When are the Value Iteration Maximizers Close to an Optimal Stationary Policy of a Discounted Markov Decision Process?

Closing the Gap between the Borel Space Theory and Actual Computations

RAÚL MONTES-DE-OCA
Departamento de Matemáticas
Universidad Autónoma Metropolitana-Iztapalapa
San Rafael Atlixco 186, 09340 México D.F.
MÉXICO
momr@xanum.uam.mx
Corresponding Author

ENRIQUE LEMUS-RODRÍGUEZ
Escuela de Actuaría
Universidad Anáhuac México-Norte
Av. Universidad. Anáhuac 46,
Col. Lomas Anáhuac
Huixquilucan
C.P. 52786, Edo.de México
MÉXICO
elemus@anahuac.mx

Abstract: Markov Decision Processes [MDPs] have been repeteadly used in Economy and Engineering but apparently are still far from achieving its full potential due to the computational difficulties inherent to the subject due to the usual impossibility of finding explicit optimal solutions. Value iteration is an elegant, theoretical method of approximating an optimal solution, frequently mentioned in Economy when MDPs are used. To extend its use and benefits, improved understanding of its convergence is needed still even if it would appear not to be the case. For instance, the corresponding convergence properties of the policies is still not well understood. In this paper we further analyze this issue: using Value Iteration, if a stationary policy f_N is obtained in th N-th iteration, such that the optimal discounted rewards of f^* and f_N are close, we would like to know whether are the corresponding actions $f^*(x)$ and $f_N(x)$ necessarily close for each state x? To our knowledge this question is still largely open. In this paper it is studied when it is possible to stop the value iteration algorithm so that the corresponding maximizer stationary policy f_N approximates an optimal policy both in the total discounted reward and in the action space (uniformly over the state space). In this article the action space is assumed to be a compact set and the reward function bounded. An ergodicity condition on the transition probability law and a structural condition on the reward function are needed. Under these conditions, an upper bound on the number of steps needed in the value iteration algorithm, such that its corresponding maximizer is a uniform approximation of the optimal policy, is obtained.

Key–Words: Markov decision process, compact action space, bounded reward, expected total discounted reward, approximation of optimal policies by means of value iteration policies

1 Introduction

Markov Decision Processes [MDPs] are frequently mentioned in the engineering and economic literature, lets quote Ada and Cooper *Dynamic Economics: Quantitative Methods and Applications*, MIT Press, 2003, p. 7:

> The mathematical theory of dynamic programming as a means of solving dynamic optimization problems dates to the early contributions of Bellman (1957) and Bertsekas (1976). For economists, the contributions of Sargent (1987) and Stokey and Lucas (1989) provide a valuable bridge to this literature.

Nevertheless, the wider application of this techniques seems to be way below its actual potential. This is the case partly due to the fact that it is often virtually impossible to find an explicit optimal stationary policy f^* with an optimality criterion as the total discounted reward case.

Bellman himself was well aware of this situation, and, in *Dynamic Programming and Modern Control Theory*, Academic Press, 1965, writes with his coauthor Robert Kalaba, p. vii:

> The processes studied in physics, engineering, economics, biology, and operations research possess a bewildering array of special features (. . .). It is essential for the suc-

> cessful analyst to know how to incorporate any of all of these features, (. . .). Furthermore, he must have an awareness of the capabilities and limitations of modern computing machines and of the interfaces between theoretical formulation and numerical solution.

Nowadays the computational power far exceeds that of the machines Bellman used, but apparently his advised has been not often followed. The lack of the *interfaces* Bellman mentions is partially responsible for this situation.

In this case, dynamic programming is embodied in Value iteration, a popular algorithm for theoretically finding an optimal policy. Its simplicity, elegance and relation to Banach's fixed point theorem, makes it an excellent choice, under convenient circumstances, that constitutes a feasible procedure for obtaining, in the N-th iteration, a stationary policy f_N that approximates the optimal policy f^* in the sense that their total discounted rewards are close.

For applications it would be important to know if indeed this policies are close, but the above mentioned fact does not necessarily imply that in each state x the corresponding actions $f^*(x)$ and $f_N(x)$ are close. To the best of our knowledge, this question is not only still largely open but also deserves more attention: it is important to further theoretical research in MDPs that may have a positive impact in practical computational matters that may foster their further application. In this paper the problem is fully stated in Section 2 in the context of a larger research program, ,and in Section 3 the solution is presented with full detail and with some relevant remarks: Basically it is studied the problem of determining when it is possible to stop the value iteration algorithm so that the corresponding maximizer stationary policy f_N approximates an optimal one both in the action space and in the total discounted reward. In fact, under the conditions to be described in the problem statement in Section 2, these policies could be stable in the following sense: choosing an action close to $f_N(x)$ while in the state x would still provide a useful approximate policy (which constitutes a matter of further research). It is important to stress once more that this kind of results shed light on important computability issues of great practical interest. In fact, they are related to the interesting Theorem 6.8.1, p. 218 and Corollary 6.8.4 quoted by M. L. Puterman in his classic work in [11], where a lower bound is found on the number of steps necessary to stop Value Iteration and obtain an optimal policy for finite state and finite action MDPs. Henceforth both these results combined will be called the Planning Horizon Theorem [PHT]. Such a result suggests some interesting problems: how it is possible to establish similar results for the Borel case. In Section 2 we will see how this question is addressed. The solution presented in Section 3 relates the lower bound to the value function of the problem. Finally, in the last Section the concluding remarks on the conditions are given and further future research is discussed.

Lets stress then that it would be desirable to follow Bellmans advise, and in this particular case, close the gap between the abstract Borel state space theory and actual computations. This paper, hopefully, is a small step in that direction.

2 Problem Formulation

The control problem in this paper is stochastic. Why should incur in such a complication? Let's recall, quoting Ada and Cooper *Dynamic Economics: Quantitative Methods and Applications*, MIT Press, 2003, p. 29, that:

> While the nonstochastic problem may be a natural starting point, in actual applications it is necessary to consider stochastic elements. The stochastic growth model, consumption/savings decisions by households, factor demand by firms, pricing decisions by sellers, search decisions, all involve the specification of dynamic stochastic enviroments.

A Markov Decision Process [MDP] is a flexible tool for the above mentioned purpose that is used in applications both in engineering and economics as both a modelling and optimization framework whose potential has not been yet fully explored. In particular, the total discounted reward case frequently appears in that kind of applications. But, it would seem that notwithstanding its mathematical elegance and its wide applicability as a tool, MDPs are not used as often they should, one reason beeing the heavy computational burden regarding the actual solution of the optimization problem.

As it is well known, iterative procedures as Value or Policy iteration present an elegant mathematical framework where a sequence (indirectly or directly) of stationary policies f_N is obtained, that under suitable conditions, converge to an f^*, an optimal stationary policy [3]. As Value Iteration is related to Banach's Fixed Point Theorem and Policy Iteration to Newton-Raphson-Kantorovich's Method (under suitable conditions), there is considerable purely theoretically research regarding this algorithms. Nevertheless, some of this research, as it stands, is not suitable for a more direct actual use. For instance. sometimes the hypothesis under which convergence of the stationary policies f_N hold can not be directly verified from the description of the model, i.e., its dynamics and reward function. That is, the hypothesis are not **structural**, they are not directly stated in terms of the actual structure of the MDP. This has to be stressed: if the conditions can not be verified directly from the structure of the systm, how could it be possible to analyse the system without actually simulating or experimentally observing it with the corresponding extra cost?

On the other hand, many of those results disregard the actual convergence of the sequence f_N of stationary policies, for instance, in Value Iteration, the case dealt with in this paper. This may happen because, apparently, if the value function of an approximating stationary policy f_{ap} to the optimal value V^*, then it would be irrelevant whether $f_{ap}(x)$ is actually close to $f^*(x)$ (the optimal stationary policy). Nevertheless, in the actual implementation, the action

$$f_{ap} + \epsilon,$$

(where ϵ is an enviromental error) is the one that will be actually used, and hence the corresponding value no longer needs to be close to the optimal. It is important to further explore this problem, that in a sense is a stability and robustness one. One important small step in this direction would be to understand the asymptotic behaviour of the sequence f_N of stationary policies, what, to our knowledge, is often ignored in the literature.

As it appears, this happens because it is not always clear that there exists a unique optimal policy or, on the contrary, there are at least two different optimal policies and consequently classical convergence of the f_N may not hold. But recent work has established some important cases where from uniqueness of the optimal policy [2], not only convergence of this sequence f_N can be established, but even a stopping rule can be stated. In plain words, such a research helps answer the question: when are the Value Iteration maximizers close to an optimal stationary policy, and how much?

The problem solved in this paper is now stated:

> **Problem:**
>
> To find a meaningful family of Markov Decision Processes such that given a positive tolerance ϵ it is possible to find a bound on the number of steps that Value Iteration needs so that its corresponding stationary policy $f_N(x)$ is within tolerance from an optimal policy $f^*(x)$, that is, the distance between $f_N(x)$ and $f^*(x)$ is less than ϵ uniformly on the state space X of the Markov Decision Process. In particular, this family of Markov Decision Processes should include Borel State and Action spaces.

The last requirement will hopefully allow some elegant but theoretical work in Borel space MDPs to be extended in a compatible fashion with numerical mathematics and consequently applied in concrete cases. In the solution to be presented in the next section an important case is studied: the reward is strictly concave. This clearly is relevant as covers a situation important when the reward is a utility function and hence its concavity allows for risk aversion or the law of diminishing rewards, etc ...

Giving an explicit and structural (this has to be stressed) set of conditions that solve the problem in a Borel space setting partially answers how close are the maximizer to an optimal policy and indeed helps to close the gap between the elegant Borel space theory and actual computational issues.

Let's recall that due to the discount, it is plausible that a stationary optimal policy for a finite horizon problem should be close to the optimal one in infinite horizon for a sufficiently large finite horizon. That would be the heuristic motivation of the Planning Horizon approach: find stationary solutions to an increasing sequence of control horizons T_n and establish the corresponding convergence of their values to the optimal value of the infinite horizon problem. If we are given a set of conditions that ensure that convergence, we get a Planning Horizon Theorem (PHT).

The statement and solution of our problem is inspired by Planning Horizon approach, adapted to the Borel case, with the extra improvement that the corresponding bound on the planning horizon does not depend explicitly on the value function of the corresponding Markov Decision Process, but depends on information easy to deduce directly from its structure. In particular, the present paper deals with infinite action spaces.

The most recent antecedent on this topic is [10],which deals with discounted MDPs with a Borel state space and finite action sets, and under suitable versions of the Ergodicity Condition (EC) and the Individuality Condition (IC), stated in section 3, it is established that the corresponding policy f_n obtained in the n-th step of the Value Iteration Algorithm (VIA) for some n is indeed an optimal stationary policy. Note that the approximation result presented here is different from the one above in which the approximation is estimated by means of the difference between the expected total discounted reward of a maximizer of the VIA and the optimal value function (see Theorem 3.1 in [7]). In this paper we are able to provide insight on when the value iteration maximizer policies are close enough to an optimal one pointwisely on the action space.

3 Problem Solution

Due to the complexity of the Borel state space setting in Markov Decision Processes, this section will be partitioned in order to somehow ease its heavy technical burden. So, this section is divided as follows:

- **The Control Model**, where the model is explicitly presented with detail.
- **Conditions**, where they are presented and discussed.
- **The Actual Solution**, where the theorem that solves the problem stated in the previous section, and its proof are presented.
- **Concavity**, where the case of strictly concace reward is studied, due to its importance in economy and engineering.
- **Remarks on the Conditions**, where many technical details of the solution and their minutiae are discussed.

3.1 The Control Model

The standard Borel model is analyzed in this paper.

Let $(X, A, \{A(x) : x \in X\}, Q, r)$ be the basic discrete-time Markov decision model (see [6], [11]), which consists of the state space X , the action space A, the transition law Q, and the reward function r. Both X and A are assumed to be Borel spaces ($\mathbb{B}(X)$ and $\mathbb{B}(A)$ denote the Borel sigma-algebras of X and A, respectively). Furthermore, for every $x \in X$ there is a nonempty set $A(x) \in \mathbb{B}(A)$ whose elements are the admissible actions when the state of the system is x. In this article it is assumed that, for every state $x, A(x) = A$, and also that A is a *compact* set containing more than one element (as the case with one element considered is trivial). The transition law $Q(B|x, a)$, where $B \in \mathbb{B}(X)$ and $(x, a) \in X \times A$, is a stochastic kernel on X, given $X \times A$. The reward function $r(\cdot, \cdot)$ is a nonnegative, upper bounded (by a bound denoted by M), and a measurable function on $X \times A$.

Let Π be the set of all (possibly randomized, history-dependent) admissible policies (see [6], [11], for details). By standard convention, a *stationary* policy is taken to be a measurable function $f : X \to A$. The set of the stationary policies is denoted by $\mathbb{F}$.

For every $\pi \in \Pi$ and state $x \in X$, let

$$V(\pi, x) = E_x^\pi \left[\sum_{t=0}^{+\infty} \alpha^t r(x_t, a_t) \right] \tag{1}$$

be the *expected total discounted reward*. The number $\alpha \in (0, 1)$ is called the *discount factor*. In our case, there will be some restrictions on α to be explicitly stated later (see Remark 3). Here, $\{x_t\}$ and $\{a_t\}$ denote the state and the control sequences, respectively, and E_x^π is the corresponding expectation operator with respect to the probability measure P_x^π defined on the space $\Omega := (X \times A)^\infty$ in a canonical way. The *optimal control problem* is to find a policy π^* such that $V(\pi^*, x) = \sup_{\pi\in\Pi} V(\pi, x)$, for all $x \in X$, in which case π^* is said to be *optimal*.

As usual, the *optimal value function* is defined as

$$V^*(x) = \sup_{\pi\in\Pi} V(\pi, x), \tag{2}$$

$x \in X$.

Notation 1

- **(a)** Denote the metric in A by d, and for $\epsilon > 0$, $x \in X$ and $a^* \in A$, let $B_\epsilon(a^*) := \{a \in A : d(a, a^*) < \epsilon\}$, and $B_\epsilon^c(a^*) := \{a \in A : d(a, a^*) \geq \epsilon\}$. Observe that $B_\epsilon^c(a^*)$ is a (possibly empty) compact set.
- **(b)** Let γ be the diameter of the set A, i.e., $\gamma := \sup\{d(z, w) : z, w \in A\}$. (Note that since A is a compact set, $\gamma < \infty$ and there exist $b, c \in A$ such that $\gamma = d(b, c)$; moreover, since A contains more than one element, then $\gamma > 0$.)

Lemma 1. *For each* ϵ *such that* $0 < \epsilon < \gamma/2$, *and* $a^* \in A, B_\epsilon^c(a^*) \neq \phi$.

Proof. Let b and c be as in Notation 1 (b). Fix $a^* \in A$ and $\epsilon \in (0, \gamma/2)$. Then $d(b, a^*) \geq \gamma/2$ or $d(c, a^*) \geq \gamma/2$ (otherwise, if $d(b, a^*) < \gamma/2$ and $d(c, a^*) < \gamma/2$, the triangle inequality implies that $\gamma = d(b, c) \leq d(b, a^*) + d(c, a^*) < \gamma/2 + \gamma/2 = \gamma$, which is a contradiction). Hence $d(b, a^*) \geq \epsilon$ or $d(c, a^*) \geq \epsilon$, i.e., $B_\epsilon^c(a^*) \neq \phi$. □

The following assumption will define in part the class of Markov Decision Processes where the problem is solved.

When does Assumption 1 hold?

For that, see the end of this subsection.

Assumption 1

(a) The optimal value function V^*, defined in (2), satisfies the *Dynamic Programming Equation* (DPE), i.e. for all $x \in X$,

$$V^*(x) = \max_{a \in A} \left[r(x,a) + \alpha \int V^*(y) Q(dy|x,a) \right]. \tag{3}$$

There also exists $f^* \in \mathbb{F}$ such that:

$$V^*(x) = r(x, f^*(x)) + \alpha \int V^*(y) Q(dy|x, f^*(x)), \tag{4}$$

$x \in X$, and f^* is optimal.

(b) The *Value Iteration Algorithm* is valid. That is, the *value iteration functions* inductively defined as

$$v_n(x) = \max_{a \in A} \left[r(x,a) + \alpha \int v_{n-1}(y) Q(dy|x,a) \right], \tag{5}$$

$x \in X$ and $n = 1, 2, \ldots$, with $v_0 = 0$, are well-defined, and for each $x \in X, v_n(x) \to V^*(x)$. Besides, for each $n = 1, 2, \ldots$,there exist $f_n \in \mathbb{F}$ such that, for each $x \in X$,

$$v_n(x) = r(x, f_n(x)) + \alpha \int v_{n-1}(y) Q(dy|x, f_n(x)). \tag{6}$$

Let

$$G(x,a) := r(x,a) + \alpha \int V^*(y) Q(dy|x,a),$$

$(x,a) \in X \times A$ (note that $G(x, f^*(x)) = V^*(x), \ x \in X$).

Remark 1 Using (4) and the fact that $\sup_{x \in X} |V^*(x) - v_n(x)| \leq (\alpha^n M)/(1-\alpha)$, for all $n \geq 1$, it is easy to verify that for each $x \in X$ and $n = 1, 2, \ldots$,

$$|G(x, f_n(x)) - G(x, f^*(x))| \leq (2M\alpha^n)/(1-\alpha). \tag{7}$$

Let $\Phi(X) = \{$real-valued bounded measurable functions on $X\}$.

The *span* of a function $\psi \in \Phi(X)$ is defined by $sp(\psi) := \sup_{x \in X} \psi(x) - \inf_{x \in X} \psi(x)$.

Remark 2
(a) Observe that $-sp(V^*) \geq -M(1-\alpha)$.

(b) In the proof of Lemma 3.5, p. 59 in [6], it is obtained that for any (x,a) and (x',a') in $X \times A$ and any $\psi \in \Phi(X)$,

$$\int \psi(y) Q(dy|x,a) - \int \psi(y) Q(dy|x',a') \leq \lambda sp(\psi). \tag{8}$$

The reader may legitimately ask:

When does Assumption 1 hold?

A classical set of conditions, stated in p. 18 [6] do the trick:

- For each state $x \in X$, the set $A(x)$ of admissible controls is (non-empty) compact subset of A.
- For some constant M,

  $$|r(k)| \leq M,$$

  for all $k = (x,a), x \in X, a \in A(x)$, and moreover, for each $x \in X$, $r(x,a)$ is a continuous function of $a \in A(x)$.
- And:

  $$\int u(y) Q(dy|x,a)$$

  is a continuous function of $a \in A(x)$ for each $x \in X$ and each bounded and measurable function u from X to the set of real numbers.

It is to be stressed that all of the above properties are **structural**, that is, can be verified directly from the control model

$$(X, A, \{A(x) : x \in X\}, Q, r).$$

3.2 Conditions

It will clear in the proof ot the theorem that solves our problem, that some technical comparison between integrals is needed. Some minimal ergodicity allows us to make that comparison, and hence, the following condition will be decisive:

Ergodicity Condition (EC)

There exists a number $\lambda < 1$ such that

$$\sup_{k,k'} \|Q(\cdot|k) - Q(\cdot|k')\|_V \leq 2\lambda, \tag{9}$$

where the sup is taken over all $k, k' \in X \times A$, and $\|\cdot\|_V$ denotes the variation norm for signed measures.

For $x \in X$, and $\epsilon > 0$ such that $B^c_\epsilon(a^*) \neq \phi$, define

$$D_{x,\epsilon} := \inf_{a^* \in A} \inf_{a \in B^c_\epsilon(a^*)} |r(x,a) - r(x,a^*)|,$$

auxiliary expression needed in the statement of the:

> **Individuality Condition (IC)** There exist ϵ and D_ϵ such that $0 < \epsilon < \gamma/2$ and $D_\epsilon > 0$, and $D_{x,\epsilon} \geq D_\epsilon$, for each $x \in X$.

Remark 3 The IC determines the admissible values for the parameter α. Let $H(\alpha) = \alpha/(1-\alpha), \alpha \in (0,1)$.Then, from $D_\epsilon > 0$ and solving $H(\alpha) < D_\epsilon/(\lambda M)$, where λ is given in (9) and M is an upper bound for r, it follows that $\alpha \in (0, H^{-1}(D_\epsilon/(\lambda M))$ (note that $H(\cdot)$ is increasing on (0,1)). In fact, throughout the present article, a fixed value of α in this interval is considered. Observe that hence

$$K^*_\epsilon := D_\epsilon - (\alpha\lambda M)/(1-\alpha) > 0. \qquad (10)$$

As α may not be close to 1, further research will be needed if techniques like the vanishing discount approach are used to obtain a PHT-like results in the average reward case.

Lemma 2. *Suppose that Assumption 1, the EC and the IC hold. Then, for each $x \in X$,*

$$\inf_{a \in B^c_\epsilon(f^*(x))} |G(x,a) - G(x,f^*(x))| \geq K^*_\epsilon. \qquad (11)$$

Proof Take $x \in X$. Then using Remark 2 it follows that

$$\begin{aligned}
&\inf_{a \in B^c_\epsilon(f^*(x))} |G(x,a) - G(x,f^*(x))| \\
&\quad \geq \inf_{a^* \in A} \inf_{a \in B^c_\epsilon(a^*)} |G(x,a) - G(x,a^*)| \\
&\quad = \inf_{a^* \in A} \inf_{a \in B^c_\epsilon(a^*)} \left[\begin{array}{c} |r(x,a) - r(x,a^*) + \\ \alpha \int V^*(y) Q(dy|x,a) - \\ \alpha \int V^*(y) Q(dy|x,a^*)| \end{array} \right] \\
&\quad \geq \inf_{a^* \in A} \inf_{a \in B^c_\epsilon(a^*)} \left[\begin{array}{c} |r(x,a) - r(x,a^*)| - \\ \alpha| \int V^*(y) Q(dy|x,a^*) - \\ \int V^*(y) Q(dy|x,a)| \end{array} \right] \\
&\quad \geq \inf_{a^* \in A} \inf_{a \in B^c_\epsilon(a^*)} [|r(x,a) - r(x,a^*)| - \alpha\lambda sp(V^*)] \\
&\quad \geq \inf_{a^* \in A} \inf_{a \in B^c_\epsilon(a^*)} \left[\begin{array}{c} |r(x,a) - r(x,a^*)| - \\ (\alpha\lambda M)/(1-\alpha) \end{array} \right] \\
&\quad \geq D_\epsilon - (\alpha\lambda M)/(1-\alpha) = K^*_\epsilon. \quad \square
\end{aligned}$$

Remark 4 In the finite action case, (11) would imply the uniqueness of f^* (see [10]), but in the present situation that would not be necessarily the case. On the other hand, different optimal policies should be ϵ-pointwise close.

3.3 The actual solution

Now it is possible to state and prove a Theorem that constitutes a solution to the problem, and is indeed a small step to closing the gap between the elegant but highly theoretical research in the Borel state framework in MDPs and actual computations.

So, the maximizers eventually get pointwise close to the optimal stationary policy, in other words:

Theorem 1
Suppose that Assumption 1, the EC and the IC hold. Let ϵ be a positive number that satisfies IC. Let $N(\epsilon) = [(\ln((1-\alpha))K^*_\epsilon/2M)/\ln\alpha] + 1$ (here $[z]$ is the integer part of z). Then $d(f_{N(\epsilon)}(x), f^*(x)) < \epsilon$, for all $x \in X$.

Proof Let x be a fixed state. Firstly, denote

$$\delta := \inf_{a \in B^c_\epsilon(f^*(x))} |G(x,a) - G(x,f^*(x))|.$$

(Observe that from (10) and (11), $\delta \geq K^*_\epsilon > 0$.) Consequently, if $d(a, f^*(x)) \geq \epsilon$, then

$$|G(x,a) - G(x,f^*(x))| \geq \delta \geq K^*_\epsilon, \text{ i.e.,}$$

if

$$|G(x,a) - G(x,f^*(x))| < K^*_\epsilon, \qquad (12a)$$

then

$$d(a, f^*(x)) < \epsilon. \qquad (12b)$$

Secondly, choose the minimal positive integer $N(\epsilon)$ such that,

$$(2M\alpha^{N(\epsilon)})/(1-\alpha) < K^*_\epsilon. \qquad (13)$$

Now, it follows from (7), (12a), (12b) and (13) that $N(\epsilon) = [(\ln((1-\alpha))K^*_\epsilon/2M)/\ln\alpha] + 1$, and $d(f_{N(\epsilon)}(x), f^*(x)) < \epsilon$. Since x is arbitrary, the result follows. □

Remark 5

a Note tha if A is a finite set, then $\gamma = 1$, and since $0 < \epsilon < \gamma/2 = 1/2$, Theorem 1 implies that

$$f_{N(\epsilon)}(x) = f^*(x),$$

for $x \in X$.

Recall that in this case, the metric used is the discrete metric, that is $d(z,w)$ is 0 if $z = w$, otherwise is 1.

b In Theorem 3.1(a) of [7] it is proved that for each positive integer n,

$$|V^*(x) - V(f_n, x)| \leq \frac{M\alpha^n}{1-\alpha},$$

for $x \in X$.

3.4 Concavity

As mentioned in the second section, it is important to know if the previous solution holds in at least one case where the reward function is strictly concave. As a matter of fact, that is the case. Two partial solutions are more or less straighforward under the two following sets of assumptions. Both of them imply the lemma in this subsection.

The following set is inspired in some previous research in uniqueness of optimal policies [2].

Assumption 2

a) X is a convex subset of $\mathbb{R}$, and A is a closed interval $[w, z]$, $w, z \in \mathbb{R}, w < z$.

b) Q is induced by a difference equation:

$$x_{t+1} = F(x_t, a_t, \xi_t),$$

$t = 0, 1, \ldots,$ where $F : X \times A \times S \to X$ is a measurable function, and $\{\xi_t\}$ is a sequence of i.i.d. random variables with values in $S \subseteq \mathbb{R}$. In addition, it is supposed that $F(\cdot, \cdot, s)$ is a convex function on $X \times A$, and if $x < y$, then $F(x, a, s) \leq F(y, a, s)$ for each $a \in A$ and $s \in S$.

c) r is strictly concave on $X \times A$, and if $x < y$, then $r(x, a) \geq r(y, a)$ for each $a \in A$.

The next set of assumptions is an alternative way of obtaining uniqueness, see [2].

Assumption 3

a) Same as Assumption $2(a)$.

b) Q is given by the relation

$$x_{t+1} = \rho x_t + \sigma a_t + \xi_t,$$

$t = 0, 1 \ldots,$ where $\{\xi_t\}$ is a sequence of i.i.d. random variables with values in $S \subseteq \mathbb{R}$, and where ρ and σ are real numbers.

c) r is strictly concave on $X \times A$.

Lemma 3. *Suppose that Assumption 1 holds. Then each Assumption* 2 *or* 3 *implies that for each* $x \in X, G(x, \cdot)$ *is strictly concave, for each* $\psi > 0$ *there exists a closed interval* $J = J(x, \psi)$ *and* $\tau = \tau(x, \psi) > 0$ *such that* $a \in J$ *and* $|G(x, a) - G(x, f^*(x))| < \tau$ *implies that* $|a - f^*(x)| < \psi$*, and* f^* *is unique.*

Proof

The proofs of the strictly concavity of $G(x, \cdot)$, $x \in X$, and the uniqueness of the optimal policy f^* are similar to the proofs of Lemmas 6.1 and 6.2, and Theorem 3.4(i) in [2]. Now, fix $x \in X$. Consider $G(x, a)$, $a \in (w, z)$.

Hence, in that interval, $G(x, \cdot)$ is a continuous function (see Theorem 3, p.113 in [1]), and since $G(x, \cdot)$ is strictly concave, $f^*(x) \in (w, z)$. Take w' and z' such that $w < w' < f^*(x) < z' < z$.

From Lemma 4.48, p.202 in [13], it follows that $G(x, \cdot)$ is strictly increasing in $[w', f^*(x)]$, and strictly decreasing in $[f^*(x), z']$. Firstly, consider $G(x, a)$, $a \in [w', f^*(x)]$. From Proposition 2.18, p. 80 in [8], it results that $G(x, a)$, $a \in [w', f^*(x)]$ has an inverse function which is also continuous and increasing. In particular, this inverse function is right-continuous in $G(x, f^*(x))$, i.e. given $\psi > 0$, there is $\tau_1 > 0$ such that $G(x, a) \in [G(x, w'), G(x, f^*(x)]$ and $|G(x, a) - G(x, f^*(x))| < \tau_1$ implies that $|a - f^*(x)| < \psi$ or equivalently, given $\psi > 0$ there is $\tau_1 > 0$ such that $a \in [w', f^*(x)]$ and $|G(x, a) - G(x, f^*(x))| < \tau_1$ implies that $|a - f^*(x)| < \psi$.

Secondly, in a similar way, it is possible to obtain that for each $\psi > 0$ there exists $\tau_2 > 0$ such that $a \in [f^*(x), z')]$ and $|G(x, a) - G(x, f^*(x))| < \tau_2$ implies that $|a - f^*(x)| < \psi$.

Finally,taking $J = [w', z']$, and $\tau = min\{\tau_1, \tau_2\}$, and since x is arbitrary, Lemma 3 follows. □

This theoretical version of the solution of our problem is in form and spirit very close to results stated by Stokey and Lucas in his work [12] for a general abstract optimization problem. This Lemma would be then a non-trivial translation (as this result in **not** a particular case of Stokey and Lucas) to the Borel MDPs that would deserve further attention.

3.5 Remarks on the Conditions

Some technical remarks are in order.

Remark 6 In [6] pp. 56-60 (see, also [7] p. 1123) several sufficient conditions for the EC are presented. In particular, Condition 3.1 (1) states : there exists a state $x^* \in X$ and a positive number β such that $Q(\{x^*\}|x, a) \geq \beta$ for all $(x, a) \in X \times A$ (see [6] p. 56), which implies the EC, and holds in the following two examples (see examples 1 and 2 below).

PHT-like results are useful in MDPs related to Economics and Finance models, as many researchers from these fields need explicit computing procedures to find approximations of optimal policies. In these cases the reward function is usually a utility function. A simple relevant example of such a possible application is presented:

Example 1 (Adapted from example 3.1 in [5]) A household with monthly income u (a positive constant) has to decide on the proportion $\bar{\lambda}$ (in (0,1)) of its current wealth x_t to be consumed. The remaining wealth $(1-\bar{\lambda})x_t$ is to be invested in a security with random return rate ξ_t. In this case it is possible to take the state space as $X = [u,\infty)$ and the action space as $A = [0,1]$. The transition probability law is given by

$$x_{t+1} = \xi_t(1-\bar{\lambda})x_t + u, \tag{14}$$

$t = 0, 1, \ldots,$ where $\xi_0, \xi_1, \ldots$ are i.i.d. nonnegative random variables. Let ξ be a generic element of the sequence $\{\xi_t\}$.Suppose that due to severe economic conditions $P[\xi = 0] = \kappa > 0$, and, for each $B \in \mathbb{B}((0,\infty)), P[\xi \in B] = \int_B \eta(z)dz$, where $\eta : \mathbb{R} \to \mathbb{R}$ is a measurable and nonnegative function such that $\int_{(0,\infty)} \eta(z)dz = 1-\kappa$. Here, using (14), it is direct to prove that $Q(\{u\}|x,a) = P[\xi = 0] = \kappa$ for all $(x,a) \in X \times A$. Hence Condition 3.1 (1) in [6] holds with $\beta = \kappa$. In this case $r(x,\bar{\lambda}) = U(\bar{\lambda}x)$, where U is a utility function, that is, the household reward is the utility of the consumption $\bar{\lambda}x$. Most utility functions easily satisfy Assumption 2 below and hence this model would satisfy all the Assumptions of this article.

Example 2 Let η and $n > 1$ be a positive constant and a positive integer, respectively. Consider the transition probability law induced by a system equation of the type

$$x_{t+1} = \min\{[x_t + a_t - \xi_t]^+, \eta\}, \tag{15}$$

$t = 0, 1, \ldots,$ where $\xi_0, \xi_1, \ldots$ are i.i.d. random variables with values in $S = [0,\infty)$, and $[z]^+$ denotes the positive part of z. Here the control space is $A = [\eta+1, \eta+n]$, $X = [0,\eta]$. It is not difficult to verify that under the condition:

$$P[\xi_0 > 2\eta + n] > 0,$$

implies that the contidion 3.1 in [6] holds with $\beta = P[\xi_0 > 2\eta + n] > 0$.

Remark 7 It is worth mentioning that in some very important cases IC holds independently of x, for instance, when there are two functions Ψ and $\Lambda, \Psi : X \to \mathbb{R}$ and $\Lambda : A \to \mathbb{R}$ such that $r(x,a) = \Psi(x) + \Lambda(a), (x,a) \in X \times A$ (notice that in this case $|r(x,a) - r(x,a^*)| = |\Lambda(a) - \Lambda(a^*)|$ for $x \in X, a^* \in A$, and $a \in B^c_\epsilon(a^*)$).

The IC can be verified in the special case described in Lemma 4 below.

Assumption 4

(a) $A \subseteq \mathbb{R}$;
(b) For each $x \in X$, there exists an open set I such that $A \subseteq I$, and $r(x,\cdot)$ is defined and is of class C^1 on I. Moreover, there exists a positive constant θ such that $r_a \geq \theta$, for all $x \in X$ (r_a denotes the first derivative of r with respect to the variable a).

Lemma 4. *Under Assumption 4, $D_\epsilon = \epsilon\theta$. Hence, IC holds for each ϵ which satisfies $0 < \epsilon < \gamma/2$.*

Proof. For $x \in X$, $a^* \in A$, $a \in B^c_\epsilon(a^*)$, and using the classical one-variable Mean Value Theorem,

$$\begin{aligned} |r(x,a) - r(x,a^*)| &= |r_a(x,a')||(a-a^*)| \\ &\geq \theta|a-a^*| \qquad (16) \\ &\geq \epsilon\theta, \qquad (17) \end{aligned}$$

where a' is a point between a and a^*. Hence,

$$\inf_{a^* \in A} \inf_{a \in B^c_\epsilon(a^*)} |r(x,a) - r(x,a^*)| \geq \epsilon\theta,$$

for each $x \in X$, i.e., $D_\epsilon = \epsilon\theta$. □

Remark 8
(a) Notice that if inequality (16) is valid for all $a, a^* \in A$, then $r(x,\cdot)$ is injective for each $x \in X$. Let n and m be positive integers $(m < n)$. Inequalities of type (17) applied to a continuous function $W : \mathbb{R}^n \to \mathbb{R}^n$ (that is, for some positive constant θ, $\|W(s) - W(z)\| \geq \theta\|s - z\|$, for all s, z in $\mathbb{R}^n$; here $\|\cdot\|$ denotes the usual norm in $\mathbb{R}^n$) have been used to study the existence of an inverse function for W (see, [4] p. 12, problem 71 and [14] p. 105, problem 5 (c)). Moreover, it is interesting to note that in the case of a continuous function $W : \mathbb{R}^n \to \mathbb{R}^m$, there is not an inverse for W (see, [9] p. 368 exercise 2).
(b) It is important to observe that neither the C^1 property (or the continuity property) nor the existence of an inverse for a reward function is necessary for a reward function to satisfy the IC.

For an example,for X a Borel space, take $A = A(x) = [0,1] \times [0,1/4]$ and let $r : X \times A \to \mathbb{R}$ be defined, for each $x \in X$, by $r(x,a) = 0$ if $a \in \Theta_0 = [0,1/4) \times [0,1/4], r(x,a) = 1$ if $a \in \Theta_1 = [1/4,2/4) \times [0,1/4], r(x,a) = 2$, if $a \in \Theta_2 = [2/4,3/4) \times [0,1/4]$, and $r(x,a) = 3$ if $a \in \Theta_3 = [3/4,1] \times [0,1/4]$.

Note that, in this example, for each $x \in X$,$\Gamma_h = \{a \in A : r(x,a) \geq h\}$ is closed for every $h \in \mathbb{R}$. In fact, $\Gamma_h = [0,1] \times [0,1/4]$ if $h \leq 0$, $\Gamma_h = [1/4,1] \times [0,1/4]$ if $0 < h \leq 1$, $\Gamma_h = [2/4,1] \times [0,1/4]$ if $1 < h \leq 2$, $\Gamma_h = [3/4,1] \times [0,1/4]$ if $2 < h \leq 3$, and $\Gamma_h = \phi$ if $h > 3$. Hence, $r(x,\cdot)$ is upper semicontinuous, for each $x \in X$.

This upper semicontinuity of $r(x, \cdot)$ for each $x \in X$ is necessary for Assumption 1 to be valid (see [6] pp.18-19). Secondly, observe that $\gamma/2 = (17)^{1/2}/8$ (in fact,γ is equal to the distance between the points $P(0, 0)$ and $Q(1, 1/4)$). Take $\epsilon = 1/2$ (note that $0 < 1/2 < (17)^{1/2}/8$). Now, take $x \in X$, and $a^* \in A$. Without loss of generality assume that $a^* \in \Theta_0$ (the proof of the other cases is similar). It is easy to verify that $\Theta_0 \subset B_{1/2}(a^*)$, or equivalently, $B^c_{1/2}(a^*) \cap A \subset \Theta_1 \cup \Theta_2 \cup \Theta_3$ (observe that the diameter of each square $[0, 1/4] \times [0, 1/4], [1/4, 2/4] \times [0, 1/4], [2/4, 3/4] \times [0, 1/4]$ and $[3/4, 1] \times [0, 1/4]$ is $(2)^{1/2}/4$, and that $(2)^{1/2}/4 < 1/2 < (17)^{1/2}/8$). Consequently, $\inf_{a \in B^c_{1/2}(a^*)} |r(x, a) - r(x, a^*)| \geq 1$. Since a^* and x are arbitrary, it follows that $D_{1/2} = 1$.

4 Conclusion

4.1 Further research

PHT-like results as those developed in this article deserve further attention for their possible application in practical computing schemes and in stability issues related to small errors in optimal actions. For instance, if in order to find an approximation to an optimal policy by digital computer, it is necessary to discretize the action space A or select a finite subset of “adequate” representatives of A, PHT-like results could suggest desirable properties of the resulting new finite action so that the Value iteration procedure would be more efficient.

Let it be mentioned that the present article has its roots in previous research on pointwise convergence of value iteration maximizers to the optimal policy, see [3], work that was partially inspired by some theoretical material in [12].

Finally, some open questions should be mentioned.

- In the case the reward does not depend on the state x, but only on the action a, as often happens in many economical applications, it should be straightforward, from the main result, to obtain uniform convergence of the maximizers. In particular, if the Assumptions in Lemma 3 hold and the transition probability law is also independent of the state of the system, does the sequence of maximizers $\{f_n\}$ converges uniformly to f^* on each closed interval $J' \subset (w, z)$ with $f^*(x) \in J'$ for each $x \in X$?
- It would be very interesting to weaken the **Ergodicity Condition (EC)** needed in the proof of Theorem 1. In particular, that would help to extend the results of the present paper to deterministic frameworks.
- Bounds of the kind:

  $$|V^*(x) - V(f_n, x)| \leq \frac{M\alpha^n}{1 - \alpha},$$

  $x \in X, n = 1, 2, 3, \ldots$, mentioned in Remark 5 may have an interpretation as stability or robustness results worth studying.

Recently, the study of problems in the mathematical modelling and optimization of sustainable development strategies has revived interest in the study of Average Reward Markov Decision Processes, as a matter of fact, a Discounted Reward Markov Decision approach would guarantee a non-sustainable solution, in the sense that first generations in a community would tend to earn all possible reward as soon as possible leaving a compromised future for posterior generations. It is natural to ask in this case if a vanishing discount approach would be feasible in the extension of the results of this paper to the average reward case [6]. The results presented in this paper impose some restrictions on the discount factor with respect to the ergodicity index, and hence this issue is open.

4.2 Final words

Bellman, in his seminal work *Dynamic Programming*, Princeton University Press, 1957, p. **ix**, states, regarding the solution of a problem in dynamic programming (and hence, perfectly relevant to the contents of this paper):

> *The problem is not to be considered solved*
> *in the mathematical sense until the structure*
> *of the optimal policy is understood.*

Much further research needs to be done in this subject if we are to follow Bellmans mandate.

References

[1] J. P. Aubin, *Applied Abstract Analysis*, Wiley, New York, 1977.

[2] D. Cruz-Suárez, R. Montes-de-Oca and F. Salem-Silva, Conditions for the Uniqueness of Optimal Policies of Discounted Markov Decision Processes, *Mathematical Methods of Operations Research*, Vol. 60, 2004, pp. 415-436.

[3] D. Cruz-Suarez and R. Montes-de-Oca, Uniform Convergence of the Value Iteration Policies for Discounted Markov Decision Processes, *Boletín de la Sociedad Matemática Mexicana*, Vol. 12, 2006, pp. 133-148.

[4] B. Gelbaum, *Problems in Analysis*, Springer-Verlag, New York, 1982.

[5] E. Gordienko, E. Lemus-Rodríguez and R. Montes-de-Oca, Discounted Cost Optimality Problem: Stability with Respect to Weak Metrics, *Mathematical Methods of Operation Research*, Vol. 68, 2008, pp. 77-96.

[6] O. Hernandez-Lerma, *Adaptive Markov Control Processes*, Springer-Verlag, New York, 1989.

[7] O. Hernandez-Lerma and J.B. Lasserre, Error Bounds for Rolling Horizon Policies in Discrete-Time Markov Control Processes, *IEEE Transactions on Automatic Control*, Vol. 35, 1990, pp. 1118-1124.

[8] G. Klambauer, *Aspects of Calculus*, Springer-Verlag, New York, 1986.

[9] S. Lang, *Analysis I*, Addison-Wesley, Massachusets, 1971.

[10] R. Montes-de-Oca, E. Lemus-Rodríguez and D.Cruz-Suárez, A Stopping Rule for Discounted Markov Decision Processes with Finite Action Sets. *Kybernetika (Prague)*, Vol 45, 2009, pp. 755-767 , 2009.

[11] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, New York, 1994.

[12] N. L. Stokey and R. E. Lucas, *Recursive Methods in Economics Dynamics*, Harvard University Press, USA, 1989.

[13] K.L.Stromberg, *An Introduction to Classical Real Analysis*, Wadsworth International Mathematical Series, Belmont, Cal., 1981.

[14] V. Zorich, *Mathematical Analysis II*, Springer, Germany, 2004.