \quad v_{n+1} - v_n - g^{(d_n)^\infty}, \quad v_n/n \quad \textit{and} \quad v_n/n - g^{(d_n)^\infty}$$

respectively, evaluated at state 4. Clearly, these sequences do not have a constant sign, and are not monotonously converging.

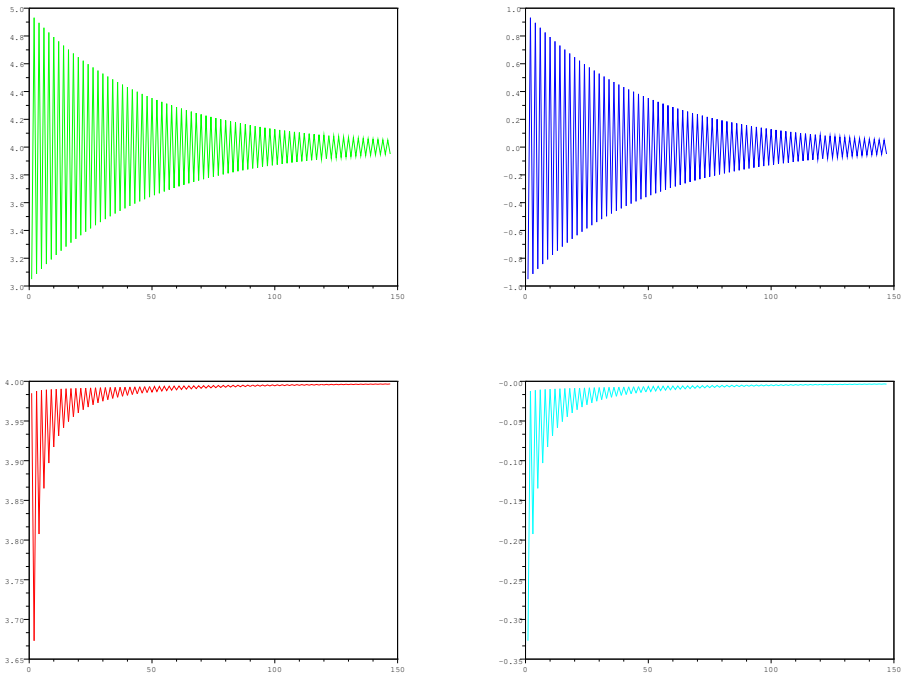


Figure 1: Evolution of $v_{n+1} - v_n$ (top), $v_{n+1} - v_n - g^{(d_n)^\infty}$ (top, right), v_n/n (bottom, left) and $v_n/n - g^{(d_n)^\infty}$ (bottom, right), for state 4

3 Convergence of the VI and RH procedures

The issue of convergence of the **VI** algorithm has attracted quite some attention in the literature. In Section 3.1, we review some of the conditions that have been proposed for ensuring the convergence of either **VI** or **RH**. In Section 3.2, we state the convergence results, including a new one which we prove (Theorem 2). In Section 3.3, we have a look at convergence rates. In Section 3.4, we discuss the relative strength of these conditions.

3.1 Convergence conditions

The two following conditions are stated by Schweitzer and Federgruen in [15].

Condition 1. *There exists a randomized maximal gain policy whose transition probability matrix is aperiodic (but not necessarily unichain) and has $R^* = \{i \in S : i$ is recurrent for some pure maximal gain policy $\}$ as its set of recurrent states.*

Condition 2. *Every optimal (pure) stationary policy gives rise to an aperiodic (but not necessarily unichain) transition matrix.*

The following condition, known as *weak unichain condition* appears in Tijms [17, p. 199] as Assumption 3.3.1.

Condition 3. *Every optimal stationary policy has a transition probability matrix unichain and aperiodic*

Puterman in [12, p. 370], presents the following one.

Condition 4. *Every stationary policy is unichain and gives rise to an aperiodic transition matrix.*

The following condition appears in Hernández Lerma and Lasserre [7] as Assumption 5.1.

Condition 5. *There exists a positive number $\delta < 1$ such that*

$$\boldsymbol{sp}(p(.|s,a) - p(.|s',a')) \leq 2\delta$$

for every (s,a) and (s',a') with $s, s' \in S$, $a \in A_s$, $a' \in A_{s'}$ and for a measure λ, $\boldsymbol{sp}(\lambda)$ denote the norm

$$\boldsymbol{sp}(\lambda) := \sup_B \lambda(B) - \inf_B \lambda(B)$$

for $B \subset S$.

Remark 1. *In Section 3.4 we show that Condition 5 $\Rightarrow$ Condition 4. It is easy to see that Condition 4 $\Rightarrow$ Condition 3 $\Rightarrow$ Condition 2 $\Rightarrow$ Condition 1.*

Condition 1 is the weakest condition under which the convergence of $v_n - ng^*$ to h^* is guaranteed: it is established in [15] that this is a necessary and sufficient condition of convergence.
Condition 5 for the convergence of **RH** is related to some ergodicity properties of the chain structure: see Appendix in [7], or [11, Chapter 15]. Other convergence conditions have been proposed in [1, 6]. More precisely, the hypothesis of Alden and Smith in [1], related to the conditions described above, is similar to a Doeblin condition while in [6], Guo and Shi assume that $\beta := \sup\{1 - \inf_{j \in S} \inf_{a \in A_s} P^n(j|i,a) < 1 : n \geq 0\}$. Both assumptions, similarly to Condition 5 (see in Theorem 3 below) imply that the model is unichain.

3.2 Convergence results

It is known from the literature that, in general, there is no convergence of the sequence $\{v_n\}_n$, which is unbounded (it actually grows asymptotically linearly, r being positive, since v_n/n converges to some g^* as mentioned earlier), nor of $\{v_n - v_{n-1}\}_n$. Also, it is proved, see [9, Lemma 5.5, p. 157], that if the sequence $\{v_n - ng^*\}_n$ is bounded,

$$g^* = \lim_{n\to\infty} \frac{1}{n} v_n = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} (v_k - v_{k-1}) \ ,$$

but this is not enough to compute an ε-optimal policy or an ε-approximation of g^*. To have one, the convergence of $\{v_n - ng^*\}_n$ is needed, and it is known that it may fail to happen if some of the matrices involved in the MDP are periodic.

The analysis of average-cost MDPs often involves the notion of "span" of a function (or vector), defined as: $\mathbf{sp}(w) := \max_{s\in S}(w(s)) - \min_{s\in S}(w(s))$. The geometric convergence of the sequence $\mathbf{sp}(v_{n+1} - v_n)$ to 0 implies the geometric convergence of **VI**, and the limit of $v_{n+1} - v_n$ is a vector with zero span, that is, a constant vector.

Theorem 1 (Convergence of **VI**). *The **VI** algorithm converges geometrically under any of Conditions 1– 5.*

Proof. Obviously, in view of Remark 1, it is sufficient to prove the result for Condition 1. It is interesting for this review to point out that proofs under specific conditions have been obtained independently, since these may involve different techniques and possibly provide different estimations for the convergence rate. See Section 3.3 below.

The convergence under Conditions 1 and 2 is proved in [15], Theorems 5.1 and 5.5 respectively. The fact that the convergence is geometric is proved by the same authors in [16, Theorem 4.2].

The claim for Condition 3 is proved in [17, Theorem 3.4.2, p. 209].

We can find the proof of the convergence under Condition 4 in [12, Theorem 8.5.4, p. 370]. The arguments do not include geometric convergence, but this property holds since Condition 4 implies Condition 3 (for this last one, again, see Section 3.3 below).

Finally, the result under Condition 5 is not proved, but commented in [7]. However, it is not hard to check that Condition 5 is equivalent to Condition a) in [12, Theorem 8.5.3, p. 368] and this theorem, together with Theorems 8.5.1 and 8.5.2, provide the geometric convergence of $\mathbf{sp}(v_{n+1} - v_n)$ to 0. □

The previous theorem means that under the conditions mentioned the sequence $\{v_n - ng^*\}$ converges geometrically to h^*. Now we ask if under any of these conditions the sequence $\{g^{(d_n)^\infty} - g^*\}$ converges geometrically to zero.

Theorem 2 (Convergence of **RH**). *If the **VI** algorithm converges geometrically then also does the **RH** procedure.*

Proof. By assumption, $\exists N_1 \in \mathbb{N}$, $C_1 > 0$ and $\alpha < 1$ such that whenever $n \geq N_1$,

$$||v_n - ng^* - h^*||_\infty \ < \ C_1\alpha^n \ .$$

Since $v_{n+1} - v_n - g^* = (v_{n+1} - (n+1)g^* - h^*) - (v_n - ng^* - h^*)$, with $C = 2C_1$ we have, for $n \geq N_1$,

$$||v_{n+1} - v_n - g^*||_\infty \ < \ C\alpha^n \ ,$$

or put differently,

$$g^* - C\alpha^n\mathbf{1} \ < \ v_{n+1} - v_n \ < \ g^* + C\alpha^n\mathbf{1} \ . \tag{7}$$

Let d_n be defined as (see (5)):

$$d_n \in \arg\max_d\{r_d + P_d v_{n-1}\}.$$

Denoting with $P^*_{d_n} := \lim_{m\to\infty} \frac{1}{m}\sum_{k=1}^m P^k_{d_n}$, the Cesàro limit for P_{d_n}, we have $g^{(d_n)^\infty} = P^*_{d_n} r_{d_n}$ and $g^{(d_n)^\infty} = P^*_{d_n} g^{(d_n)^\infty}$.

Moreover, according to [12, Lemma 9.4.3], there exists N_2 such that for all $n > N_2$, $P_{d_n} g^* = g^*$, which implies $P^*_{d_n} g^* = g^*$. Then, since $P^*_{d_n} P_{d_n} = P^*_{d_n}$,

$$g^{(d_n)^\infty} = P^*_{d_n} r_{d_n} = P^*_{d_n}(r_{d_n} + P_{d_n} v_n - v_n) = P^*_{d_n}(v_{n+1} - v_n). \tag{8}$$

It is clear that for all $s \in S$, $g^*(s) \geq g^{(d_n)^\infty}(s)$. Consequently, for $n > \max\{N_1, N_2\}$ and any s, (8) and (7) imply respectively

$$P^*_{d_n}(v_{n+1} - v_n)(s) = g^{(d_n)^\infty}(s) \leq g^*(s) \tag{9}$$

$$P^*_{d_n}(v_{n+1} - v_n)(s) > P^*_{d_n}(g^* - C\alpha^n \mathbf{1})(s) = g^*(s) - C\alpha^n. \tag{10}$$

From (9) and (10), for $n > \max\{N_1, N_2\}$, it follows that

$$0 \leq g^*(s) - g^{(d_n)^\infty}(s) < C\alpha^n, \tag{11}$$

which concludes the proof. □

Theorems 1 and 2 imply:

Corollary 1. *Any of Condition 1 to 5 implies the geometric convergence of the **RH** procedure.*

The reciprocal of Theorem 2 cannot hold: there are MDP models in which **RH** converges, whereas **VI** does not. The simplest such example is perhaps that of an uncontrolled, two state, periodic Markov chain (see Example 7 below): since there is only one policy, **RH** converges, but since the model does not satisfy Condition 1, **VI** cannot converge.

3.3 Convergence rates

Having an estimate of the convergence rate is useful in many practical situations, especially for determining horizon lengths or stopping rules (see Section 5). We discuss this point now.

The question is to find values of N, C and δ such that, for all $n \geq N$:

$$||ng^* - v_n - h^*|| \ < \ C\delta^n \tag{12}$$

or for the **RH** procedure,

$$||g^* - g^{(d_n)^\infty}|| \ < \ C\delta^n \ . \tag{13}$$

Under Condition 5, Inequality (13) holds with $N = 1$, $C = 2||r||/(1-\delta)$ and δ given accordingly to [7, Proposition 5.1].

Under Condition 3, [5, Theorem 5] states that for $M = \frac{1}{2}|S|(|S|-1)$, and any pair of M-tuples of decision rules $\pi_1, \pi_2 \in D^M$,

$$\min_{s_1, s_2 \in S} \sum_{j \in S} \min\left\{P^M_{\pi_1}(s_1, j), P^M_{\pi_2}(s_2, j)\right\} \ =: \ \rho_{\pi_1,\pi_2}(M) > 0. \tag{14}$$

As S and D are considered finite in this discussion, taking

$$\tilde{\delta} = 1 - \min_{\pi_1, \pi_2 \in D^M} \rho_{\pi_1,\pi_2}(M), \tag{15}$$

we have $\tilde{\delta} < 1$ and from [12, Theorem 8.5.2.a., p. 368] T is a M-step contraction operator with coefficient $\tilde{\delta}$ and then

$$0 \leq g^* - g^{(d_{nM})^\infty}(s) \;\leq\; \mathbf{sp}(g^* - v_{nM}) \leq C\tilde{\delta}^n, \tag{16}$$

and $C = \mathbf{sp}(g^* - v_0)$. Therefore, (13) holds with $\delta = (\tilde{\delta})^{1/M}$, at least when n is a multiple of M. In [16] the authors claim that, for the same value of M, and $d \in D$,

$$\min_{s_1,s_2 \in S} \sum_{j \in S} \min\left\{P_d^M(s_1,j), P_d^M(s_2,j)\right\} \;=\; \rho_d(M) > 0,$$

and that there is geometric convergence with rate

$$\gamma \;=\; 1 - \min_{d \in D} \rho_d(M). \tag{17}$$

They do not present a proof but they also refer to [5, Theorem 5]. While it is clear that $\tilde{\delta} \geq \gamma$, the fact that γ is also a rate of convergence is not obvious.

In the following example we compute both $\tilde{\delta}$ and γ defined in (15) and (17), in order to illustrate the facts that: these numbers are different, and also that the importance of this result is essentially theoretic. It is clear that the complexity of the task grows exponentially with the number of states and actions.

Example 2. *Let us consider a model with three states $S = \{s_1, s_2, s_3\}$ and two available action in each state, $A_{s_i} = \{a_1^i, a_2^i\}$, $i = 1, 2, 3$. The (positive) transitions probabilities are defined as follows:*

$$p(s_{i+1}|s_i, a_1^i) = p(s_{i-1}|s_i, a_2^i) = 0.9 \text{ and } p(s_i|s_i, a_1^i) = p(s_i|s_i, a_2^i) = 0.1$$

(where the sum within subindices is modulo 3).

Since we have a 3-state model, the value of M is 3. Since there are two actions in each state, there are 8 different stationary policies, and 8 different matrices appearing in the definition of $\rho_{\pi_1,\pi_2}(M) > 0$ in (14)*.*

In the computation of γ, we obtain a value $\rho = 0.297$ for two policies: (a_1^1, a_1^2, a_1^3) and (a_2^1, a_2^2, a_2^3), and a value $\rho = 0.487$ for the rest. We find then the value $\gamma = 1 - 0.297 = 0.703$.

In the computation of $\tilde{\delta}$, we obtain a largest value of ρ_{π_1,π_2} as 0.244, obtained for instance with $\pi_1 = (a_1^1, a_1^2, a_2^3)$ and $\pi_2 = (a_1^1, a_2^2, a_1^3)$. We then have $\tilde{\delta} = 0.756$. As established above, $\tilde{\delta} \geq \gamma$.

The corresponding convergence rates are respectively $\tilde{\delta}^{1/3} \simeq 0.911$ and $\gamma^{1/3} \simeq 0.889$.

In summary, only Condition 5 readily provides bounds on the convergence rate. It is an open question to design algorithms which compute bounds on the convergence rate in an efficient way for generic models. This does not preclude the possibility that bounds be established based on (15) for specific models.

3.4 Discussion and comparison of convergence conditions

Discussion on the practicality of convergence conditions. Conditions 1-4 of Section 3.1 all involve aperiodicity, irreducibility and/or the classification of MDP models. Although determining whether a MDP is irreducible is polynomially solvable, there is no polynomial algorithm to determine whether a MDP is unichain or multichain

(see comments in [9, p. 127]). It is worth mentioning that a simple transformation, discussed below, makes aperiodic any transition matrix of the MDP model without changing the optimization problem. Conditions requiring aperiodicity can therefore be applied without this requirement.

Comparison. Among the five Conditions we have reviewed, Condition 5 is the only one not referring to structural properties of the underlying matrices. We investigate here these structural implications.
The following lemma provides a convenient characterization of cases where Condition 5 does *not* hold. Remember that, given a probability measure μ, supp(μ) is the smallest set B such that $\mu(B) = 1$.

Lemma 1. *Condition 5 does not hold if and only if there exist* $(s, s') \in S \times S$, $a \in A_s$, $a' \in A_{s'}$, *such that supp*$(p(\cdot|s, a)) \cap$ *supp*$(p(\cdot|s', a')) = \emptyset$.

Proof. Since the state and action spaces are finite, Condition 5 fails if and only if there exist s, s', a, a' such that: $\mathbf{sp}(\lambda) = 2$, with $\lambda(.) = p(.|s, a) - p(.|s', a')$. This in turn is equivalent to: $\sup_B \lambda(B) = 1$ and $\inf_B \lambda(B) = -1$, and finally: $\exists B, B'$: $p(B|s, a) = 1$, $p(B'|s, a) = 0$, $p(B|s', a') = 0$ and $p(B'|s', a') = 1$. The set B can be taken as supp$(p(.|s, a))$ and $B' =$ supp$(p(.|s', a'))$. □

Next, Condition 5 implies structural properties on the MDP model.

Theorem 3. *Every model where Condition 5 holds is a) unichain and b) aperiodic. The converse is not true.*

Proof. The proof of the first statement is by contradiction. Consider first part a) of the statement, and assume there is some d, a decision rule associated to a multichain Markov chain with transition probabilities P_d, B_1 and B_2 two distinct irreducible recurrent classes. Then choosing $s_1 \in B_1$, $d(s_1) \in A_{s_1}$, $s_2 \in B_2$ and $d(s_2) \in A_{s_2}$, we have supp$(p(\cdot|s_1, d(s_1))) \subset B_1$, supp$(p(\cdot|s_2, d(s_2))) \subset B_2$, implying that both sets are disjoint. Lemma 1 applies and Condition 5 cannot hold.

Next, consider part $b)$ of the statement: we prove that periodic chains do not verify Condition 5. We use for this the following result (see [4, p. 161]):

Lemma 2. *Let X an irreducible Markov chain with recurrent states of period γ. Then states can be divided into γ disjoint sets $B_1, B_2, \ldots, B_\gamma$ where $p(j|i) = 0$ unless $i \in B_1$ and $j \in B_2$ or $i \in B_2$ and $j \in B_3$, ..., or $i \in B_\gamma$ and $j \in B_1$.*

Thus, let us suppose a MDP unichain and periodic for the decision rule d. Denote with $\gamma > 1$ the period of P_d. Then, there exists $B_1, B_2, ..., B_\gamma$ sets as provided in Lemma 2. By construction, for every $s_i \in B_i$, supp$(p(\cdot|s_i, d(s_i))) \subset B_{i+1}$, where it is understood that "$\gamma + 1$" means 1. The supports are therefore disjoint for s_i and s_j if $i \neq j$. Therefore, Lemma 1 applies with any $s \in B_1$, $a = d(s)$, $s' \in B_2$, $a' = d(s')$.

Finally, to prove the converse statement, we exhibit an aperiodic model for which Condition 5 does not hold.

Example 3. *Let $S = \{s_1, s_2, s_3\}$, $A_{s_1} = \{a_1^1, a_2^1\}$, $A_{s_2} = \{a_1^2\}$, $A_{s_3} = \{a_1^3\}$. The controlled transition probabilities can be associated to the rules $d_1 = (a_1^1, a_1^2, a_1^3)$ y $d_2 = (a_2^1, a_1^2, a_1^3)$ and the*

$$P_{d_1} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}, \quad P_{d_2} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}.$$

Taking $k = (s_1, a_1^1)$ and $k' = (s_2, a_1^2)$, we have supp$(p(\cdot|k)) = \{s_2\}$ and supp$(p(\cdot|k') = \{s_3\}$. These sets are disjoint, so that by Lemma 1, Condition 5 fails.

□

To conclude this section, we discuss the idea of possibly relaxing the aperiodicity assumption. We show through the following example (obtained as a simplification of Example 4 in [10]) that the **RH** procedure may not converge on MDPs for which the optimal policy gives rise to an unichain periodic Markov process. The following section will explain how to handle this problem.

Example 4. *Let $S = \{s_1, s_2, s_3\}$ be the state space, and the action sets $A_{s_1} = \{a_1^1, a_2^1\}$, $A_{s_2} = \{a_1^2\}$, $A_{s_3} = \{a_1^3\}$. The rules we can construct are $d_1 = (a_1^1, a_1^2, a_1^3)$ and $d_2 = (a_2^1, a_1^2, a_1^3)$. The transition matrix are P_{d_1} and P_{d_2}. P_{d_2} is periodic of period 2.*

$$P_{d_1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \; P_{d_2} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

We consider the rewards

$$r(s_1, a_1^1) = 2, \; r(s_1, a_2^1) = 2, \; r(s_2, a_1^2) = 5, \; r(s_3, a_1^3) = 1.$$

*The **RH** algorithm leads to a sequence where d_2 appears for odd horizons and d_1 for even horizons. When d_2 is considered as stationary policy in the infinite horizon problem, it produces an average value equal to $(3, 3, 3)$, while d_1, an average reward equal to $(2, 3, 3)$. In consequence, the sequence of values does not converge. Moreover, it is immediate that the subsequence corresponding to odd horizons is better than that to even horizons.*

4 Modified Rolling Horizon Procedure

We present a standard transformation under which all policies are perturbed to give rise to aperiodic Markov chains without modifying the corresponding gain. This transformation is similar to what Puterman presents in [12, Section 8.5.4] and, up to our knowledge, it is originally due to Schweitzer (see [14]). Although these authors proposed it in the context of unichain models, it is clear that it works also in multichain models.

4.1 Transformation and modified procedure

Let $0 < \tau < 1$. Define $S_\tau = S$, $A_{s,\tau} = A_s$ for all $s \in S$. For all s and all $a \in A_{s,\tau}$,

$$r_\tau(s, a) = r(s, a)$$

and for all $j \in S$,

$$p_\tau(j|s, a) = (1 - \tau)\delta_s(\{j\}) + \tau p(j|s, a) \,, \tag{18}$$

where $\delta_s(\cdot)$ is the Dirac measure concentrated at s. Thus, for every decision rule d,

$$P_{d,\tau} = (1 - \tau)I + \tau P_d, \quad r_{d,\tau} = r_d$$

and $P_{d,\tau}$ is the transition matrix of a Markov chain with aperiodic recurrent classes.

Remark 2. *Since the state set and the action sets are those of the original problem the policy sets also coincide.*

For the sake of completeness we present the following results already proved in [12]. We want to highlight that these results are valid also for the multichain case.

Proposition 1. *Let us suppose that S finite, so that for each $(d)^\infty$, a stationary policy, P_d^*, the Cesàro limit matrix, is stochastic. Then it follows that*

$$g^{(d)^\infty}(s) = P_d^* r_d(s).$$

Proposition 2. *For every decision rule d*

$$P_{d,\tau}^* = P_d^* \text{ and } g_\tau^{(d)^\infty}(s) = g^{(d)^\infty}(s)$$

for every $s \in S$.

Proof. We shall use the fact that, since S is finite, given d any decision rule, the limit matrix $P_{d,\tau}^*$ is the unique one that satisfies

$$P_{d,\tau} P_{d,\tau}^* = P_{d,\tau}^* P_{d,\tau} = P_{d,\tau}^* P_{d,\tau}^* = P_{d,\tau}^*.$$

Note that $\forall \tau \in (0,1]$, we have

$$P_d^* P_{d,\tau} = (1-\tau) P_d^* I + \tau P_d^* P_d = P_d^*$$

and

$$P_{d,\tau} P_d^* = (1-\tau) I P_d^* + \tau P_d P_d^* = P_d^*.$$

By well-known results, described for example in [12, Appendix A, p. 595], it follows that $P_{d,\tau}^* = P_d^*$.
To establish the second equality, notice that, by Proposition 1, $g_\tau^{(d)^\infty} = P_{d,\tau}^* r_{d,\tau}$, so

$$g_\tau^{(d)^\infty} = P_{d,\tau}^* r_{d,\tau} = P_d^* r_d = g^{(d)^\infty}.$$

□

Corollary 2. *The optimal stationary policies d^* for the original problem and for the transformed one are the same. In addition, $g_\tau^{d^*}(s) = g^{d^*}(s)$ for all $s \in S$ and all $\tau \in (0,1]$.*

Proof. Similar to that of [12, Corollary 8.5.9]. □

Our contribution in this direction is to propose to use this transformation as a pre-processing of the problem in order to deal only with aperiodic models.

More precisely, we propose the following procedure. Consider a MDP with state set S, actions A_s for $s \in S$, transition probabilities $p(j|s,a)$ for $j, s \in S$, $a \in A_s$ and rewards $r(s,a)$, $s \in S$, $a \in A_s$. Being $\tau \in (0,1)$, transform the problem to a new one with $S_\tau = S$, $A_{s,\tau} = A_s$ for $s \in S_\tau$, $r_\tau(s,a) = r(s,a)$ $s \in S_\tau$, $a \in A_s$ and $p_\tau(\cdot|j,a)$ given by Equation (18).

Modified Rolling Horizon Procedure (MRH)

1. Given $0 < \tau < 1$, make the transformation described above.
2. Apply **RH** procedure to the new problem.

4.2 On the theoretical convergence of MRH

The **MRH** procedure can be applied to MDP whose optimal stationary policy gives rise to chains with several irreducible classes. Through the transformation at **Step 1**, every finite model becomes, in the most general case, a multichain aperiodic model. Since the number of states and actions are finite and due to aperiodicity, the model satisfies the condition of Schweitzer and Federgruen (Condition 1). Then, by Corollary 1, the geometric convergence of **RH** is assured. In the particular case where the model is unichain (Condition 4), a parameter associated to the contraction might be computed explicitly, but likely just in small or very simple examples.

4.3 Practical convergence of the MRH procedure

Clearly, the transformed transition matrix converges towards original transition matrix when $\tau \to 1$. How is the convergence of $v_n(s)/n$ toward $g^*(s)$ modified when τ increases? We have a quantitative look at the question with the two following examples.

Example 5. *Consider again the example with five states described in the Appendix. The optimal gain* $g^* = (2, 2, 4, 4, 4)$ *and it is produced by the stationary policy* $d = (a_2, a_2, a_1, a_2, a_1)$*,*

We have applied the τ*-transformation described above. With* $\tau = 0$ *the model is transformed in an uncoupled one where all the states are absorbing. As it is shown in Figure 2, when* τ *increases to 1 the periodicity effects are more evident.*

Example 6. *Consider again the example of the Appendix. This model has a multichain and periodic structure. The **RH** procedure applied directly on this problem produces infinitely (and periodically) many times two policies, one of them is not two policies,* $(a_2, a_2, a_1, a_1, a_1)$ *and* $(a_2, a_2,\ a_1, a_2, a_1)$*. The first one produces a gain* $g = (2, 2, 3, 3, 3)$ *and then it is not optimal since* $g^* = (2, 2, 4, 4, 4)$*.*

When we pre-process the data, the Rolling Horizon procedure gives the optimal policy for the original problem, for any value of $\tau \in (0, 1)$*.*

4.4 Preservation of Condition 5

Next we investigate the connection of the transformation with Condition 5. Does the transformation preserve this property? Can it be expected obtaining Condition 5 through this transformation? The examples below show that this transformation does not destroy but weakens Condition 5, and that it does not produce it necessarily when it is not initially present, even with uncontrolled Markov chains. This is possibly due to the known fact that Condition 5 is related to some ergodicity property of the Markov chain involved and the transformation proposed by Schweitzer does not produce ergodicity properties, only aperiodicity.

Condition 5 is preserved by the aperiodicity transformation

Theorem 4. *If a MDP model satisfies Condition 5, then its transformation according to Section 4.1 still satisfies this condition. The constant "*δ_τ*" of the transformed model can be chosen as:* $1 - \tau + \delta\tau$*.*

Proof. According to Lemma 1, Condition 5 holds for the MDP if and only if, for all $(s, s') \in S \times S$, all $a \in A_s$, all $a' \in A_{s'}$, $\text{supp}(p(\cdot|s, a)) \cap \text{supp}(p(\cdot|s', a')) \neq \emptyset$. From the transformation (18), it is easy to see that $\text{supp}(p_\tau(\cdot|s, a)) = \text{supp}(p(\cdot|s, a)) \cup$

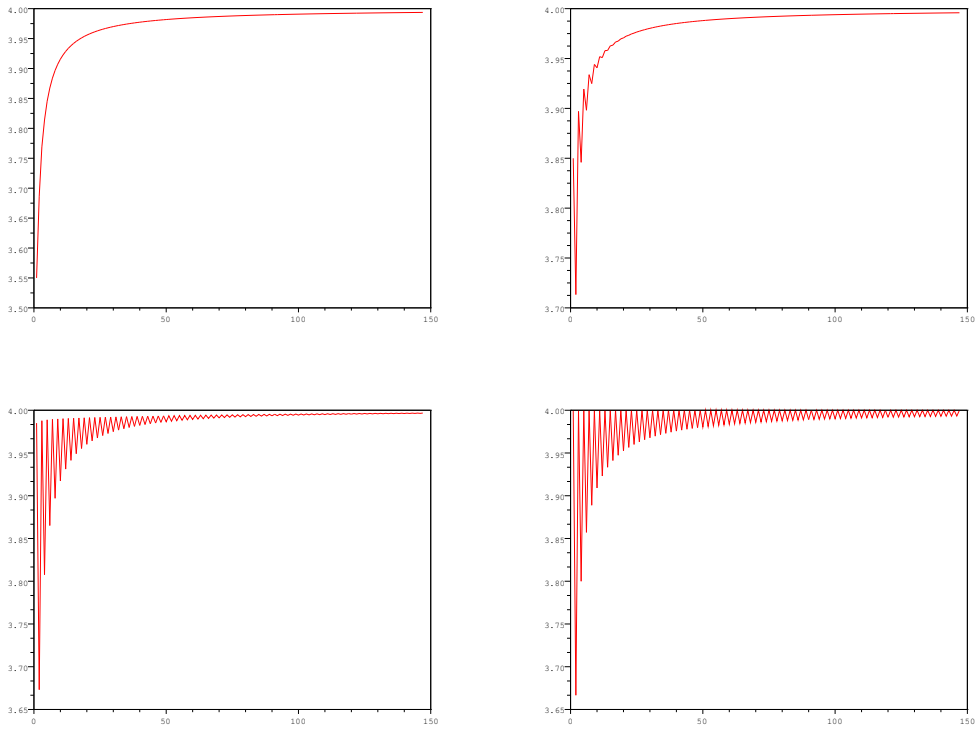


Figure 2: Evolution of $v_n(4)/n$ for $n = 1...150$, with $\tau = 0.7$ (top left), 0.9 (top right), 0.99 (bottom left) and 1 (bottom right)

$\{s\}$. Therefore, supp$(p(\cdot|s,a)) \cap$ supp$(p(\cdot|s',a')) \neq \emptyset$ implies supp$(p_\tau(\cdot|s,a)) \cap$ supp$(p_\tau(\cdot|s',a')) \neq \emptyset$, and Condition 5 holds also for the transformed MDP. We proceed with estimating the value of the constant "δ" corresponding to this transformed model.

Let us consider a MDP where Condition 5 is verified and a parameter τ such that $0 < \tau < 1$. Then, for all $B \subset S$, and all admissible pair (s,a), after transforming the problem we have the new transition probabilities p_τ defined by

$$p_\tau(B|s,a) = (1-\tau)\delta_s(B) + \tau p(B|s,a),$$

or equivalently

$$p(B|s,a) = \frac{1}{\tau} p_\tau(B|s,a) - \frac{1-\tau}{\tau}\delta_s(B).$$

Now, for all subsets $B_1, B_2 \subset S$,

$$\begin{aligned}
&\mathbf{sp}(p(\cdot|s,a) - p(\cdot|s',a')) \\
&\quad = \max_B (p(B|s,a) - p(B|s',a')) - \min_B (p(B|s,a) - p(B|s',a')) \\
&\quad \geq (p(B_1|s,a) - p(B_1|s',a')) - (p(B_2|s,a) - p(B_2|s',a')) \\
&\quad = \frac{1}{\tau} p_\tau(B_1|s,a) - \frac{1-\tau}{\tau}\delta_s(B_1) - \frac{1}{\tau} p_\tau(B_1|s',a') + \frac{1-\tau}{\tau}\delta_{s'}(B_1) \\
&\qquad - \frac{1}{\tau} p_\tau(B_2|s,a) + \frac{1-\tau}{\tau}\delta_s(B_2) + \frac{1}{\tau} p_\tau(B_2|s',a') - \frac{1-\tau}{\tau}\delta_{s'}(B_2) \; .
\end{aligned}$$

The value of the preceding expression depends on the facts that the states s and s' belong or not to the subsets B_1 and B_2. Since $0 \leq \delta_s(\cdot) \leq 1$ for any s, we have the following inequalities:

$$\begin{aligned}
&\mathbf{sp}(p(\cdot|s,a) - p(\cdot|s',a')) \\
&\quad \geq \frac{1}{\tau} p_\tau(B_1|s,a) - \frac{1-\tau}{\tau} - \frac{1}{\tau} p_\tau(B_1|s',a') - \frac{1}{\tau} p_\tau(B_2|s,a) + \frac{1}{\tau} p_\tau(B_2|s',a') - \frac{1-\tau}{\tau} \\
&\quad = \frac{1}{\tau} \left[p_\tau(B_1|s,a) - p_\tau(B_1|s',a') \right] - \frac{1}{\tau} \left[p_\tau(B_2|s,a) - p_\tau(B_2|s',a') \right] - 2\frac{1-\tau}{\tau} \; .
\end{aligned}$$

Then, as Condition 5 holds, there exists δ, $0 < \delta < 1$ such that, for any pair of subsets B_1 and B_2,

$$2\delta \geq \frac{1}{\tau} \left[p_\tau(B_1|s,a) - p_\tau(B_1|s',a') \right] - \frac{1}{\tau} \left[p_\tau(B_2|s,a) - p_\tau(B_2|s',a') \right] - 2\frac{1-\tau}{\tau}$$

and taking the maximum over B_1 and minimum over B_2 gives

$$2\delta \geq \frac{1}{\tau} \mathbf{sp} \left(p_\tau(\cdot|s,a) - p_\tau(\cdot|s',a) \right) - 2\frac{1-\tau}{\tau} \; ,$$

or equivalently:

$$\mathbf{sp} \left(p_\tau(\cdot|s,a) - p_\tau(\cdot|s',a) \right) \; \leq \; 2(1 - \tau + \delta\tau) \; .$$

As expected, Condition 5 holds with a constant $\delta_\tau = 1 - \tau + \delta\tau$. □

It is readily checked that $\delta < \delta_\tau < 1$ since $0 < \tau < 1$. Hence, the transformation makes Condition 5 weaker: the smaller τ is, the weakest is the condition.

To conclude this section, we provide two examples which show that the aperiodicity transformation may or may not produce Condition 5, even for uncontrolled unichain problems (multichain models remain multichain after transformation). It is known that this condition is related to the structure of the Markov chain and not to the policies considered.

Example 7. *This example is a case where Condition 5 appears after transformation.*

Let us consider a MDP with state space $S = \{s_1, s_2\}$, and one admissible action in each state, i.e. an (uncontrolled) Markov Chain, where the transition probabilities are given by the matrix

$$P_d = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} .$$

This is a periodic and unichain model, and then it does not verify Condition 5.

After apply the aperiodicity transformation we have the new transition probability matrix

$$P_{d,\tau} = \begin{pmatrix} 1-\tau & \tau \\ \tau & 1-\tau \end{pmatrix}.$$

It is not hard to see that Lemma 1 applies to this matrix: supp$(p(\cdot|s)) = \{s_1, s_2\}$ *for* $s = s_1, s_2$*. More precisely,*

$$\begin{aligned} p_\tau(s_1|s_1, d(s_1)) - p_\tau(s_1|s_2, d(s_2)) &= (1-\tau) - \tau = 1 - 2\tau \\ p_\tau(s_2|s_1, d(s_1)) - p_\tau(s_2|s_2, d(s_2)) &= \tau - (1-\tau) = 2\tau - 1 \\ \textbf{\textit{sp}}(p_\tau(.|s_1, d(s_1)) - p_\tau(.|s_2, d(s_2))) &= \begin{cases} 2-4\tau & \text{if } \tau \le 1/2 \\ 4\tau - 2 & \text{if } \tau \ge 1/2. \end{cases} \end{aligned}$$

Condition 5 therefore holds for the transformed model with a constant $\delta_\tau = |1-2\tau| < 1$.

Example 8. *In this example, Condition 5 does not appear after transformation.*

Let $S = \{s_1, s_2, s_3, s_4, s_5, s_6\}$*, and again consider just one action at each state. Transition probabilities are specified in the matrix*

$$P_d = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The chain is unichain and aperiodic and again Condition 5 does not hold. In fact we can see that as a consequence of Lemma 1, since there exist the pair $(s_1, s_4) \in S \times S$*, for which*

$$\textit{supp}(p(\cdot|s_1, d(s_1))) \cap \textit{supp}(p(\cdot|s_4, d(s_4))) = \{s_1, s_2\} \cap \{s_4, s_5\} = \emptyset\,.$$

Through transformation for some $\tau \in (0, 1)$*, we have the new transition matrix*

$$P_{d,\tau} = \begin{pmatrix} 1-\tau/2 & \tau/2 & 0 & 0 & 0 & 0 \\ 0 & 1-\tau & \tau & 0 & 0 & 0 \\ 0 & 0 & 1-\tau & \tau & 0 & 0 \\ 0 & 0 & 0 & 1-\tau/2 & \tau/2 & 0 \\ 0 & 0 & 0 & 0 & 1-\tau & \tau \\ \tau & 0 & 0 & 0 & 0 & 1-\tau \end{pmatrix}$$

where again the pair $(s_1, s_4) \in S \times S$*, gives us*

$$\textit{supp}(p_\tau(\cdot|s_1, d(s_1))) \cap \textit{supp}(p_\tau(\cdot|s_4, d(s_4))) = \{s_1, s_2\} \cap \{s_4, s_5\} = \emptyset$$

and Condition 5 does not hold.

5 On stopping rules

When a geometric convergence result exists, with a computable convergence bound, a simple stopping rule for **VI** is easily derived. Likewise, if **RH** converges geometrically

with a known rate, the value of the horizon n can be chosen so that the **RH** policy is ϵ-optimal.

Assume that some computable δ and C exist such that, for **VI**,

$$||ng^* - v_n - h^*|| \ < \ C\delta^n$$

or for the **RH** procedure,

$$||g^* - g^{(d_n)^\infty}|| \ < \ C\delta^n \ .$$

Then consider the following stopping rule (or horizon choice rule for **RH**):

Stopping Rule 1. *Stop if* $n > \log(\epsilon/C)/\log(\delta)$.

Obviously, when **VI** stops under this rule, the policy $(d_n, \ldots, d_1, d_n, \ldots)$ is ε-optimal, and when the time horizon for **RH** is chosen according to this rule, the policy $(d_n)^\infty$ is ε-optimal. We have discussed in Section 3.3 when this rule can be used in practice.

When no explicit convergence bound is known, the following practical convergence rule is proposed in [12].

Stopping Rule 2. *Stop if* $\boldsymbol{sp}(v_{n+1} - v_n) \leq \epsilon$.

For the aperiodic irreducible, unichain, communicating and weakly communicating models, or even if any of these properties is required only for optimal policies, all of them with g^* a constant vector, Stopping Rule 2 is adequate as proved in [12, Section 8.5.4, p. 370]. More precisely, it is proved that if Stopping Rule 2 applies, the value of the policy is ε-optimal. Indeed, the classical proof of this claim involves passing to the limit in the following inequalities

$$\min_{s\in S}(v_n - v_{n-1})(s) \leq g^{(d_n)^\infty}(s) \leq g^*(s) \leq \max_{s\in S}(v_n - v_{n-1})(s).$$

Clearly, $(d_n)^\infty$ will be ε-optimal for n large enough if

$$\lim_{n\to\infty} \max_{s\in S}(v_n - v_{n-1})(s) = \lim_{n\to\infty} \min_{s\in S}(v_n - v_{n-1})(s).$$

When we deal with arbitrary MDP's, we do not have *a priori* information about its structure and, in consequence we cannot guarantee that g^* is a constant vector. In this case, Stopping Rule 2 does not provide a suitable stopping criterion, since the span $\mathbf{sp}(v_{n+1} - v_n)$ fails to converges to zero.

Puterman [12, Section 9.4.2, p. 477] states a conjecture about a stopping rule for the multichain case. It involves the following rule.

Stopping Rule 3. *Stop if* $\boldsymbol{sp}(v_{n+1} - v_n) - \boldsymbol{sp}(v_n - v_{n-1}) \leq \epsilon$.

The next result shows that this rule not appropriate either.

Theorem 5. *For any* $\epsilon > 0$*, there exist MDP models such that, for the sequence* $\{v_n\}$ *obtained using the* ***VI*** *algorithm, it is possible to have, for some* $n \in \mathbb{N}$, $\boldsymbol{sp}(v_{n+1} - v_n) - \boldsymbol{sp}(v_n - v_{n-1}) < \epsilon$ *and* $\|g^* - g^{(d_n)^\infty}\| \geq \epsilon$.

Proof. We construct such a model as follows. Consider the two-state model $S = \{s_1, s_2\}$, $A_{s_1} = \{a_1^1, a_2^1\}$ y $A_{s_2} = \{a_1^2\}$, with transition probabilities

$$p(s_1|s_1, a_1^1) = 1; \ \ p(s_2|s_1, a_2^1) = 1; \ \ p(s_2|s_2, a_1^2) = 1$$

and gains

$$r(s_1, a_1^1) = 10; \;\; r(s_1, a_2^1) = 1; \;\; r(s_2, a_1^2) = 10 + \varepsilon.$$

In this model, there exist only two stationary policies: $d_1 = (a_1^1, a_1^2)$ and $d_2 = (a_2^1, a_1^2)$. As the gains that they produces are, respectively, $(10, 10+\varepsilon)$ and $(10+\varepsilon, 10+\varepsilon)$, the second policy is the optimal one.
The sequence of functions v_m obtained with the **VI** algorithm is such that, for $m \leq N(\varepsilon) := 9/\varepsilon + 1$, $v_{m+1} - v_m = v_m/m = \frac{1}{m}\sum_{k=1}^{m}(v_k - v_{k-1}) = g^{(d_m)^\infty} = (10, 10+\varepsilon)$.
On the other hand, as we have said, for any $\varepsilon > 0$, there exist $N(\varepsilon)$ (who tends to infinity as ε tends to zero), such that, if $m < N(\varepsilon)$ implies $g^*(s_1) - g^{(d_m)^\infty}(s_1) = \varepsilon > 0$. □

Actually, this example serves to show that a larger family of stopping rules, containing Stopping Rule 3, is not adequate.

In fact, consider the values $\alpha_m = v_m/m$, $\beta_m = v_{m+1} - v_m$, $\gamma_m = \frac{1}{m}\sum_{h=1}^{m} v_h - v_{h-1}$ and $\delta_m = g^{(d_m)^\infty}$, for $m = 1, ..., n$. Define the rule $A : (\mathbb{R}^{|S|})^{4n} \mapsto \mathbb{R}$,

$$A(x_1, x_2, ..., x_{4n}) \;=\; \sum_{i=1}^{4n}\sum_{j=1}^{4n} a_{ij}\, \mathbf{sp}(x_i - x_j)\;,$$

where the a_{ij} are arbitrary constants.

Theorem 6. *There exist finite-state, finite-action MDP models such that, for some n, $A(\alpha_1, \alpha_2, ..., \alpha_m, \beta_1, \beta_2, ..., \beta_m, \gamma_1, \gamma_2, ..., \gamma_m, \delta_1, \delta_2, ..., \delta_m) < \varepsilon$, and*

$$\|g^* - g^{(d_m)^\infty}\| \geq \varepsilon.$$

As a consequence of Theorem 5, Stopping Rule 3 cannot work in general, and Theorem 6 even shows that no immediate generalization will work either. No alternative seems to be known in the literature, and Kallenberg mentions in [9] that formulating a stopping rule for **VI** without chain analysis for multichain MDPs and provide a valid stopping rule remains an open problem.

As it is observed in [17, p. 208], in the proof of Theorem 3.4.1, for models which verify Condition 3 (under which g^* a constant vector) we obtain

$$\frac{\mathbf{sp}(v_n - v_{n-1})}{\min_{s\in S}(v_n - v_{n-1})} \leq \varepsilon \quad \Rightarrow \quad 0 \leq \frac{g^{(d_n)^\infty} - g^*}{g^*} \leq \varepsilon,$$

which induces the following

Stopping Rule 4. *Stop if* $\mathbf{sp}(v_{n+1} - v_n) \leq \varepsilon \; \min_{s\in S}(v_n - v_{n-1})$.

Observe that, in general, this is not an admissible rule, since it is not effective in the cases with non constant optimal rewards.

This stopping rule differs form the other ones in the sense that, in this case, when the **VI** algorithm algorithm terminates iterating, it returns a policy whose *relative* error is not greater than ε.

6 Conclusions

We have reviewed in this paper convergence conditions for the Value Iteration procedure, and applied them to the question of "convergence" of the Rolling Horizon procedure.

Our analysis concludes that Condition 1 is a sufficient condition to ensure geometric convergence of the **RH** procedure in finite models. Condition 5, proposed earlier in [7] is less general.

In addition, we introduce for the Rolling Horizon procedure a standard pre-processing of the problem for eliminating periodicities, resulting in the **MRH** procedure. We show that this transformation does not change the near optimal policies nor their values.

Theorem 2 claims that convergence is geometric, including for Multichain models. However, since a general stopping rule is not available, this result remains mostly *theoretical*. It remains to find an *adequate* bound for the error.

References

[1] J. M. Alden and R. L. Smith, "Rolling Horizon Procedures in Nonhomogeneous Markov Decision Processes", *Operations Research*, **40**(Supp. 2), S183–S194, 1992.

[2] D. P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models*, Prentice Hall, Englewood Cliffs, NJ, 1987.

[3] C. Derman, *Finite State Markovian Decision Processes*, Academic Press, N.Y., 1970.

[4] E. Çinlar, *Introduction to Stochastic Processes*, Prentice Hall Press, 1975.

[5] A. Federgruen, P. Schweitzer and C. Tijms, "Contraction mappings underlying undiscounted Markov decision problems", *Journal of Mathematical Analysis and Applications*, **65**, 711–730, 1978.

[6] X. Guo and P. Shi, "Limiting average criteria for non stationary Markov decision processes", *SIAM J. Optim.*, **11**(4), 1037–1053, 2001.

[7] O. Hernández-Lerma and J.B. Lasserre, "Error Bounds for Rolling Horizon Policies in Discrete-Time Markov Control Processes", *IEEE Transactions on Automatic Control*, **35**(10), 1118–1124, 1990.

[8] L. Kallenberg, *Finite state and action MDPS*, in *Handbook of Markov Decision Processes. Methods and applications*, E. Feinberg and A. Shwartz (Eds), Kluwer's international Series, 2002.

[9] L. Kallenberg, *Markov decision processes*, Lectures Notes, University of Leiden, 2009, in `www.math.leidenuniv.nl/~kallenberg/Lecture-notes-MDP.pdf`

[10] E. Lanery, "Etude asymptotique des systèmes markoviens à commande", *Rev.Informat. Recherche Operat.*, 1, 3-56, 1967.

[11] S.P. Meyn and R.L. Tweedie, *Markov chains and stochastic stability*, Cambridge University Press, Second Edition, 2009.

[12] L. Puterman, *Markov Decision Processes*, Wiley and Sons, 1994.

[13] S. M. Ross, *Applied Probability Models with Optimization Applications*, Holden-Day, 1970.

[14] P. J. Schweitzer, "Iterative solution of the functional equation of undiscounted Markov renewal programming", *Journal of Mathematical Analysis and Applications*, **34**, 495–501, 1971.

[15] P. J. Schweitzer and A. Federgruen, "The asymptotic behavior of undiscounted value iteration in Markov decision problems", *Mathematics of Operations Research*, **2**(4), 360–381, 1977.

[16] P. J. Schweitzer and A. Federgruen, "Geometric convergence of the value iteration in multichain Markov decision problems", *Adv. Appl. Prob.*, **11**, 188–217, 1979.

[17] H.C. Tijms, *Stochastic Modelling and Analysis, A Computational Approach*, Wiley, New York, 1986.

[18] D. J. White, *Markov decision processes*, John Wiley and sons, 1993.

Appendix

Each month an individual must decide how to allocate his wealth between different consumptions and investments. Each state represents a level of individual's wealth at the start of a month. Wealth levels give access to two different investment opportunities, prudent or risky. Choosing an investment profile at each level results in a probability transition for the next wealth level, as well as an instantaneous gain. The individual's objective is to maximize the average gain.

There are five levels of wealth, ordered from the smallest to the largest. At the medium level, connected to the risky behavior, there exists positive probability to pass to the next inferior level of wealth. It is also possible to cycle among the two inferior levels, but there is no action which permit the access to the three superior levels from the inferior ones. Besides, being at the poorest level, by some external help we achieve level 2. There is a common action space $A = \{a_1, a_2\}$, where a_1 represents the prudent investment profile and a_2 the risky attitude. We show the data below. $P_{a_k}(s, j)$ is the transition probability from the state s to state j when action a_k is used, i.e. $P_{a_k}(s, j) = p(j|s, a_k)$.

$$P_{a_1} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0.4 & 0.6 & 0 & 0 & 0 \\ 0 & 0 & 0.7 & 0.3 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \qquad P_{a_2} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

The gains can summarize as follows: $r(s, a_k)$ in the matrix below is the gain when at state s, the action a_k is chosen.

$$\begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 1 \\ 3 & 2 \\ 6 & 6 \end{pmatrix}.$$

Through the implementation of the **MRH** procedure the optimal average wealth can be computed: $g^* = (2, 2, 4, 4, 4)$. It is produced by the stationary policy associated to the decision rule $d = (a_2, a_2, a_1, a_2, a_1)$ whose transition matrix is:

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.7 & 0.3 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Clearly, it is a multichain periodic model. When **RH** procedure is applied directly, there is no convergence: the procedure gives infinitely (and periodically) many times two policies, $(a_2, a_2, a_1, a_1, a_1)$ and $(a_2, a_2, a_1, a_2, a_1)$. The first one produces a gain $g = (2, 2, 3, 3, 3)$ and then it is not optimal.



Centre de recherche INRIA Sophia Antipolis – Méditerranée
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399