Stochastic Optimal Control for Small Noise Intensities: The Discrete-Time Case

HUGO CRUZ-SUÁREZ
Universidad Autónoma de Puebla
Facultad de Ciencias Físico-Matemáticas
Av. San Claudio y 18 Sur, Puebla
MÉXICO
hcs@fcfm.buap.mx

ROCIO ILHUICATZI-ROLDÁN
Universidad Autónoma de Puebla
Facultad de Ciencias Físico-Matemáticas
Av. San Claudio y 18 Sur, Puebla
MÉXICO
rroldan@alumnos.fcfm.buap.mx

Abstract: This paper deals with Markov Decision Processes (MDPs) on Borel spaces with an infinite horizon and a discounted total cost. It will be considered a stochastic optimal control problem which arises by perturbing the transition law of a deterministic control problem, through an additive random noise term with coefficient epsilon. In the paper, we will analyze the behavior of the optimal solution (optimal value function and optimal policy) of the stochastic system, when the coefficient epsilon goes to zero. Specifically, conditions given in the paper guarantee the uniform on compact sets convergence of both the optimal value function and the optimal policy of the stochastic system to the optimal value function and the optimal policy of the deterministic one, when epsilon goes to zero, respectively. Finally, two examples which illustrate the developed theory are presented.

Key–Words: Stochastic Optimization, Markov Decision Process, Dynamic Programming, Total Discounted Cost, Deterministic Approximation, Inventory/Production System

1 Introduction

This paper will deal with discrete-time Markov Decision Processes (MDPs) with an infinite horizon and a total discounted cost (see [2], [10] and [11]). MDPs are widely used to model controlled dynamical systems in control theory, operations research, image fusion (see [3]), artificial intelligence (see [3], [15] and [19]) and others. Besides, the MDPs in question possess the objective function known as the discounted cost function.The principal goal of MDPs is to determine the optimal policy f and to obtain the optimal value function V. One widely studied methodology to characterize and determine f and V is called the dynamic programming equation (see [10] and [11]).

In this article, for the MDPs taken into account, there will be assumed the existence of optimal policies and the validity of the Dynamic Programming Equation (see [2] and [11]). Besides, the MDPs in question posses, possibly unbounded cost functions.

There will be considered a deterministic Markov Decision Process (MDP), and a family of the stochastic MDPs indexed by a coefficient ε with values in a certain compact set of real numbers containing zero, and for each element of this family the probability law is the transition law of the deterministic MDP perturbed by an additive random noise multiplied by ε. It will be interesting to analyze the behavior of the optimal value function and the optimal policy of the stochastic system, when the coefficient ε goes to zero.

Specifically, conditions given in the paper guarantee the uniform on compact sets convergence of both the optimal value function and the optimal policy of the stochastic MDP to the optimal value function and the optimal policy of the deterministic one, when ε goes to zero, respectively.

This article was inspired by the papers of Flemming (see [8]), Lipster, Runggaldier and Taksar (see [16]), and Cruz-Suárez & Montes-de-Oca (see [5]). The first article is the one related to the theory of small disturbances for problems of control in continuous time with a finite horizon. In this paper this approach is used to obtain expansions of a optimal value function of a stochastic MDP in powers $\varepsilon, \varepsilon^2, ...$ This work was one of the pioneers regarding the analysis of problems of control with small disturbances. Nowadays, this approach has been applied to models of economic growth (see [6], [7], [17] and [20]) and in asymptotic methods (see [13] and [14]). Lipster et al's paper deals with a stochastic control system in continuous time with a finite horizon and with nonnegative costs. In [16] the stochastic problem is approximated by a deterministic system when the noise intensitiy ε is small. In the paper of Cruz-Suárez & Montes-de-Oca (see [5]), a stochastic control system via a deterministic one is analyzed. In this case, the solution of the stochastic system (the optimal value function and the optimal policy) is induced by the determinis-

tic control system.

This paper is organized as follows. In Section 2, basics concepts and results in the theory of MDPs are presented. In Section 3, the statement of the problem is established and some results of Lipschitz functions are presented. In Section 4, the theory about the uniform convergence is developed. Finally, in Section 5, two examples which illustrate the developed theory are presented, and, in Section 6, some concluding remarks are provided.

2 Discounted Markov Decision Processes

A *Markov control model* (see [2], [10] and [11]) is a five-tuple $(X, A, \{A(x) : x \in X\}, Q, c)$ consisting of

(a) a Borel space X, called the *state space*;

1. a Borel space A, called the *action set*;

(b) a family $\{A(x) : x \in X\}$ of nonempty measurable subsets $A(x)$ of A, where $A(x)$ denotes the set of *feasible actions* when the system is in state $x \in X$. The set $\mathbb{K} = \{(x, a) : x \in X, a \in A(x)\}$ of admissible state-actions pairs is assumed to be a measurable subset of the cartesian product $X \times A$;

(c) a stochastic kernel Q on X given $\mathbb{K}$ called the *transition law*. Specifically, $Q(\cdot|x, a)$ is a probability measure on X for every $(x, a) \in \mathbb{K}$, and $Q(B|\cdot)$ is a measurable function on $\mathbb{K}$ for every $B \in \mathcal{B}(X)$ ($\mathcal{B}(X)$ denotes the Borel $\sigma-$algebra of X);

(d) c is a real-valued measurable function on $\mathbb{K}$ called the *cost-per-stage* (or one-stage cost) *function*.

For each $t = 0, 1, ...$, let x_t and a_t be the state and the control at time t, respectively. If the system is in the state $x_t = x$ at time t and the control action $a_t = a \in A(x)$ is applied, then a cost $c(x, a)$ is paid and the system moves to a new state x_{t+1} by means of the probability distribution $Q(\cdot\,|\,x, a)$ on X (i.e. $Q(B\,|x, a) = \Pr(x_{t+1} \in B\,|x_t = x, a_t = a)$, $B \in \mathcal{B}(X)$ and $(x, a) \in \mathbb{K}$)

In this article, the transition law is specified by a dynamic model of the form

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

$t = 0, 1, ...$, where the random perturbations $\{\xi_t\}$ is a sequence of independent and identically distributed (i.i.d.) random elements with values in some nonempty Borel subset of an Euclidean space S and a common distribution μ. Meanwhile, $M : \mathbb{K} \times S \to X$ is a measurable function. In this case, the transition law Q is given by

$$\begin{aligned} Q(B\,|x, a) &= \Pr(x_{t+1} \in B\,|x_t = x, a_t = a) \\ &= \mu\left(\{s \in S : M(x, a, s) \in B\}\right), \end{aligned}$$

for all $B \in \mathcal{B}(X)$ and $(x, a) \in \mathbb{K}$.

A control policy π is a sequence $\{\widehat{\pi}_t : t = 0, 1, ...\}$, where, for each $t = 0, 1, ...$, $\widehat{\pi}_t(\cdot\,|h_t)$ is a conditional probability on the Borel $\sigma-$algebra $\mathcal{B}(A)$, given the history $h_t := (x_0, a_0, ..., x_{t-1}, a_{t-1}, x_t)$, such that $\widehat{\pi}_t(A(x_t)\,|h_t) = 1$. In this paper, the set of policies will be denoted by Π.

Let $\mathbb{F} := \{\theta : X \to A$ such that θ is measurable and $\theta(x) \in A(x), x \in X\}$. A sequence $\pi = \{\theta_t : t = 0, 1, ...\}$ of functions $\theta_t \in \mathbb{F}$ is called a *Markov policy*. A Markov policy $\pi = \{\theta_t : t = 0, 1, ...\}$ is said to be a *stationary policy* if $\theta_t = \theta \in \mathbb{F}$, for all t.

Given the initial state $x_0 = x$, $x \in X$, and any policy $\pi \in \Pi$, there is a probability measure P_x^π induced by the pair (x, π) on the space $\Omega = (X \times A)^\infty$, with $\mathcal{F}$ as the product sigma-algebra, in a canonical way (see [10]). The corresponding expectation operator will be denoted by E_x^π. The pair (x, π) determines a stochastic process $(\Omega, \mathcal{F}, P_x^\pi, \{x_t\})$ called a *Markov Decision Process* (MDP).

Let $(X, A, \{A(x) : x \in X\}, Q, c)$ be a fixed control model. For each policy π and initial state $x \in X$, consider

$$v(\pi, x) = E_x^\pi \left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t)\right].$$

$v(\pi, x)$ from the equation above is called the *total expected discounted cost*, where $\alpha \in (0, 1)$ is the discount factor.

The *optimal control problem* is then to find a policy $\pi^* \in \Pi$, such that

$$v(\pi^*, x) = \inf_{\pi \in \Pi} v\left(\pi, x\right),$$

$x \in X$, and in this case π^* is called an *optimal policy*. The function $\widehat{v}$ defined by

$$\widehat{v}(x) = v(\pi^*, x),$$

$x \in X$, is called the *optimal value function*.

The *value iteration functions* are defined as

$$V_n(x) = \min_{a \in A(x)} \left\{c(x, a) + \alpha \int V_{n-1}(M(x, a, s))d\mu(s)\right\}, \tag{1}$$

$x \in X$, and $n = 1, 2, ...$, with $V_0\left(\cdot\right) \equiv 0$.

Let $w : X \to [1, \infty)$ be a measurable function. If m is a real-valued function on X, then its $w-$*norm* is defined as

$$\|m\|_w := \sup_{x \in X} \frac{|m(x)|}{|w(x)|}.$$

(w is called a *weight function*.)

Assumption I

(a) $A(x)$ is a compact subset for each $x \in X$.

(b) $c(x,a)$ is lower semicontinuous (l.s.c.) in $a \in A(x)$ for each $x \in X$. (i.e. for each $x \in X$, $\liminf_{n\to\infty} c(x,a_n) \geq c(x,a)$, for any sequence $\{a_n\}$ in A that converges to a.)

(c) For each $x \in X$, the function

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

is continuous in $a \in A(x)$, for every bounded measurable function u on X.

(d) There exist nonnegative constants r and β, with $1 \leq \beta < 1/\alpha$, and a weight function $w \geq 1$ on X such that, for every state $x \in X$,

i) $\sup_{a\in A(x)} |c(x,a)| \leq rw(x)$ and

ii) $\sup_{a\in A(x)} \int w(y)Q(dy\,|x,a) \leq \beta w(x)$.

(e) For every state $x \in X$, the function

$$w'(x,a) = \int w(y)Q(dy\,|x,a)$$

is continuous in $a \in A(x)$.

Lemma 1. *Suppose that Assumption I holds. Then:*

(a) The optimal value function $\widehat{v}$ is a solution for the following equation:

$$\widehat{v}(x) = \min_{a\in A(x)} \left\{ c(x,a) + \alpha \int \widehat{v}(y)Q(dy\,|x,a) \right\},$$

$x \in X$. (This equation is called the Dynamic Programming Equation.)

(b) There exists $\theta \in \mathbb{F}$ such that

$$\widehat{v}(x) = c(x,\theta(x)) + \alpha \int \widehat{v}(y)Q(dy\,|x,\theta(x)), \tag{2}$$

$x \in X$, and θ is optimal.

(c) $V_n(x) \to \widehat{v}(x)$, when $n \to \infty$ for each $x \in X$.

Remark 2. *The proof of the previous Lemma could be consulted in [11] pp. 51-53. In particular, in this source the following inequality is proved,*

$$|\widehat{v}(x)| \leq \frac{r}{1-\gamma} w(x), \tag{3}$$

$x \in X$ and $\gamma := \alpha\beta$. Inequality (3) will be used in the following sections.

Consider a deterministic MDP with a space state X, a space control A, with admissible sets $A(x) \subset A$, $x \in X$. Suppose that the dynamic of the system is given by the difference equation

$$x_{t+1} = F(x_t, a_t), \tag{4}$$

$t = 0, 1, ...$, where $F : \mathbb{K} \to X$ is a given measurable function, and the cost function $c : \mathbb{K} \to \mathbb{R}$ is measurable as well. In this case, Q is given by

$$Q(B\,|x,a) = I_B(F(x,a)),$$

for all $B \in \mathcal{B}(X)$ and $(x,a) \in \mathbb{K}$, where $I_B(\cdot)$ denotes the indicator function of B. The transition law of the deterministic problem will be denoted by Q_F. Then the Markov control model is given by $(X, A, \{A(x) : x \in X\}, Q_F, c)$.

Remark 3. *Observe that, when F is a continuous function on $\mathbb{K}$, the transition law Q_F is weakly continuous, i.e.*

$$\int u(y)Q_F(dy\,|x,a) = u(F(x,a))$$

is a continuous function of $(x,a) \in \mathbb{K}$ for every $u \in \{\zeta : X \to \mathbb{R} : \zeta$ is a bounded continuous function$\}$.

Assumption II

(a) Same as Assumption I (a). Moreover, the multifunction $x \longmapsto A(x)$ is upper semicontinuous (u.s.c.).

(b) The cost function c is l.s.c. on $\mathbb{K}$.

(c) F is a continuous function on $\mathbb{K}$.

(d) There exist nonnegative constants r_d and β_d with $1 \leq \beta_d < 1/\alpha$, and a continuous weight function $w_d \geq 1$ on X such that for every state $x \in X$:

i) $\sup_{a\in A(x)} |c(x,a)| \leq r_d w_d(x)$ and

ii) $\sup_{a\in A(x)} w_d(F(x,a)) \leq \beta_d w_d(x)$.

Remark 4. *Under Assumption II a similar version of Lemma 1 holds for a deterministic MDP (see [11], pp. 65-67). (In this case, in Lemma 1 it is just necessary to change the transition law Q for Q_F.)*

3 Statement of the Problem

Let $(X, A, \{A(x) : x \in X\}, Q_F, c)$ be a deterministic Markov control model as introduced in the previous section. Besides, consider a stochastic control system with the same: state space X, control space A, admissible sets $A(x)$, $x \in X$, and the cost function c, but with the following dynamic of the system:

$$x_{t+1} = F(x_t, a_t) + \varepsilon\xi_t, \tag{5}$$

$t = 0, 1, ...$, where $\{\xi_t\}$ is a sequence of i.i.d. random elements taking values in a Borel space $S \subseteq X$ with a distribution function μ and $\varepsilon \in \Xi$, where Ξ is a compact subset of the real numbers containing zero. Note that in this case a family of Markov control models indexed by ε: $\{(X, A, \{A(x) : x \in X\}, Q_\varepsilon, c) : \varepsilon \in \Xi\}$ is considered, where the transition law Q_ε is given by

$$Q_\varepsilon(B\,|x, a) = \int I_B(F(x, a) + \varepsilon s)d\mu(s),$$

$B \in \mathcal{B}(X)$ and $(x, a) \in \mathbb{K}$.

Remark 5. *Observe that in the stochastic transition law (5), when $\varepsilon \to 0$, the stochastic system goes to the deterministic system (4). The rest of the paper will be focused on analyzing conditions which allow both the optimal value function and the optimal policy of the stochastic MDP tend to the corresponding optimal value function and the optimal policy of the deterministic MDP.*

In the next sections, the value function of the stochastic system will be denoted by V^ε and the deterministic system by V. In the same way, the optimal policy of the stochastic system will be denoted by f^ε and the deterministic optimal policy by f. Moreover, it will be supposed that Assumptions I or II hold for each MDP considered. (The weight functions will be denoted by w and w_ε for the deterministic MDP and for the stochastic one, respectively, and similarly for V_n and V_n^ε.) Also, there will be assumed the existence of a unique stationary optimal policy f for the deterministic control system (see [4]).

Let $(\widehat{E}, \widehat{d})$ be a metric space. If B is a subset of $\widehat{E}$ and $a \in \widehat{E}$, it is defined that

$$\widehat{d}(a, B) = \inf\left\{\widehat{d}(a, b) : b \in B\right\}.$$

Let B_1 and B_2 be two nonempty closed subsets of $\widehat{E}$. Define

$$\widehat{d}(B_1, B_2) = \sup\left\{\widehat{d}(b, B_2) : b \in B_1\right\},$$

and

$$\widehat{d}(B_2, B_1) = \sup\left\{\widehat{d}(b, B_1) : b \in B_2\right\}.$$

The function

$$H(B_1, B_2) = \max\left\{\widehat{d}(B_1, B_2), \widehat{d}(B_2, B_1)\right\} \tag{6}$$

where B_1 and B_2 are nonempty closed subsets of $\widehat{E}$, is called the *Hausdorff metric*. It has the properties of a metric on the family of nonempty closed subsets of $\widehat{E}$.

In the rest of this Section it will be supposed that d_1 and d_2 denote, respectively, the metrics on X and A. Besides, let d be the metric on $\mathbb{K}$ defined by $d := \max\{d_1, d_2\}$.

Assumption III

There is $L_1 > 0$ such that

$$H(A(x), A(x')) \leq L_1 d_1(x, x'),$$

for every x and x' in X where H is the Hausdorff metric (6).

Lemma 6. *Under Assumption III, the multifunction $x \to A(x)$ is u.s.c..*

Proof. Let $\{x_n\} \subset X$ be a sequence such that $x_n \to x$, $x \in X$ and let $\{a_n\}$ be a sequence of elements of $A(x_n)$, $n \geq 1$. Using Assumption III, it results that

$$H(A(x_n), A(x)) \leq L_1 d_1(x_n, x),$$

$n \geq 1$. Then, when $n \to \infty$ in the last inequality and using the definition of H, it is obtained that $d_2(A(x_n), A(x)) \to 0$. In particular,

$$\lim_{n\to\infty} d_2(a_n, A(x)) = 0. \tag{7}$$

Fix $n \geq N$. For each $k \geq 1$, there exists $b_k^n \in A(x)$ such that

$$d_2(a_n, A(x)) + \frac{1}{k} > d_2(a_n, b_k^n). \tag{8}$$

Since $A(x)$ is a compact subset of A, there exists a subsequence $\{b_{k_l}^n\}$ of $\{b_k^n\}$ and $b^n \in A(x)$ such that $b_{k_l}^n \to b^n$. Then, substituting k by k_l and letting $l \to \infty$ in (8) it results that

$$d_2(a_n, A(x)) \geq d_2(a_n, b^n). \tag{9}$$

Using a similar argument (now, considering n as variable), there exist a subsequence $\{b^{n_z}\}$ and $a \in A(x)$, such that $b^{n_z} \to a$ when $z \to \infty$. Then, by (9) and (7) it is obtained that

$$\lim_{z\to\infty} d_2(a_{n_z}, b^{n_z}) = 0.$$

On the other hand,

$$d_2(a_{n_z}, a) \leq d_2(a_{n_z}, b^{n_z}) + d_2(b^{n_z}, a).$$

Therefore, when $z \to \infty$, $a_{n_z} \to a$. Now, from Lemma 2.20 in [3], it follows that $x \to A(x)$ is u.s.c.. This concludes the proof of Lemma 6. □

Now, let $G : \mathbb{K} \to \mathbb{R}$ be a measurable function. It is supposed that there is a function $\Lambda : X \to \mathbb{R}$ such that $G(x,a) \geq \Lambda(x)$, for all $x \in X$ and $a \in A(x)$. Define $g : X \to \mathbb{R}$ by

$$g(x) = \inf_{a \in A(x)} G(x,a), \tag{10}$$

$x \in X$.

The following Lemma is similar to Proposition 24 in [1]. The proof of this Lemma is presented here for the completeness of the paper.

Lemma 7. *Suppose that Assumptions I a) and III hold. Let $G : \mathbb{K} \to \mathbb{R}$ be a Lipschitz function on $\mathbb{K}$ with a Lipschitz's constant L. Then g given by (10) is a Lipschitz function on X with a Lipschitz's constant $L \max\{L_1, 1\}$.*

Proof. It will be known that due to Assumption I a), Lemma 6, and the fact that G is Lipschitz there exists $h \in \mathbb{F}$ such that $g(x) = G(x, h(x))$, $x \in X$ (see Proposition D.3, Appendix D in [10]). Fix $x, x' \in X$. Using Assumption III, it results that

$$\inf\left\{d_2(h(x), a) : a \in A(x')\right\} \leq L_1 d_1(x, x').$$

Then, there exists $a' \in A(x')$ such that

$$d_2(h(x), a') \leq L_1 d_1(x, x'). \tag{11}$$

Therefore,

$$\begin{aligned} g(x') - g(x) &= g(x') - G(x, h(x)) \\ &\leq G(x', a') - G(x, h(x)) \\ &\leq L \max\left\{d_1(x, x'), d_2(h(x), a')\right\} \end{aligned}$$

$$= \begin{cases} L d_1(x, x'), \text{ if } d_1(x, x') \geq d_2(h(x), a') \\ L d_2(h(x), a'), \text{ if } d_1(x, x') < d_2(h(x), a') \end{cases}. \tag{12}$$

Now, using (11) in (12), it is obtained that

$$g(x') - g(x) \leq L \max\{L_1, 1\} d_1(x, x').$$

In a similar way it is possible to demonstrate that

$$g(x') - g(x) \leq L \max\{L_1, 1\} d(x, x').$$

Therefore, since x and x' are arbitrary the result follows. □

Now the following assumption will be presented.

Assumption IV

There are constants L_0 and L_2 such that:

(a) $|c(k) - c(k')| \leq L_0 d(k, k')$ for every k and k' in $\mathbb{K}$.

(b) $|F(k) - F(k')| \leq L_2 d(k, k')$ for every k and k' in $\mathbb{K}$.

Now, let $G : \mathbb{K} \to \mathbb{R}$ be a function defined by

$$G(k) = c(k) + \alpha \int V^{\varepsilon}(F(k) + \varepsilon s) d\mu(s), \tag{13}$$

$k = (x, a) \in \mathbb{K}$. And for each n, let $G_n : \mathbb{K} \to \mathbb{R}$ be a function defined by

$$G_n(k) = c(k) + \alpha \int V_n^{\varepsilon}(F(k) + \varepsilon s) d\mu(s), \tag{14}$$

$k = (x, a) \in \mathbb{K}$.

Lemma 8. *Suppose that Assumptions I, III, and IV hold. Then for each $\varepsilon \in \Xi$,*

(a) V_n^{ε} (see (1)) is a Lipschitz function with a constant

$$K_n = (L_0 + \alpha K_{n-1} L_2) \max\{1, L_1\},$$

for $n = 1, 2, ...$, with $K_0 = 0$.

(b) The optimal value function V^{ε} is a Lipschitz function.

Proof. (a) Fix $\varepsilon \in \Xi$. The proof will be made by induction. For $n = 1$ it results that, due to Lemma 7 and Assumption IV (a), V_1^{ε} is a Lipschitz function on X with a constant $K_1 = L_0 \max\{1, L_1\}$. Suppose that V_{n-1}^{ε} is a Lipschitz function with a constant

$$K_{n-1} = (L_0 + \alpha K_{n-2} L_2) \max\{1, L_1\},$$

for $n > 1$. Let $k, k' \in \mathbb{K}$. Then, using (14) and Assumption IV, it results that

$$\left|G_n(k) - G_n\left(k'\right)\right| \leq \left|c(k) - c(k')\right| +$$

$$\alpha \int \left|V_{n-1}^{\varepsilon}(F(k) + \varepsilon s) - V_{n-1}^{\varepsilon}(F(k') + \varepsilon s)\right| d\mu(s)$$

$$\begin{aligned} &\leq L_0 d(k, k') + \alpha K_{n-1} \left|F(k) - F(k')\right| \\ &\leq L_0 d(k, k') + \alpha K_{n-1} L_2 d(k, k') \\ &\leq (L_0 + \alpha K_{n-1} L_2)\, d(k, k'). \end{aligned}$$

Hence, G_n is a Lipschitz function and then, using Lemma 7 and the fact that ε is arbitrary, the result follows.

(b) Firstly, it will be proved that the sequence $\{K_n\}$ given in a) is convergent. The sequence $\{K_n\}$ satisfies the following equation:

$$K_{n+1} = (L_0 + \alpha K_n L_2) \max\{1, L_1\}, \tag{15}$$

for $n \geq 1$ and $K_0 = 0$, since $V_0 \equiv 0$. Iterating (15), it results that

$$K_n = L_0 \max\{1, L_1\} \sum_{i=0}^{n-1} B^i,$$

where $B = \alpha L_2 \max\{1, L_1\}$ and $n > 1$. Without losing generality, it is assumed that $0 < B < 1$. Then, when $n \to \infty$, it follows that

$$\begin{aligned} K &= \lim_{n\to\infty} K_n \\ &= \frac{L_0 \max\{1, L_1\}}{1 - \alpha L_2 \max\{1, L_1\}} \\ &= \frac{L_0 \max\{1, L_1\}}{1 - B}. \end{aligned}$$

Now, using a) and (13) it results that

$$\left|G(k) - G(k')\right| \leq K d(k, k'),$$

for $k, k' \in \mathbb{K}$. Therefore, using Lemma 7 the result follows. □

Remark 9. *The Lipschitz continuity in the context of MDPs using the Kantorovich metric can be consulted in Hinderer [12].*

Lemma 10. *The optimal policy f is continuous.*

Proof. The proof is made by contradiction: suppose that f is not continuous. Then there exist $x \in X$ and a sequence $\{x_n\}$ in X such that x_n converges to x, and $f(x_n) \nrightarrow f(x)$. It is possible to obtain a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ such that

$$d_2(f(x_{n_k}), f(x)) \geq \tau, \tag{16}$$

for some $\tau > 0$ and for all $k = 1, 2, ...$ Since $y_{n_k} = f(x_{n_k}) \in A(x_{n_k}), k = 1, 2, ...$ and the multifunction $x \to A(x)$ is compact-valued and is also u.s.c., there exists a subsequence $\{y_{n_{k_l}}\}$ of $\{y_{n_k}\}$ such that $y_{n_{k_l}} \to y$, for some $y \in A(x)$.

On the other hand, using (1) applied to the deterministic MDP, it results that

$$V(x_{n_{k_l}}) = c(x_{n_{k_l}}, y_{n_{k_l}}) + \alpha V(F(x_{n_{k_l}}, y_{n_{k_l}}), \tag{17}$$

$l = 1, 2, ...$ Then, when $l \to \infty$ in (17), it follows that

$$V(x) = c(x, y) + \alpha V(F(x, y)).$$

But the deterministic optimal policy f is unique, so $y = f(x)$. This last conclusion is a contradiction, since $d_2(y, f(x)) \geq \tau$, due to (16). Therefore, f is continuous.

In the following section the main results of the paper will be presented: see Theorems 11 and 12, below. □

4 Main Results

Assumption V

$w(\cdot)$ and $w_\varepsilon(\cdot)$ are continuous functions on X.

Let Υ be a compact subset of X and $\mathbb{K}_\Upsilon := \{(x, a) : x \in \Upsilon, a \in A(x)\}$.

Theorem 11. *Suppose that Assumptions I-V hold. Let $\{\varepsilon_n\} \subset \Xi$ be a sequence such that $\varepsilon_n \to 0$. Then $\{V^{\varepsilon_n}\}$ converges uniformly to V on every nonempty compact subset of X.*

Proof. Firstly, observe that for each $n \geq 1$,

$$\|V^{\varepsilon_n} - V\|_{\overline{w}_{\varepsilon_n}} \leq R, \tag{18}$$

where $\overline{w}_{\varepsilon_n}(x) = w(x) + w_{\varepsilon_n}(x), x \in X$, and $R = max\{r/(1-\gamma), r_d/(1-\gamma_d)\}$ and r and γ are the constants of the stochastic systems given by Assumption I, meanwhile r_d and γ_d are the constants of the deterministic system. Inequality (18) is a direct consequence of inequality (3) applied to both problems: the deterministic and the stochastic one.

Let $x \in \Upsilon$, where Υ a fixed compact subset of X. Then using Lemma 8 it results that

$$\begin{aligned} &|V^{\varepsilon_n}(x) - V(x)| \\ &\leq \alpha \min_{a\in A(x)} \left| \int V^{\varepsilon_n}(F(x,a) + \varepsilon_n s) d\mu(s) - V(F(x,a)) \right| \\ &\leq \alpha \min_{a\in A(x)} \int |V^{\varepsilon_n}(F(x,a) + \varepsilon_n s) - V(F(x,a))| \, d\mu(s). \end{aligned} \tag{19}$$

On the other hand, for $a \in A(x)$ and $s \in S$,

$$\begin{aligned} &|V^{\varepsilon_n}(F(x,a) + \varepsilon_n s) - V(F(x,a))| \\ &\leq |V^{\varepsilon_n}(F(x,a)) - V(F(x,a))| \\ &\quad + |V^{\varepsilon_n}(F(x,a) + \varepsilon_n s) - V^{\varepsilon_n}(F(x,a))| \\ &\leq K|\varepsilon_n||s| + \|V^{\varepsilon_n} - V\|_{\overline{w}_{\varepsilon_n}} \overline{w}_{\varepsilon_n}(F(x,a)) \\ &\leq K|\varepsilon_n||s| \\ &\quad + \|V^{\varepsilon_n} - V\|_{\overline{w}_{\varepsilon_n}} \sup_{(k,\varepsilon_n)\in\mathbb{K}_\Upsilon\times\Xi} \overline{w}_{\varepsilon_n}(F(k)). \end{aligned} \tag{20}$$

Then, using (20) in (19), it results that

$$|V^{\varepsilon_n}(x) - V(x)| \leq \alpha \widehat{L} \|V^{\varepsilon_n} - V\|_{\overline{w}_{\varepsilon_n}} + \alpha K |\varepsilon_n| E|\xi|,$$

where $\widehat{L} = \sup_{(k,\varepsilon_n)\in\mathbb{K}_\Upsilon\times\Xi} \overline{w}_{\varepsilon_n}(F(k))$. Since $\overline{w}_{\varepsilon_n} \geq 1$, it results that

$$\|V^{\varepsilon_n} - V\|_{\overline{w}_{\varepsilon_n}} \leq \frac{\alpha}{1 - \alpha\widehat{L}} K |\varepsilon_n| E|\xi|.$$

Then,

$$\sup_{x\in\Upsilon} |V^{\varepsilon_n}(x) - V(x)| \leq \frac{\alpha}{1 - \alpha\widehat{L}} K\widehat{L} |\varepsilon_n| E|\xi|. \tag{21}$$

Therefore, when $n \to \infty$ in (21), and since Υ is arbitrary the result follows. □

Theorem 12. *Suppose that Assumptions I-V hold. Let* $\{\varepsilon_n\} \subset \Xi$ *be a sequence such that* $\varepsilon_n \to 0$. *Then* $\{f^{\varepsilon_n}\}$ *converges uniformly to* f *on every nonempty compact subset of* X.

Proof. The proof will be made by contradiction. Let $\Upsilon \subset X$ be a fixed compact set and take $g_n(x) = f^{\varepsilon_n}(x), x \in X, \varepsilon_n \in \Xi, n \geq 1$ with $\varepsilon_n \to 0$. Suppose that

$$\sup_{x\in\Upsilon} d_2(g_n(x), f(x)) \nrightarrow 0.$$

Then there exists $\tau > 0$ such that for all m, there exist $n \geq m$, such that

$$\sup_{x\in\Upsilon} d_2(g_n(x), f(x)) \geq \tau > 0.$$

Let $\{n_k\}$ be a subsequence of $\{n\}$ such that

$$\sup_{x\in\Upsilon} d_2(g_{n_k}(x), f(x)) \geq \tau.$$

For each, $k \geq 1$, there exists $x_{n_k} \in \Upsilon$, such that

$$d_2(g_{n_k}(x_{n_k}), f(x_{n_k})) \geq \tau. \tag{22}$$

Because Υ is compact, there exists $x \in \Upsilon$, such that $x_{n_k} \to x$. Since the multifunction $\overline{x} \to A(\overline{x})$ is u.s.c. and $g_{n_k}(x_{n_k}) \in A(x_{n_k}), k \geq 1$, there exists $a \in A(x)$ such that $g_{n_{k_l}}(x_{n_{k_l}}) \to a$. Now, taking x_{n_k} equal to $x_{n_{k_l}}$ in (22), using the continuity of f, and letting $l \to \infty$ in this inequality, it results that

$$d_2(a, f(x)) \geq \tau > 0, \tag{23}$$

$x \in X$. On the other hand, by (2) it is obtained that

$$V_{n_{k_l}}(x_{n_{k_l}}) = c(x_{n_{k_l}}, g_{n_{k_l}}(x_{n_{k_l}}))$$
$$+\alpha \int V_{n_{k_l}}(F(x_{n_{k_l}}, g_{n_{k_l}}(x_{n_{k_l}})) + \varepsilon_{n_{k_l}} s)d\mu(s).$$

Hence, when $l \to \infty$,

$$V(x) = c(x, a) + \alpha V(x, a),$$

$x \in X$. But, the uniqueness of f implies that $a = f(x)$, which is a contradiction to (23). Therefore, since Υ is arbitrary, Theorem 12 follows. □

In the following examples there will be verified the assumptions given in this paper.

5 Examples

In this section d_1 and d_2 are considered as the usual metric in $\mathbb{R}$, that is $d_1(z_1, z_2) = d_2(z_1, z_2) = |z_1 - z_2|, z_1, z_2 \in \mathbb{R}$.

Example 13. *The dynamic of the system is given by*

$$x_{t+1} = a_t + \varepsilon\xi_t,$$

$t = 0, 1, ...,$ *and* $\{\xi_t\}$ *is a sequence of random variables i.i.d. taking values in* $S = [0, M/2]$ *and* $\varepsilon \in [0, 1]$. *Observe that in this case, trivially,* $E|\xi_0| < +\infty$. *The state space is* $X = [0, M]$, *where* M *is a fixed positive number; the control space is* $A = [0, M/2]$; *the set of admissible controls in a state* x *is* $A(x) = [0, \min\{x, M/2\}]$, *and the cost function is*

$$c(x, a) = e^{x-a},$$

$(x, a) \in \mathbb{K}$.

Remark 14. *For Example 13 Assumptions I, II and V trivially hold as a consequence of the boundness of the cost function* c *and of the compactness of* $A(x), x \in X$.

Lemma 15. *For this Example Assumptions III and IV hold.*

Proof. Assumption III holds since the Hausdorff metric is given by: $H(A(x), A(x')) = |x - x'|$, if $x, x' \in [0, M/2]$, $H(A(x), A(x')) = 0$; if $x, x' \in [M/2, M]$, $H(A(x), A(x')) = |x - M/2|$; if $x \in [0, M/2]$ and $H(A(x), A(x')) = |x' - M/2|$; if $x' \in [0, M/2]$ and $x' \in [M/2, M]$. In all the previous cases the continuity of Lipschitz holds.
Now, let

$$F(x, a) = a,$$

$(x, a) \in \mathbb{K}$ and $s \in S$. Then

$$\begin{aligned} \left|F(k) - F(k')\right| &= \left|a - a'\right| \\ &\leq \max\{\left|x - x'\right|, \left|a - a'\right|\} \\ &= d(k, k'), \end{aligned}$$

where $k = (x, a), k' = (x', a') \in \mathbb{K}$.
Afterwards, the Lipschitz's continuity of the cost function c will be proved. Let $k, k' \in \mathbb{K}$, then

$$\left|c(k) - c(k')\right|$$

$$\begin{aligned} &= \left|e^{-a}\left(e^{x} - e^{x'}\right) + e^{x'}\left(e^{-a} - e^{-a'}\right)\right| \\ &\leq e^{-a}\left|e^{x} - e^{x'}\right| + e^{x'}\left|e^{-a} - e^{-a'}\right|. \end{aligned}$$

Using the Mean Value Theorem, it results that there exist constants M_1 and M_2 such that

$$\begin{aligned} \left|c(k) - c(k')\right| &\leq \left|e^{x} - e^{x'}\right| + e^{M/2}\left|e^{a} - e^{a'}\right| \\ &\leq M_1\left|x - x'\right| + M_2\left|a - a'\right| \\ &\leq Md(k, k'), \end{aligned}$$

where $k, k' \in \mathbb{K}$ and $M = \max\{M_1, M_2\}$. □

Example 16. *An Inventory/Production System. Consider a finite capacity $C < \infty$ of an inventory/production system in which the state variable x_t is the stock level at the beginning of the period t, where $t = 0, 1, 2, ...$ The control variable a_t is the quantity ordered or produced at the beginning of the period t, and the disturbance process $\{\xi_t\}$ is the corresponding demand. $\{\xi_t\}$ is a sequence of i.i.d. random variables with values in the space $S = [0, \infty)$ with distribution μ. It will be supposed that the demand distribution μ is absolutely continuous with density Δ, i.e.*

$$\mu(B) = \int_B \Delta(s)ds,$$

$B \in \mathcal{B}(\mathbb{R})$. Denoting the amount sold during the period t by

$$x_{t+1} = x_t + a_t - \varepsilon\xi_t,$$

and letting the initial state be some given inventory level x_0 independent of $\{\xi_t\}$. It will be assumed that $E|\xi| < +\infty$, where ξ is a generic element of the sequence $\{\xi_t\}$. The state space is $X = \mathbb{R}$, the control space is $A = [0, C]$, the set of admissible controls in the state x is $A(x) = A$, and the cost function is

$$c(x,a) = \psi(a) + \int [p\max(0, x + a - \varepsilon s) + h\max(0, -x - a + \varepsilon s)]\Delta(s)ds, \quad (24)$$

$(x, a) \in \mathbb{K}$, where $\psi : A \to \mathbb{R}$ is the cost production (i.e. $\psi(a)$ represents the cost to order a units), h is the unit holding cost for excess inventory, and p is the shortage cost for unfilled demand. These unit costs are all positive. Moreover, it will be assumed that $\psi(0) = 0$, ψ is strictly convex and Lipschitz continuous. Take $\alpha \leq 1/2$. ε belongs to some fixed nonempty subset containing zero.

Remark 17. *Assumptions I (a) and (b), and II (a) and (b) trivially hold. The proof of the uniqueness of the policy f is similar to the proof given for Example 4.5 in [4], and Assumption I (c) is also verified in that reference. The rest of the Assumptions are shown in the following Lemmas.*

Lemma 18. *The weight-functions w and w_ε are given by*

$$w_\varepsilon(x) = h|x| + T_\varepsilon,$$

$x \in X$, and

$$w(x) = h|x| + T,$$

$x \in X$, where $T_\varepsilon = \eta + Ch + \varepsilon pE[\xi] + 1$, $T = \eta + Ch + 1$, and $\eta = \sup_{a\in A}\psi(a)$.

Remark 19. *Observe that the weight-functions w and w_ε satisfy the following assumptions: Assumption I (d) and II (c), with $r = r_d = 1$ and $\beta = \beta_d = 2$ (these are consequences of the fact that $\alpha \leq 1/2$). Besides Assumptions I (e), II (b) and IV, trivially hold.*

Proof. Taking $y = x + a$ in (24), it results that

$$c(x, y - x) = \psi(y - x) + \lambda(y),$$

where

$$\begin{aligned}\lambda(y) &= hE[\max(0, y - \varepsilon\xi)] \\ &\quad + pE[\max(0, -y + \varepsilon\xi)].\end{aligned}$$

Using the Change Variable Theorem (assuming that $\varepsilon \neq 0$), it is obtained that

$$\begin{aligned}\lambda(y) &= hy\mu\left(\frac{y}{\varepsilon}\right) - \varepsilon(h + p)\int_{-\infty}^{y/\varepsilon} sd\mu(s) \\ &\quad + p\varepsilon E[\xi] - py\left(1 - \mu\left(\frac{y}{\varepsilon}\right)\right).\end{aligned}$$

Then,

$$\lambda(y) \leq hy\mu\left(\frac{y}{\varepsilon}\right) + p\varepsilon E[\xi].$$

Since

1. $0 \leq \mu\left(\frac{y}{\varepsilon}\right) \leq 1$; and
2. $y \in [x, x + C], x \in X$,

it results that

$$\begin{aligned}|c(x,a)| &\leq \eta + Ch + hx + p\varepsilon E[\xi] \\ &\leq \eta + Ch + h|x| + p\varepsilon E[\xi] + 1,\end{aligned}$$

$(x, a) \in \mathbb{K}$. Then $w_\varepsilon(x) = h|x| + T_\varepsilon, x \in X$, and in a similar way it is possible to prove that $w(x) = h|x| + T$, $x \in X$. □

Lemma 20. *For Example 16, Assumptions III,IV, and V hold.*

Proof. To prove this Lemma it is just necessary to demonstrate that the cost function c and the function

$$F(x, a) = x + a,$$

$(x, a) \in K$ are Lipschitz functions. Observe that in this case the Hausdorff metric is constant, i.e. $H(A(x), A(x')) = 0$, $x, x' \in X$. First of all, it will be proved that F is a Lipschitz function. To do so, let $k, k' \in \mathbb{K}$ and

$$\begin{aligned}\left|F(k) - F(k')\right| &\leq \left|a - a'\right| + \left|x - x'\right| \\ &\leq 2\max\left\{\left|a - a'\right|\left|x - x'\right|\right\} \\ &= 2d(k, k').\end{aligned}$$

Now, it will be proved that c is a Lipschitz function. Since, the function ψ is Lipschitz, it is just necessary to prove that the function

$$\begin{aligned}\widehat{H}(k) &:= \int p\max(0,k-\varepsilon s)\,\Delta(s)ds \\ &\quad + \int h\max(0,-k+\varepsilon s)\,\Delta(s)ds,\end{aligned}$$

$k \in \mathbb{K}$, is Lipschitz. Let $k, k' \in \mathbb{K}$. Then, using the identity $\max(0,z) = (z+|z|)/2, z \in \mathbb{R}$, it results that

$$\begin{aligned}&\left|\widehat{H}(k) - \widehat{H}(k')\right| \\ &\leq \frac{p+h}{2}\int |k-k'|\,\Delta(s)ds \\ &\quad + \frac{p+h}{2}\int \left||k-\varepsilon s| - |k'-\varepsilon s|\right|\Delta(s)ds \\ &\leq (p+h)\,|k-k'| \\ &\leq (p+h)\left(|x-x'| + |a+a'|\right) \\ &\leq 2(p+h)d(k,k').\end{aligned}$$

Therefore, the result follows. □

6 Concluding Remarks

In this article there have been established conditions which guarantee the uniform of compact sets convergence of both the optimal value functions and the optimal policies of a certain class of stochastic systems to the optimal value functions and the optimal policies of the deterministic systems associated, in an convenient sense, to the stochastic ones, respectively.

With the results obtained in this article, it is now possible to study the perturbation methodology in the context of MDPs (see [13] and [14]) and to find inequalities to estimate the stability (robustness) between the stochastic system and the deterministic one.

Research in these directions is still in progress.

References

[1] J. P. Aubin, and I. Ekeland, *Applied Nonlinear Analysis*, Wiley, New York, 1984.

[2] D. P. Bertsekas, *Dynamic Programming: Deterministic and Stochastic Models*, Prentice-Hall, Inc., New Jersey, 1987.

[3] F. Besharati, M. Tarjoman, H. Ghassemian, Quad Tree Decomposition of Fused Image of Sunspots for Classifying The Trajectories, *Proceedings of the 7th WSEAS International Conference on Automation and Information, Cavtat, Croatia*, 2006, pp. 105-109.

[4] D. Cruz-Suárez, R. Montes-de-Oca, and F. Salem-Silva, Conditions for the uniqueness of optimal policies of discounted Markov decision processes, *Math. Methods of Oper. Res.*, Vol. 60, No. 3, 2004, pp. 415-436.

[5] H. Cruz-Suárez and R. Montes-de-Oca, Discounted Markov control processes induced by deterministic systems, *Kybernetika (Prague)*, Vol. 42, No. 6, 2006, pp. 647-664.

[6] P. Dupuis and R. J. Mathew, Rates of convergence for approximation schemes in optimal control, *SIAM J. Control Optim.*, Vol. 36, No.2, 1998, pp. 719-741.

[7] P. Dupuis and A. Szipiro, Convergence of the optimal feedback policies in a numerical method for a class of deterministic optimal control problems, *SIAM J. Control Optim.*, Vol. 40, No. 2, 2001, pp. 393-420.

[8] W. H. Flemming, Stochastic control for small noise intensities, *SIAM J. Control Optim.*, Vol. 9, No. 3, 1971.

[9] K. Framling, Reducing state space exploration in reinforcement learning problems by rapid identification of initial solutions and progressive improvement of them, Advances in Neural Networks World. *Proceedings of 3rd WSEAS International Conference on Neural Networks and Applications (NNA'02), Interlaken, Switzerland*, 2002, pp. 83-88.

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

[11] O. Hernández-Lerma and J. B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.

[12] K. Hinderer, Lipschitz Continuity of value functions in Markovian decision processes, *Math. Methods of Oper. Res.*, Vol. 62, No. 1, 2005, pp. 3-22.

[13] K. L. Judd, *Numerical Methods in Economics*, The MIT Press, 1998.

[14] K. L. Judd and G. Sy-Ming, Asymptotic methods for aggregate growth models, *J. Econ. Dynam. Control*, Vol. 21, No. 6, 1997, pp. 1025-1042.

[15] O. D. Karaduman, A. M. Erkmen, N. Baykal, Intelligent "Health Restoration System": Reinforcement Learning Feedback to Diagnosis and Treatment Planning, *Proceedings of the 5th*

WSEAS International Conference on Telecommunications and Informatics, Istanbul, Turkey, 2006, pp. 463-468.

[16] R. Sh. Liptser, W. J. Runggaldier, and M. Taksar, Deterministic approximation for stochastic control problems, *SIAM J. Control Optim.*, Vol. 34, No. 1, 1996, pp. 161-178.

[17] S. Schmitt-Grohé, and M. Uribe, Solving dynamic general equilibrium models using a second-order approximation to the policy function, *J. Econ. Dynam. Control*, Vol. 28, No. 4, 2004, pp. 755- 775.

[18] N. L. Stokey and R. E. Lucas, *Recursive Methods in Economic Dynamics*, Harvard University Press, Massachusetts, 1989,

[19] C. Rodríguez, The problem of Robot random motion tracking learning algorithms, *Proceedings of the 6th WSEAS International Conference on Signal Processing, Robotics and Automation, Corfu Island, Greece*, 2007, pp. 219-224.

[20] N. Williams, Small noise asymptotics for a stochastic growth model, *J Econ. Th.*, Vol. 119, No. 2, 2004, pp. 271-298.