Morfismos, Vol. 12, No. 2, 2008, pp. 1–15

The vanishing discount approach to average reward optimality: the strongly and the weakly continuous cases *

Tomás Prieto-Rumeau Onésimo Hernández-Lerma

Abstract

We consider a discrete-time stochastic dynamic programming model and we propose conditions under which the limit of discount optimal policies, as the discount factor converges to one, is average optimal. We prove this result under strong and weak continuity conditions and, moreover, we relax the usual value boundedness condition on the relative values of the optimal discounted reward.

2000 Mathematics Subject Classification: 93E20, 90C40.
Keywords and phrases: dynamic programming, vanishing discount, average optimality.

1 Introduction

The basic problem dealt with in this paper is the existence of control policies π that maximize the long-run expected average reward

$$v(x,\pi) := \liminf_{T\to\infty} \mathbb{E}_x^\pi \left[\frac{1}{T}\sum_{t=0}^{T-1} r(x_t,\pi(x_t))\right] \tag{1}$$

for every initial state $x_0 = x$. (The underlying controlled system is a fairly general discrete-time stochastic control process described in Section 2; see (6).) Among the several known techniques to analyze this problem, the most common is the vanishing discount approach, which can be traced back to Taylor [16]. It is so-named because it is based on

*This research was partially supported by CONACyT Grant 45693-F.

the convergence as $\rho \uparrow 1$ $(0 < \rho < 1)$ of ρ-discounted optimal reward policies. To state this more precisely, we need some notation.

For each discount factor $\rho \in (0,1)$, let

$$v_\rho(x,\pi) := \mathbb{E}_x^\pi \left[\sum_{t=0}^{\infty} \rho^t r(x_t, \pi(x_t)) \right] \tag{2}$$

be the expected discounted reward of the admissible control policy $\pi \in \Pi$ (see Section 2) when the initial state is $x_0 = x$. The optimal ρ-discounted reward function is defined as

$$v_\rho(x) := \sup_{\pi \in \Pi} v_\rho(x,\pi) \tag{3}$$

for every state x. For a given fixed state x', consider the relative value function

$$u_\rho(x) := v_\rho(x) - v_\rho(x').$$

This function is one of the key tools in the vanishing discount approach. To obtain the convergence of ρ-discount optimal policies to average optimal policies as $\rho \uparrow 1$, it was assumed in [16] that u_ρ was uniformly bounded, that is, there exists a constant L such that

$$|u_\rho(x)| \le L$$

for every state x and $0 < \rho < 1$. This condition was later relaxed to the following weaker value boundedness condition: there exists a constant L and a function m such that

$$-m(x) \le u_\rho(x) \le L \tag{4}$$

for every state x and $0 < \rho < 1$; see, e.g., [2, Assumption A1], [5, Assumption 4.1], [12, Definition 2.1] or [15].

In this paper, we further relax (4) and assume the existence of a function m (satisfying appropriate hypotheses) such that

$$-m(x) \le u_\rho(x) \le m(x) \tag{5}$$

for every x and $0 < \rho < 1$. Such a condition can also be found in e.g. [3, Lemma 4.5], [4, Assumption 3.3] or [7, Lemma 10.4.2]. Relaxing (4) to (5) is indeed a relevant issue because (4) is, in fact, a fairly restrictive condition. For instance, to obtain (4), it is assumed in [12] that the reward function r is bounded. Moreover, condition (4) excludes the case

of an unbounded utility function (see the comment after Assumption 5.3 in [12, p. 1423]). Also, in Section 4 of this paper, we describe a control model for which (5) holds, whereas (4) does not.

Summarizing, the goal of this paper is to give conditions on the controlled process that, together with the condition (5), ensure that the limit of ρ-discount optimal policies, as $\rho \uparrow 1$, is average optimal. The basic control model is described in Section 2. In Section 3, we consider two different sets of hypotheses, namely, strong and weak continuity conditions, depending on the corresponding strong or weak continuity of the control system's transition function. Also in Section 3, we state our main results: Theorem 3.10 and Corollary 3.12, in which we mention several particular cases of interest. Finally, we present an example in Section 4, and our conclusions are stated in Section 5.

2 The control model

The formulation of the controlled process and the notation is mainly drawn from [12].

We assume that the state space S and the action set A are Borel spaces (that is, measurable subsets of complete and separable metric spaces). Let Γ be a nonempty set-valued function from S to A. For each $x \in S$, the corresponding set of feasible control actions is $\Gamma(x) \subseteq A$. The family of feasible state-action pairs is denoted by K, i.e.,

$$K := \{(x, a) \in S \times A : a \in \Gamma(x)\},$$

which is assumed to be a measurable subset of $S \times A$. (In this paper, measurability is always referred to the Borel σ-algebra.)

We consider a sequence $\{\xi_t\}_{t\geq 0}$ of i.i.d. random variables from a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $(Z, \mathcal{Z})$ with common distribution ν. Let $h : K \times Z \to S$ be a measurable function. We assume that the state of the system is updated according to the function h, meaning that if the action $a \in \Gamma(x)$ is chosen at $x \in S$ and the value of the random perturbation is ξ, then the next state of the system is $h(x, a, \xi) \in S$.

We suppose that the reward function is the measurable real-valued mapping $r : K \to \mathbb{R}$.

Let Π be the family of measurable functions $\pi : S \to A$ such that $\pi(x) \in \Gamma(x)$ for every $x \in S$. (We suppose that Π is nonempty.) We call $\pi \in \Pi$ a deterministic stationary policy. For each $\pi \in \Pi$ and every

initial state $x_0 \in S$ independent of $\{\xi_t\}_{t\geq 0}$,

$$x_{t+1} = h(x_t, \pi(x_t), \xi_t) \quad \text{for } t = 0, 1, \ldots \tag{6}$$

is a Markov process and it stands for the state of the system under the policy π. The corresponding expectation operator is denoted by $\mathbb{E}^{\pi}_{x_0}$. Although larger classes of policies may be considered, it is well known that for the control problem we are dealing with Π is a "sufficient" class of policies — see [6, Chapter 4] or [7, Chapter 8], for instance.

Given an admissible policy $\pi \in \Pi$ and an initial state $x \in S$, the corresponding long-run average reward and expected discounted reward are defined as in (1) and (2), respectively. Given a discount factor $0 < \rho < 1$, we say that $\pi \in \Pi$ is ρ-discount optimal if $v_\rho(x, \pi) = v_\rho(x)$ for every $x \in S$ (recall (3)). Similarly, $\pi \in \Pi$ is average reward optimal if

$$v(x, \pi^*) = \sup_{\pi \in \Pi} v(x, \pi) \quad \forall\ x \in S.$$

3 Main results

As already mentioned, we will consider two different sets of hypotheses, which we label as strong and weak continuity assumptions.

The strongly continuous case

We state the assumptions we make on our control model. First, we have the following Lyapunov-like condition.

Assumption 3.1 *There exists a measurable function $w : S \to [1, \infty)$, and constants $0 < \beta < 1$ and $b > 0$ such that*

$$\int_Z w(h(x, a, \xi))\nu(d\xi) \leq \beta w(x) + b \quad \forall\ (x, a) \in K.$$

The next assumption introduces some usual continuity and compactness requirements. We note that the function w in Assumptions 3.2 and 3.4 is taken from Assumption 3.1.

Assumption 3.2 **(i)** *For every $x \in S$, the set $\Gamma(x)$ is compact.*

(ii) *The reward function $r(x, a)$ is upper semicontinuous on $A(x)$ for every $x \in S$. In addition, there exists a constant M such that*

$$|r(x, a)| \leq M w(x) \quad \forall\ (x, a) \in K.$$

(iii) *The function*

$$(x,a) \mapsto \int_Z w(h(x,a,\xi))\nu(d\xi)$$

is continuous on $A(x)$ for every $x \in S$.

(iv) ***Strong continuity.*** *For every bounded and measurable $\zeta : S \to \mathbb{R}$, the function*

$$(x,a) \mapsto \int_Z \zeta(h(x,a,\xi))\nu(d\xi)$$

is continuous on $A(x)$ for every $x \in S$.

Remark 3.3 (The additive-noise case) *The strong continuity condition is satisfied, for instance, when $S = Z = \mathbb{R}$,*

$$h(x,a,\xi) = g(x,a) + \xi,$$

where g is continuous on $A(x)$ for each fixed $x \in S$, and, in addition, ν has an almost everywhere continuous bounded density with respect to the Lebesgue measure. This includes, of course, the linear case in which $g(x,a) = k_1 x + k_2 a$ for some constants k_1, k_2.

Finally, we state the value boundedness condition.

Assumption 3.4 *There exists a state $x' \in S$ and a constant $M' > 0$ such that*

$$\sup_{0<\rho<1} |v_\rho(x) - v_\rho(x')| \leq M' w(x) \quad \forall\ x \in S.$$

The weakly continuous case

Among the hypotheses made so far on the control model, the most restrictive one is the strong continuity condition in Assumption 3.2(iv). Under additional appropriate conditions, strong continuity can be relaxed to weak continuity. To this end, first, the "measurability" of w in Assumption 3.1 is replaced with "continuity".

Assumption 3.5 *There exists a continuous function $w : S \to [1,\infty)$, and constants $0 < \beta < 1$ and $b > 0$ such that*

$$\int_Z w(h(x,a,\xi))\nu(d\xi) \leq \beta w(x) + b \quad \forall\ (x,a) \in K.$$

In Assumptions 3.6 and 3.8 below, the function w is taken from Assumption 3.5.

Assumption 3.6 **(i)** *The function $\Gamma : S \to 2^A$ is upper semicontinuous and compact-valued.*

(ii) *The reward function r is upper semicontinuous on K and, moreover, there exists a constant $M > 0$ such that*

$$|r(x,a)| \leq Mw(x) \quad \forall\ (x,a) \in K.$$

(iii) *The function*

$$(x,a) \mapsto \int_Z w(h(x,a,\xi))\nu(d\xi)$$

is continuous on K.

(iv) ***Weak continuity.*** *The function*

$$(x,a) \mapsto \int_Z \zeta(h(x,a,\xi))\nu(d\xi)$$

is continuous on K for every bounded and continuous $\zeta : S \to \mathbb{R}$.

Remark 3.7 *The weak continuity assumption is satisfied, for instance, if the function $h(x,a,\xi)$ is continuous on K for each $\xi \in Z$.*

We introduce some notation. Let $\mathcal{B}_w(S)$ be the family of measurable functions $\zeta : S \to \mathbb{R}$ with finite w-norm, that is,

$$||\zeta||_w := \sup_{x \in S}\{|\zeta(x)|/w(x)\} < \infty.$$

Assumption 3.8 *The controlled process is w-uniformly ergodic on Π; that is, for each $\pi \in \Pi$, the Markov process (6) has a unique invariant probability measure μ_π on S and, in addition, there exist constants $R > 0$ and $0 < \alpha < 1$ such that for every $x \in S$, $\zeta \in \mathcal{B}_w(S)$ and $t \geq 0$*

$$\sup_{\pi \in \Pi} \left| E_x^\pi[\zeta(x_t)] - \int_S \zeta(y)\mu_\pi(dy) \right| \leq w(x)||\zeta||_w R\alpha^t.$$

In the weakly continuous case, we do not need to impose a value boundedness condition because, in fact, Assumption 3.8 implies Assumption 3.4 (the proof is easy; see, e.g., Lemma 4.5 in [3] or Lemma

10.4.2 in [7]). A sufficient condition for Assumption 3.8 is proposed in [7, Proposition 10.2.5].

In what follows, we will suppose that either the Assumptions 3.1, 3.2 and 3.4 or the Assumptions 3.5, 3.6 and 3.8 hold. In either case, we know from the results in [7, Chapter 8] that the optimal ρ-discounted reward is the unique solution in $\mathcal{B}_w(S)$ of the *discounted reward optimality equation*:

$$v_\rho(x) = \max_{a\in\Gamma(x)} \left\{ r(x,a) + \rho \int_Z v_\rho(h(x,a,\xi))\nu(d\xi) \right\} \quad \forall\, x \in S. \tag{7}$$

In addition, a policy $\pi^* \in \Pi$ is ρ-discount optimal if and only if $\pi^*(x)$ attains the maximum in (7) for every $x \in S$, i.e.,

$$v_\rho(x) = r(x,\pi^*(x)) + \rho \int_Z v_\rho(h(x,\pi^*(x),\xi))\nu(d\xi) \quad \forall\, x \in S. \tag{8}$$

The vanishing discount approach to average reward optimality is related to the following definition of limit and accumulation policies.

Definition 3.9 *Given a policy $\pi^* \in \Pi$ and a sequence $\{\pi_k\}_{k\in\mathbb{N}}$ in Π, we say that*

(i) *$\{\pi_k\}_{k\in\mathbb{N}}$ converges to π if $\lim_k \pi_k(x) = \pi(x)$ for every $x \in S$;*

(ii) *π^* is an accumulation policy of $\{\pi_k\}_{k\in\mathbb{N}}$ if, for every $x \in S$, there exists a subsequence $\{k_x\}$ such that $\pi_{k_x}(x) \to \pi(x)$;*

(iii) *$\{\pi_k\}_{k\in\mathbb{N}}$ converges continuously to π if $\lim_k \pi_k(x_k) = \pi(x)$ for every $x \in S$ and every sequence $x_k \to x$.*

The concept of accumulation policy in Definition 3.9(ii) comes from [13]. Continuous convergence and its applications to stochastic dynamic programming are analyzed in [10].

Next, we prove our main result, which states the relation between average reward optimal policies and the limit of discount optimal policies. The proof of this result, Theorem 3.10, follows the same arguments needed to obtain the so-called *average reward optimality inequality* [7, Theorem 10.3.1], although the proof is focused on the analysis of the limit of discount optimal policies.

Theorem 3.10 *Let $\{\rho_k\}_{k\in\mathbb{N}}$, with $\rho_k \uparrow 1$, be a sequence of discount factors, and let $\pi_k \in \Pi$, for every $k \in \mathbb{N}$, be a ρ_k-discount optimal policy. Then the following holds:*

(i) *If Assumptions 3.1, 3.2 and 3.4 are satisfied and* $\{\pi_k\}$ *converges to* $\pi^* \in \Pi$*, then* π^* *is an average reward optimal policy;*

(ii) *If Assumptions 3.5, 3.6 and 3.8 are satisfied and* $\{\pi_k\}$ *converges continuously to* $\pi^* \in \Pi$*, then* π^* *is an average reward optimal policy.*

Proof. From Assumption 3.1 or 3.5, an induction argument (see, e.g., [7, Lemma 10.4.1]) gives

$$\mathbb{E}_x^\pi[w(x_t)] \leq \beta^t w(x) + \frac{(1-\beta^t)}{(1-\beta)b} \quad \forall\ \pi \in \Pi,\ x \in S,\ t \geq 0. \tag{9}$$

Therefore, by Assumption 3.2(ii) or 3.6(ii), we have

$$\mathbb{E}_x^\pi |r(x_t, \pi(x_t))| \leq M\beta^t w(x) + \frac{M(1-\beta^t)}{(1-\beta)b},$$

so that $\sup_{\rho\in(0,1)} |(1-\rho)v_\rho(x')|$ is finite, with $x' \in S$ as in Assumption 3.4. Thus

$$\underline{g} := \liminf_{k\to\infty}\ (1-\rho_k)v_{\rho_k}(x')$$

is well defined.

Our proof now proceeds in two steps. In step one, we prove that

$$\underline{g} \geq \sup_{\pi\in\Pi} v(x,\pi) \quad \forall\ x \in S.$$

In step two, we show that π^* satisfies

$$\underline{g} \leq v(x,\pi^*) \quad \forall\ x \in S.$$

Average reward optimality of π^* will then follow.

Step one. By definition of u_ρ (in Section 1), a simple calculation shows that the discounted reward optimality equation (7) can be written in the equivalent form:

$$(1-\rho)v_\rho(x') + u_\rho(x) = \max_{a\in\Gamma(x)} \left\{ r(x,a) + \rho \int_Z u_\rho(h(x,a,\xi))\nu(d\xi) \right\} \tag{10}$$

for every $x \in S$. Consider now a subsequence $\{k'\}$ of $\{k\}$ such that

$$\lim_{k'\to\infty}\ (1-\rho_{k'})v_{\rho_{k'}}(x') = \underline{g}.$$

Let $\underline{u} := \liminf_{k'} u_{\rho_{k'}}$ and note that $\underline{u}$ is in $\mathcal{B}_w(S)$. Now, by (10), for the sequence $\{\rho_{k'}\}$ and every $(x,a) \in K$ we have

$$(1-\rho_{k'})v_{\rho_{k'}}(x') + u_{\rho_{k'}}(x) \geq r(x,a) + \rho_{k'} \int_Z u_{\rho_{k'}}(h(x,a,\xi))\nu(d\xi).$$

Taking the $\liminf_{k'\to\infty}$ in this inequality and using Fatou's lemma (which indeed applies as a consequence of our assumptions), we obtain

$$\underline{g} + \underline{u}(x) \geq r(x,a) + \int_Z \underline{u}(h(x,a,\xi))\nu(d\xi) \quad \forall\ (x,a) \in K. \tag{11}$$

Iteration of (11) yields that, for every initial state $x \in S$, any policy $\pi \in \Pi$ and $t \geq 0$,

$$\underline{g} \geq \mathbb{E}_x^\pi[r(x_t, \pi(x_t))] + \mathbb{E}_x^\pi[\underline{u}(x_{t+1}) - \underline{u}(x_t)].$$

Summing up these inequalities for $t = 0, \ldots, T-1$ and then dividing by T yields

$$\underline{g} \geq \mathbb{E}_x^\pi \left[\frac{1}{T}\sum_{t=0}^{T-1} r(x_t, \pi(x_t))\right] + \frac{\mathbb{E}_x^\pi[\underline{u}(x_T)] - \underline{u}(x)}{T}.$$

Letting $T \to \infty$, recalling that $\underline{u} \in \mathcal{B}_w(S)$ and using (9), we obtain $\underline{g} \geq v(x,\pi)$ and, therefore,

$$\underline{g} \geq \sup_{\pi\in\Pi} v(x,\pi) \quad \forall\ x \in S. \tag{12}$$

This completes step one.

Step two. Since π_k is a ρ_k-discount optimal policy, from (8) and (10) we have

$$(1-\rho)v_{\rho_k}(x') + u_{\rho_k}(x) = r(x, \pi_k(x)) + \rho_k \int_Z u_{\rho_k}(h(x, \pi_k(x), \xi))\nu(d\xi)$$

for every $k \in \mathbb{N}$ and $x \in S$. Consequently, for every $\varepsilon > 0$ and large enough k, we have

$$\underline{g} - \varepsilon + u_{\rho_k}(x) \leq r(x, \pi_k(x)) + \rho_k \int_Z u_{\rho_k}(h(x, \pi_k(x), \xi))\nu(d\xi) \tag{13}$$

for every $x \in S$.

Suppose now that the Assumptions 3.1, 3.2 and 3.4 are satisfied. Then, taking the lim sup in (13), recalling that $r(x,\cdot)$ is upper semicontinuous and by the extension of Fatou's lemma [7, Lemma 8.3.7], we obtain

$$\underline{g} - \varepsilon + \overline{u}(x) \leq r(x,\pi^*(x)) + \int_Z \overline{u}(h(x,\pi^*(x),\xi))\nu(d\xi),$$

where $\overline{u} := \limsup_k u_{\rho_k} \in \mathcal{B}_w(S)$. But $\varepsilon > 0$ being arbitrary, the same arguments as in the proof of step one yield that

$$\underline{g} \leq v(x,\pi^*) \quad \forall\, x \in S,$$

which combined with (12) shows that π^* is an average reward optimal policy and, besides, that $\underline{g}$ is the (constant) optimal average reward. This completes the proof of statement (i), that is, under the hypotheses in the strongly continuous case.

We now consider the weakly continuous case, which consists of Assumptions 3.5, 3.6 and 3.8. Following [8], we define the generalized lim sup of the sequence u_{ρ_k} as

$$u^*(x) := \sup\{\limsup_{k\to\infty} u_{\rho_k}(x_k)\},$$

where the supremum is taken over the family of sequences $\{x_k\} \subseteq S$ such that $x_k \to x$. Let us now go back to (13) and take the lim sup through a sequence $x_k \to x$ such that $\limsup_k u_{\rho_k}(x_k) \geq u^*(x) - \varepsilon$, so that

$$\begin{aligned} \underline{g} - 2\varepsilon + u^*(x) \quad &\leq \quad \limsup_{k\to\infty} r(x_k,\pi_k(x_k)) \\ &\quad + \limsup_{k\to\infty} \int_Z u_{\rho_k}(h(x_k,\pi_k(x_k),\xi))\nu(d\xi). \end{aligned}$$

Then we proceed as in the proof for the strongly continuous case, but this time we take into account that both r and the multifunction Γ are upper semicontinuous. Finally, we apply the Fatou lemma for a generalized lim sup (see [8, Lemma 5] and also [14, Lemma 2.3]) to obtain

$$\underline{g} - 2\varepsilon + u^*(x) \leq r(x,\pi^*(x)) + \int_Z u^*(h(x,\pi^*(x),\xi))\nu(d\xi) \quad \forall\, x \in S. \tag{14}$$

This implies, by standard arguments, that $v(x,\pi^*) \geq \underline{g}$ for every $x \in S$. The proof of Theorem 3.10 is complete. □

Remark 3.11 *The second step in the proof of Theorem 3.10 relies on the application of a Fatou-like lemma. For instance, when the usual value boundedness condition holds, then we use the standard Fatou lemma because the relative value function u_ρ is bounded above; see (4). Under the strong continuity assumptions, we use the Fatou lemma in [7, Lemma 8.3.7], while if the weak continuity conditions hold, then we use the Fatou lemma for a generalized* $\limsup$ *in [14, Lemma 2.3]. Therefore, the assumptions we make on the control model heavily depend on the hypotheses needed for the corresponding Fatou lemma and, similarly, the kind of results we reach (statements (i) and (ii) in Theorem 3.10) also depend on the kind of Fatou lemma that is applied.*

We specialize Theorem 3.10 to the following important particular cases.

Corollary 3.12 *Suppose that $\{\rho_k\}_{k\in\mathbb{N}}$ is a sequence of discount factors such that $\rho_k \uparrow 1$ and let $\pi_k \in \Pi$, for every $k \in \mathbb{N}$, be a ρ_k-discount optimal policy.*

(i) *Under the strong continuity conditions (Assumptions 3.1, 3.2 and 3.4), if for every $x \in S$ the function $\rho \mapsto u_\rho(x)$ is monotone (either increasing or decreasing), then any accumulation policy of $\{\pi_k\}_{k\in\mathbb{N}}$ is average reward optimal.*

(ii) *If the state space S is denumerable, then under either the strong or the weak continuity conditions, any accumulation policy of $\{\pi_k\}$ is average reward optimal.*

The condition in Corollary 3.12(i) can be interpreted as follows: the expected discounted reward grows faster for any $x \in S$ than for $x' \in S$ as $\rho \uparrow 1$, and it is satisfied, for instance, in the consumption-investment model in [6, Section 3.6]; see also [1].

4 An example

In this section we give an example of a control model that satisfies (5) but does not satisfy the value boundedness condition (4).

The following inventory system with permitted backlog is based on the model analyzed in [17]. The state space and the action set are $S = A = \mathbb{R}$. The distribution ν is supported on $[0, \infty)$, it satisfies the

conditions in Remark 3.3, and we assume that its expectation equals one. Furthermore, we suppose that there exists some $\delta > 0$ such that

$$\int_0^\infty e^{\delta\xi}\nu(d\xi) < \infty.$$

(Note that, for instance, the mean one exponential distribution satisfies these hypotheses.) Fix a constant $K > 1/2$ and let

$$0 < \lambda < -\frac{1}{\delta}\log\int_0^\infty e^{-\delta\xi}\nu(d\xi).$$

The action sets $\Gamma(x)$ are the intervals

$$[-x, \max\{-2x, -x + K\}] \quad \text{for } x \leq 0$$

and

$$[-x, \max\{\lambda, -x + K\}] \quad \text{for } x > 0.$$

The system's transition function h is given by $h(x, a, \xi) = x + a - \xi$. The cost function is $c(x, a) = (x + a)^2 - a$ (cf. [17, Equation (3.1)]). Finally, let $w(x) = e^{\delta|x|}$ for $x \in \mathbb{R}$. This control model satisfies the Assumptions 3.1 and 3.2.

Given a discount factor $0 < \rho < 1$, a direct calculation shows that the optimal ρ-discounted cost function (recall that we are minimizing a cost) is

$$v_\rho(x) = x - \frac{(\rho + 1)^2}{4(1 - \rho)} \quad \forall\ x \in \mathbb{R},$$

and the optimal ρ-discount policy is

$$\pi_\rho(x) = -x + \frac{1}{2}(1 - \rho) \quad \forall\ x \in \mathbb{R}.$$

Hence, the value boundedness condition (4) *does not hold*, whereas (5) (or Assumption 3.4) is satisfied.

Moreover, for every $x \in \mathbb{R}$, $\pi_\rho(x)$ converges to $-x$ as $\rho \uparrow 1$. Therefore, by Theorem 3.10(i), the policy $\pi(x) = -x$, for $x \in \mathbb{R}$, is average cost optimal. Further, from the proof of Theorem 3.10 we also obtain that the minimal average cost is

$$-1 = \lim_{\rho\uparrow 1}\ (1 - \rho)v_\rho(x).$$

5 Concluding remarks

In the previous sections, we have considered a fairly general discrete-time stochastic control model and, under two different sets of hypotheses (strong and weak continuity), we have proved that the limit of ρ-discount optimal policies, as the discount factor $\rho \uparrow 1$, is a long-run average reward optimal policy. The main contribution of this paper is to relax the usual value boundedness assumption on the relative value funtion (4) and, instead, assume the weaker condition (5). We have illustrated our results with the generalized inventory system in Section 4.

Some important issues, however, remain open. In Theorem 3.10(i) it is assumed that the discount optimal policies $\{\pi_k\}$ converge to some π^*, and then it is proved that π^* is average reward optimal. It would be interesting to know whether this convergence can be relaxed, and thus obtain a result like that in Corollary 3.12(i) under general assumptions. To this end, results on the existence of measurable selectors would be involved. Also, it would be interesting to check whether the continuous convergence in Theorem 3.10(ii) can be relaxed to (usual) convergence, perhaps by strengthening the hypotheses on the control model.

Tomás Prieto-Rumeau
Departamento de Estadística,
Facultad de Ciencias, UNED,
Senda del Rey 9, 28040,
Madrid, Spain,
tprieto@ccia.uned.es

Onésimo Hernández-Lerma
Departamento de Matemáticas,
CINVESTAV-IPN, 14-470,
México D.F. 07000,
México,
ohernand@math.cinvestav.mx

References

[1] Cruz-Suárez H. D., *A stochastic consumption-investment problem with unbounded utility function*, Morfismos **4** (2000), 19–30.

[2] Dutta P. K., *What do discounted optima converge to? A theory of discount rate asymptotics in economic models*, J. Econom. Theory **55** (1991), 64–94.

[3] Gordienko E.; Hernández-Lerma O., *Average cost Markov control processes with weighted norms: existence of canonical policies*, Appl. Math. (Warsaw) **23** (1995), 199–218.

[4] Guo X. P.; Zhu Q. X., *Average optimality for Markov decision processes in Borel spaces: a new condition and approach*, J. Appl. Prob. **43** (2006), 318–334.

[5] Hernández-Lerma O.; Lasserre J. B., *Average cost optimal policies for Markov control processes with Borel state space and unbounded costs*, Systems Control Lett. **15** (1990), 349–356.

[6] Hernández-Lerma O.; Lasserre J. B., Discrete-Time Markov Control Processes: Basic Optimality Criteria, Springer, New York, 1996.

[7] Hernández-Lerma O.; Lasserre J. B., Further Topics on Discrete-Time Markov Control Processes, Springer, New York, 1999.

[8] Jaśkiewicz A.; Nowak A. S., *On the optimality equation for average cost Markov control processes with Feller transition probabilities*, J. Math. Anal. Appl. **316** (2006), 495–509.

[9] Kawaguchi K.; Morimoto H., *Long-run average welfare in a pollution accumulation model*, J. Econom. Dynam. Control **31** (2007), 703–720.

[10] Langen H. J., *Convergence of dynamic programming models*, Math. Oper. Res. **6** (1981), 493–512.

[11] Morimoto H.; Fujita Y., *Ergodic control in stochastic manufacturing systems with constant demand*, J. Math. Anal. Appl. **243** (2000), 228–248.

[12] Nishimura K.; Stachurski J., *Stochastic optimal policies when the discount rate vanishes*, J. Econom. Dynam. Control **31** (2007), 1416–1430.

[13] Schäl M., *Conditions for optimality and for the limit of n-stage optimal policies to be optimal*, Z. Wahrs. verw. Gerb. **32** (1975), 179–196.

[14] Schäl M., *Average optimality in dynamic programming with general state space*, Math. Oper. Res. **18** (1993), 163–172.

[15] Sennott L. I., *A new condition for the existence of optimal stationary policies in average cost Markov decision processes*, Oper. Res. Lett. **5** (1986), 17–23.

[16] Taylor H. M., *Markovian sequential replacement processes*, Ann. Math. Stat. **36** (1965), 1677–1694.

[17] Vega-Amaya O.; Montes-de-Oca R., *Application of average dynamic programming to inventory systems*, Math. Methods Oper. Res. **47** (1998), 451–471.