Top
DOI 10.1007/s11750-009-0110-7

ORIGINAL PAPER

Optimal policies for constrained average-cost Markov decision processes

Juan González-Hernández · César E. Villarreal

Received: 12 December 2008 / Accepted: 2 July 2009

Abstract We give mild conditions for the existence of optimal solutions for a Markov decision problem with average cost, under m constraints of the same kind, in Borel actions and states spaces. Moreover, there is an optimal policy that is a convex combination of at most $m + 1$ deterministic policies.

Keywords Markov decision processes · Constraints · Stable measures

Mathematics Subject Classification (2000) 90C40

1 Introduction

This paper is concerned with the expected average cost optimization of a Markov decision problem with constraints. We study the case in which the state and action spaces are Borel spaces. The cost function may be unbounded, and it is subjected to m expected average cost constraints. We give general conditions for the existence of solutions of the Markov decision problem and for the existence of an optimal stable policy, which is a convex combination of $m + 1$ stable deterministic polices.

It is already known that for Markov decision constraint problems, there exist optimal randomized policies (Beutler and Ross 1985; Borkar 1994; Frid 1972;

J. González-Hernández
Departamento de Probabilidad y Estadística, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Apartado postal 20-726, Admon. No. 20, delegación Álvaro Obregón, 01000 Mexico City, D.F., Mexico
e-mail: juan@sigma.iimas.unam.mx

C.E. Villarreal (✉)
Posgrado en Ingeniería de Sistemas, Facultad de Ingeniería Mecánica y Eléctrica, Universidad Autónoma de Nuevo León, Ciudad Universitaria, 66450 San Nicolás de los Garza, N.L., Mexico
e-mail: cesarevr@gmail.com

Published online: 23 July 2009

González-Hernández and Hernández-Lerma 2005; Haviv 1996; Sennott 1991, 1993). The case of finite, denumerable, and compact state spaces has been widely dealed (Beutler and Ross 1985; Sennott 1993; Borkar 1994; Kurano et al. 2000a; Piunovskiy 1993, 1997; Piunovskiy and Khametov 1991; Tanaka 1991; Hu and Yue 2008). The discounted performance criteria has also already been dealed (Feimberg and Shwartz 1996; González-Hernández and Hernández-Lerma 2005; Hernández-Lerma and González-Hernández 2000; and Sennott 1991). We can see other related aspects (Collins and McNamara 1998; Kurano et al. 2000b; and Yushkevich 1997).

To give a characterization of some optimal solutions of the control problem (Theorems 1 and 2 bellow), we follow the procedure used by González-Hernández and Hernández-Lerma (2005, Theorem 2.6), which is a generalization of a theorem given by Winkler (1988, Theorem 2.1) and which in turn is an extension of a theorem given by Karr (1983, Theorem 2.1).

In the next section we give a brief background of the Markov decision processes, including the model with constraints and the construction of the process. In Sect. 3 we raise the control problem we are interested in, provide the hypothesis to assure the existence of solutions, and set some lemmas in order to prove the main theorem (Theorem 2), which shows the existence of optimal randomized policies in Borel spaces for average criterium. Finally, Sect. 4 shows three examples with optimal policy: the first is related to an inventory problem with optimal stable policy, the second one is an example with stable policies but without optimal stable policies, and the third is an example without stable policies.

2 Preliminaries

We shall use the following concepts and definitions throughout the article.

We suppose that a metric space $\mathbb{S}$ is always endowed with its Borel σ-algebra, which is denoted by $\mathcal{B}(\mathbb{S})$. A *Polish space* is a complete and separable metric space, and a *Borel space* is a subspace of a Polish space. The symbol $\mathcal{M}(\mathbb{S})$ stands for the linear space of finite signed measures on $\mathbb{S}$, and $\mathcal{P}(\mathbb{S})$ is the subset of probability measures.

Given two Borel spaces $\mathbb{S}$ and $\mathbb{S}'$, a *stochastic kernel* on $\mathbb{S}$ given $\mathbb{S}'$ is a real-valued function $(x, B) \mapsto K(B|x)$ on $\mathbb{S}' \times \mathcal{B}(\mathbb{S})$ such that $K(B|\cdot)$ is a measurable function on $\mathbb{S}'$ for each fixed $B \in \mathcal{B}(\mathbb{S})$ and $K(\cdot|x)$ is a probability measure on $\mathcal{B}(\mathbb{S})$ for each $x \in \mathbb{S}'$.

We shall give a brief review of the main concepts of Markov decision processes. A deeper study of these concepts can be seen in several books (Altman 1999; Borkar 1994; Hernández-Lerma and Lasserre 1996; Hu and Yue 2008; and Piunovskiy 1997).

Definition 1 Let m be a nonnegative integer. An *m-constrained Markov decision model* (m-CMDM) is a sequence of $5 + m$ components

$$(\mathbb{X}, \mathbb{A}, A, Q, c, d_1, \ldots, d_m),$$

where:

(a) $\mathbb{X}$ is a nonempty Borel space, called the *state space*.
(b) $\mathbb{A}$ also is a nonempty Borel space, called the *action space*.
(c) $A : \mathbb{X} \longrightarrow \mathcal{B}(\mathbb{A})$ is called the *function of admissible actions*. We denote by $\mathbb{K}$ the set of *admissible pairs* $\{(x, a) \in \mathbb{X} \times \mathbb{A} : a \in A(x)\}$, which we assume to be measurable and containing the graph of a measurable function from $\mathbb{X}$ to $\mathbb{A}$.
(d) Q is a stochastic kernel on $\mathbb{X}$ given $\mathbb{X} \times \mathbb{A}$. It represents the dynamics of the system.
(e) $c : \mathbb{K} \longrightarrow [0, +\infty)$ is a measurable function called the *cost function*.
(f) $d_i : \mathbb{K} \longrightarrow [0, +\infty)$ for $i = 1, \ldots, m$ are measurable functions that we use to define the constraints.

A measurable function $f : \mathbb{X} \longrightarrow \mathbb{A}$ such that $f(x) \in A(x)$ is called a *measurable selector*. Since $\mathbb{K}$ contains the graph of a measurable map (see Definition 1(c) above), the set of measurable selectors is nonempty.

Definition 2 Randomized, deterministic, and m-randomized control.

(a) A *randomized control* is a stochastic kernel φ on $\mathbb{A}$ given $\mathbb{X}$, such that $\varphi(A(x)|x) = 1$ for each $x \in \mathbb{X}$.
(b) If φ is a randomized control and there is a measurable selector f such that $\varphi(\{f(x)\}|x) = 1$ for all $x \in \mathbb{X}$, then we say that φ is a *deterministic control*, that is,

$$\varphi(\cdot|x) = \delta_{f(x)}(\cdot) \quad \text{for } x \in \mathbb{X},$$

where δ_y denotes the Dirac measure concentrated on y. (Note that each deterministic control is identified with a measurable selector and vice versa.)
(c) A randomized control φ is an *m-randomized control* if there are m measurable selectors $f_1, \ldots, f_m$ and m nonnegative numbers $\alpha_1, \ldots, \alpha_m$ such that $\sum_{i=1}^{m} \alpha_i = 1$ and

$$\varphi(\cdot|x) = \sum_{i=1}^{m} \alpha_i \delta_{f_i(x)}(\cdot).$$

We need to define $\mathbb{H}_n$, "the set of possible histories up to time n." Let $\mathbb{H}_0 := \mathbb{X}$ and $\mathbb{H}_n := \mathbb{K}^n \times \mathbb{X}$ for $n = 1, 2, \ldots$. An arbitrary element $h_n \in \mathbb{H}_n$ is represented by $h_n = (x_0, a_0, x_1, a_1, \ldots, x_n)$, where $(x_i, a_i) \in \mathbb{K}$ for $i = 0, 1, \ldots, n-1$ and $x_n \in \mathbb{X}$.

Definition 3 Policies.

(a) A *policy* is a sequence $\pi = (\pi_n)$ of stochastic kernels on $\mathbb{A}$ given $\mathbb{H}_n$, satisfying

$$\pi_n\big(A(x_n)|h_n\big) = 1$$

for each history $h_n = (x_0, a_0, \ldots, x_{n-1}, a_{n-1}, x_n)$ in $\mathbb{H}_n$. We denote by Π the set of all policies.
(b) We say that a policy $\pi = (\pi_n)$ is *stationary* if there is a randomized control φ such that, for every history $h_n = (x_0, a_0, \ldots, x_n)$, we have

$$\pi_n(\cdot|h_n) = \varphi(\cdot|x_n).$$

We specify this dependence by writing π^{φ}. Also, we denote by Φ the set of randomized controls, and by $\widehat{\Phi}$ the set of stationary policies.

(c) A stationary policy $\pi^{\varphi} = (\pi_n)$ is called *deterministic* if the corresponding randomized control φ given in (b) is deterministic, that is, there is a measurable selector f such that, for every history $h_n = (x_0, a_0, \ldots, x_n)$, we have

$$\pi_n(\cdot|h_n) = \delta_{f(x_n)}(\cdot).$$

Let us denote by Φ_1 the set of deterministic controls, and by $\widehat{\Phi}_1$ the set of deterministic stationary policies.

(d) Let m be a positive integer. A stationary policy is said to be *m-randomized* policy if it is a convex combination of m deterministic policies. The corresponding randomized control is called *m-randomized* control. We denote by Φ_m the set of m-randomized controls, and by $\widehat{\Phi}_m$ the set of m-randomized policies.

These definitions establish bijections between the set Φ of randomized controls and the set $\widehat{\Phi}$ of stationary policies; the set Φ_1 of deterministic controls and the set $\widehat{\Phi}_1$ of deterministic policies; the set Φ_m of m-randomized controls and the set $\widehat{\Phi}_m$ of m-randomized policies. We have the following inclusions diagram:

$$\widehat{\Phi}_1 \subset \widehat{\Phi}_m \subset \widehat{\Phi}_{m+1} \subset \widehat{\Phi} \subset \Pi.$$

Construction of the process Suppose that $\mathbb{S}_1$, $\mathbb{S}_2$, and $\mathbb{S}_3$ are metric spaces, $\mu \in \mathcal{P}(\mathbb{S}_1)$, φ_1 is a stochastic kernel on $\mathbb{S}_2$ given $\mathbb{S}_1$, and φ_2 is a stochastic kernel on $\mathbb{S}_3$ given $\mathbb{S}_2$. The *product measure* $\mu \otimes \varphi_1$ on $\mathbb{S}_1 \times \mathbb{S}_2$ is defined as the measure generated by the formula

$$\mu \otimes \varphi_1(B \times C) := \int_B \varphi_1(C|s_1)\mu\,(\mathrm{d}s_1) \tag{1}$$

for $B \in \mathcal{B}(\mathbb{S}_1)$ and $C \in \mathcal{B}(\mathbb{S}_2)$. Also, the *kernel product* $\varphi_1 \otimes \varphi_2$ on $\mathbb{S}_2 \times \mathbb{S}_3$ given $\mathbb{S}_1$ is defined as the kernel generated by the formula

$$\varphi_1 \otimes \varphi_2(C \times D|s_1) := \int_C \varphi_2(D|s_2)\varphi_1\,(\mathrm{d}s_2|s_1) \tag{2}$$

for $C \in \mathcal{B}(\mathbb{S}_2)$ and $D \in \mathcal{B}(\mathbb{S}_3)$.

Let us construct a discrete-time stochastic process on $\mathbb{X} \times \mathbb{A}$. Given an initial distribution $\nu \in \mathcal{P}(\mathbb{X})$ (the distribution of x_0) and a policy $\pi = (\pi_n)$, by the Ionescu–Tulcea Theorem (Ash 1972, Theorem 2.7.2; Hinderer 1970, Sect. 11; Loève 1977, pp. 137–139), we have a stochastic process on $\mathbb{X} \times \mathbb{A}$ such that the initial distribution μ_0 (of the process) is $\nu \otimes \pi_0$, and the joint distribution μ_{n+1} of $(x_0, a_0, x_1, a_1, \ldots, x_{n+1}, a_{n+1})$ is $(\mu_n \otimes \pi_n) \otimes Q$, where μ_n is the distribution of $(x_0, a_0, x_1, a_1, \ldots, x_n, a_n)$ for $n \in \{0, 1, 2, \ldots\}$. We consider the measurable space of trajectories of the process $\Omega := (\mathbb{X} \times \mathbb{A})^{\infty}$ and $\mathrm{P}_{\nu}^{\pi} \in \mathcal{P}(\Omega)$ such that $\mathrm{P}_{\nu}^{\pi}(B_n \times (\mathbb{X} \times \mathbb{A})^{\infty}) = \mu_n(B_n)$ for $B_n \in \mathcal{B}((\mathbb{X} \times \mathbb{A})^{n+1})$. Note that P_{ν}^{π} is supported on $\mathbb{K}^{\infty}$.

If π^{φ} is a stationary policy and φ its corresponding randomized control, we put $\mathrm{P}_{\nu}^{\varphi}$ in place of $\mathrm{P}_{\nu}^{\pi^{\varphi}}$. Also, if π is a deterministic policy and f its corresponding measurable selector, we put P_{ν}^{f} in place of P_{ν}^{π}. For a random variable Y on Ω, we denote the

expected value of Y by $\mathrm{E}_\nu^\pi(Y)$, that is, $\mathrm{E}_\nu^\pi(Y) = \int Y \,\mathrm{dP}_\nu^\pi$. Furthermore, in the case $\nu = \delta_x$, we denote P_ν^π and E_ν^π by P_x^π and E_x^π, respectively, instead.

3 Solution of a control problem

The control problem (CP) consists in finding a policy π and an initial distribution ν that minimize the *objective function* (or *performance index function*) $J : \Pi \times \mathcal{P}(\mathbb{X}) \longrightarrow \mathbb{R} \cup \{+\infty\}$ given by

$$\text{minimize:} \quad J(\pi, \nu) := \limsup_{N\to\infty} \frac{1}{N+1} \mathrm{E}_\nu^\pi \left(\sum_{t=0}^{N} c(x_t, a_t) \right), \tag{3}$$

subject to

$$J_i(\pi, \nu) := \limsup_{n\to\infty} \frac{1}{N+1} \mathrm{E}_\nu^\pi \left(\sum_{t=0}^{N} d_i(x_t, a_t) \right) \leq k_i \tag{4}$$

for $i \in \{1, \ldots, m\}$ and $k_1, \ldots, k_m \geq 0$. We could interpret the functions $J_i(\pi, \nu)$ as long-run average costs that needed to keep bounded.

Let Δ be the set of feasible pairs for the CP, that is,

$$\Delta := \big\{(\pi, \nu) \in \Pi \times \mathcal{P}(\mathbb{X}) : J(\pi, \nu) < \infty \text{ and } J_i(\pi, \nu) \leq k_i,\ i \in \{1, \ldots, m\}\big\}. \tag{5}$$

When π^* is a policy and ν^* is an initial distribution such that $(\pi^*, \nu^*) \in \Delta$ and

$$J(\pi^*, \nu^*) = \inf\big\{J(\pi, \nu) : (\pi, \nu) \in \Delta\big\}, \tag{6}$$

we say that the pair (π^*, ν^*) is an *optimal solution* of the CP.

Hypothesis 1

(a) CP is consistent. That is, the set of feasible pairs Δ is non empty.
(b) $c \geq 0$ is an inf-compact function, that is, for each $r \in \mathbb{R}$, the set $\{(x, a) \in \mathbb{K} : c(x, a) \leq r\}$ is compact.
(c) Each d_i is a nonnegative lower semicontinuous function.
(d) The stochastic kernel Q is weakly continuous, that is, $\int_{\mathbb{X}} u(y) Q(\mathrm{d}y|\cdot) \in C_b(\mathbb{K})$ for every function $u \in C_b(\mathbb{X})$ (where for a topological space $\mathbb{S}$, $C_b(\mathbb{S})$ denotes the space of bounded continuous real functions).

Under this hypothesis, we have the following theorem (Hernández-Lerma et al. 2003, Theorem 3.2).

Theorem 1 *Under Hypothesis 1, there is an optimal solution of the* CP.

If $\mu \in \mathcal{M}(\mathbb{X} \times \mathbb{A})$, there are a randomized control φ and a signed measure $\hat{\mu} \in \mathcal{M}(\mathbb{X})$ such that

$$\mu(B \times C) = \hat{\mu} \otimes \varphi(B \times C) = \int_B \varphi(C|x) \hat{\mu}(\mathrm{d}x) \tag{7}$$

for $B \in \mathcal{B}(\mathbb{X})$ and $C \in \mathcal{B}(\mathbb{A})$. The measure $\hat{\mu}$ in (7) is called the *marginal* measure of μ on $\mathbb{X}$, and it is defined as $\hat{\mu} := \mu(\cdot \times \mathbb{A})$. Moreover, for each $C \in \mathcal{B}(\mathbb{A})$, the function $\varphi(C|\cdot)$ is the Radon–Nikodým derivative of $\mu(\cdot \times C)$ with respect to $\hat{\mu}$. Conversely, if φ is a randomized control and $\hat{\mu} \in \mathcal{M}(\mathbb{X})$, there is a signed measure $\mu \in \mathcal{M}(\mathbb{X} \times \mathbb{A})$ such that (1) is satisfied.

If $g : \mathbb{K} \longrightarrow \mathbb{R}$ is measurable, φ is a randomized control, and f is a measurable selector, for simplicity, we denote

$$g(x, f) := g\big(x, f(x)\big)$$

and

$$g(x, \varphi) := \int g(x, a)\varphi\,(\mathrm{d}a|x).$$

Definition 4 A measure $\mu = \hat{\mu} \otimes \varphi \in \mathcal{M}(\mathbb{X} \times \mathbb{A})$ is said to be *stable* if

(a) $$\langle \mu, c\rangle := \int c(x, \varphi)\hat{\mu}\,(\mathrm{d}x) < +\infty$$

and

(b) $$\hat{\mu}(B) = \int Q(B|x, \varphi)\hat{\mu}\,(\mathrm{d}x) \quad \text{for } B \in \mathcal{B}(\mathbb{X}).$$

Also, a randomized control φ is called *stable control* if there exists $\hat{\mu} \in \mathcal{M}(\mathbb{X})$ such that the signed measure $\hat{\mu} \otimes \varphi$ is stable. We denote the set of probability measures stables concentrated on $\mathbb{K}$ by $\mathcal{P}_s(\mathbb{K})$.

Remark 1 For a policy $\pi = (\pi_n) \in \Pi$ and initial distribution ν, let us consider the occupation measure in the n-step $\mu_n(B) := \mathrm{E}_\nu^\pi(\delta_{(x_n,a_n)}(B))$, and let us disintegrate this measure as $\mu_n = \hat{\mu}_n \otimes \varphi_n$; then the Markovian policy $\pi' = (\varphi_n)$ is equivalent to the former policy in the sense that $J(\pi, \nu) = J(\pi', \nu)$ and $J_i(\pi, \nu) = J_i(\pi', \nu)$ for $i \in \{1, 2, \ldots, m\}$. Hence, Markovian policies are sufficient for CP. Even more, Lemma 1 below shows that the search of optimal policies can be reduced to stationary and stable policies.

Let $\hat{\mu}$ be an initial distribution, and φ a stable control. The Individual Ergodic Theorem (Yoshida 1978, p. 338; Hernández-Lerma and Lasserre 1996, Theorem E11) implies that if $\hat{\mu} \otimes \varphi$ is stable, then the long-run average value $J(\varphi, \hat{\mu})$ of $c(x_t, a_t)$ in (3) is given just by the limit rather than the upper limit, i.e.,

$$J\big(\pi^\varphi, \hat{\mu}\big) = \lim_{N\to\infty} \frac{1}{N+1} \mathrm{E}_{\hat{\mu}}^{\pi^\varphi} \left(\sum_{t=0}^{N} c(x_t, a_t) \right) = \langle \mu, c\rangle,$$

and analogously $J_i(\pi^\varphi, \hat{\mu}) = \langle \mu, d_i\rangle$ for $i \in \{1, 2, \ldots, m\}$. In brief, we have

$$\mu \in \mathcal{P}_s(\mathbb{K}) \quad \Longrightarrow \quad \begin{cases} \text{(a)}\, J(\pi^\varphi, \hat{\mu}) = \langle \mu, c\rangle & \text{and} \\ \text{(b)}\, J_i(\pi^\varphi, \hat{\mu}) = \langle \mu, d_i\rangle & \text{for} i \in \{1, 2, \ldots, m\}. \end{cases} \tag{8}$$

The key of the CP is that we can reduce the search of optimum policies to the search of optimal stable controls. If this is the case, the CP can be reformulated as

$$\begin{aligned} &\text{minimize:} \quad \langle \mu, c \rangle, \\ &\text{subject to} \quad \langle \mu, d_i \rangle \leq k_i, \quad \text{for } i \in \{1, 2 \ldots, m\}. \end{aligned} \tag{9}$$

We shall denote the problem (9) by CP′.

As we can see, CP′ is a linear programming problem with m constraints whose dimension is not necessarily finite. The following lemma (Hernández-Lerma et al. 2003, Lemma 3.5) gives us a guide to do such a reformulation.

Lemma 1 (Reduction of CP to the set of stable controls) *Under Hypothesis 1, for each feasible pair* $(\pi, \nu) \in \Delta$ *of the* CP, *there is a stable measure* $\mu = \hat{\mu} \otimes \varphi$, *such that*

(a) $(\pi^{\varphi}, \hat{\mu}) \in \Delta$, *and*
(b) $J(\pi, \nu) \geq J(\pi^{\varphi}, \hat{\mu}) = \langle \mu, c \rangle$.

Remark 2 If we consider the sequence (μ_n) of empirical measures on $\mathbb{X} \times \mathbb{A}$ given by

$$\mu_n(B) := \frac{1}{n+1} \mathrm{E}^{\pi}_{\hat{\mu}} \left(\sum_{t=0}^{n} \delta_{(x_t, a_t)}(B) \right) \quad \text{for } n \in \mathbb{N},$$

then we need Hypothesis 1(b) to apply Prokhorov's Theorem. In last section we provide two examples showing that without this hypothesis we cannot assure the reduction of CP to stable policies. Even more, if we can apply Prokhorov's Theorem to the sequence (μ_n), then there is a convergent subsequence (μ_{k_n}) of (μ_n). So, if μ is the limit of (μ_{k_n}), then we need Hypothesis 1(c) and (d) to show that the limit measure μ satisfies Lemma 1(a) and (b).

Lemma 1 and Theorem 1 yield the existence of an optimal solution $(\pi^{\varphi^*}, \nu^*) \in \Phi \times \mathcal{P}(\mathbb{X})$ for the CP, with $\mu^* = \nu^* \otimes \varphi^*$ stable.

With the next hypothesis we can characterize some optimal policies. Let us define some concepts that will be used in Hypothesis 2.

Let μ be a finite (nonnegative) measure on $\mathcal{B}(Y)$. The measure μ is said to be *regular* if, for every $B \in \mathcal{B}(Y)$, $\mu(B) = \sup\{\mu(F) : F \subset B \text{ and } F \text{ is closed}\}$. The measure μ is said to be τ-*smooth* if, for each decreasing net (F_α) of closed subsets of Y, we have $\mu(\bigcap_\alpha F_\alpha) = \inf_\alpha \mu(F_\alpha)$. A probability P or a probability space $(\Omega, \mathcal{F}, P)$ is *nonatomic* if $P(A) > 0$ implies that there is $B \in \mathcal{F}$ such that $B \subset A$ and $0 < P(B) < P(A)$ (Billingsley 1995, p. 35). In our particular case in which $\mathcal{F}$ is the Borel σ-algebra of some Borel space, the fact that P is nonatomic is reduced to $P(\{x\}) = 0$ for all $x \in \Omega$.

Hypothesis 2

(a) $\mathbb{A}$ is a topological space such that every probability measure in $\mathcal{B}(\mathbb{A})$ is τ-smooth and regular.

(b) The stochastic kernel Q is nonatomic, that is, for every couple $(x, a) \in \mathbb{K}$, the probability measure $Q(\cdot|x, a)$ is nonatomic.

Theorem 2 *Under Hypotheses 1 and 2, there is an optimal solution* $(\pi^{\varphi^*}, \nu^*) \in \Delta$ *for* CP *such that* φ^* *is a stable* $(m+1)$*-randomized control.*

We can see that, for a Borel space, each probability measure is regular and τ-smooth (Munkres 1975, Theorems 1.2 and 1.3, and Exercise 7 of Sect. 4.1). To prove Theorem 2, we shall use several lemmas. In the sequel of this section we fix $\nu^* \in \mathcal{P}(\mathbb{X})$ such that, for some $\varphi^* \in \Phi$, we have an optimal solution (π^{φ^*}, ν^*) for CP, with $\nu^* \otimes \varphi^*$ stable. By Lemma 1 and Theorem 1, we get that the minimum value ρ^* of CP can be written as

$$\rho^* = \inf\big\{\langle \nu^* \otimes \varphi, c\rangle : \varphi \in \Phi_s\big\},$$

where $\Phi_s := \{\varphi \in \Phi : \nu^* \otimes \varphi \in \mathcal{P}_s(\mathbb{K})\}$.

Let $\Phi_{1,s} := \{\varphi \in \Phi_1 : \nu^* \otimes \varphi \in \mathcal{P}_s(\mathbb{K})\}$.

Lemma 2 *The set of extreme points of* Φ_s *is* $\Phi_{1,s}$.

Proof Let $\varphi \in \Phi_s$. Suppose that there are $x \in \mathbb{X}$ and $B \in \mathcal{B}(\mathbb{A})$ such that $0 < \varphi(B|x) < 1$. Then φ is not an extreme point of Φ_s, because we can express φ as $\varphi = \alpha\varphi_1 + (1-\alpha)\varphi_2$ with $0 < \alpha < 1$ and $\varphi_1 \neq \varphi_2$. Indeed, taking $\alpha := \varphi(B|x)$, let

$$\varphi_1(\cdot|y) := \begin{cases} \frac{\varphi(\cdot \cap B|x)}{\alpha} & \text{if } y = x, \\ \varphi(\cdot|y) & \text{if } y \neq x, \end{cases}$$

and

$$\varphi_2(\cdot|y) := \begin{cases} \frac{\varphi(\cdot \cap B^c|x)}{1-\alpha} & \text{if } y = x, \\ \varphi(\cdot|y) & \text{if } y \neq x, \end{cases}$$

where $B^c := \mathbb{A} \setminus B$. Also, Hypothesis 2(b) and the definition of stable measure imply that $\nu^* \otimes \varphi_1$ and $\nu^* \otimes \varphi_2$ are stable. □

Let $\Phi' := \{\varphi \in \Phi : \nu^* \otimes \varphi \in \mathcal{P}_s(\mathbb{K})$ and $(\pi^\varphi, \nu^*) \in \Delta\}$ and $\Phi'_{m+1} := \{\varphi \in \Phi_{m+1} : \nu^* \otimes \varphi \in \mathcal{P}_s(\mathbb{K})$ and $(\pi^\varphi, \nu^*) \in \Delta\}$. We have the following lemma, which is an analogous result to that given for Markov decision processes with discounted cost in González-Hernández and Hernández-Lerma (2005, Theorem 2.6).

Lemma 3 *The set* Φ' *is convex, and* Φ'_{m+1} *is the set of its extreme points.*

Proof The proof of a theorem given by González-Hernández and Hernández-Lerma (2005, Theorem 2.6) works in the present case if we use Lemma 2, taking stable policies, initial distribution ν^*, and posing Φ_s, $\Phi_{1,s}$, Φ' and Φ'_{m+1} in place of Φ, $\mathbb{F}$, Δ and $\mathcal{R}^0_{m+1}$ in the cited article, respectively. □

Let $\nu^* \otimes \Phi' := \{\nu^* \otimes \varphi : \varphi \in \Phi'\}$ and $\nu^* \otimes \Phi'_{m+1} := \{\nu^* \otimes \varphi : \varphi \in \Phi'_{m+1}\}$. Note that by Lemma 3, the set of extreme points of $\nu^* \otimes \Phi'$ is $\nu^* \otimes \Phi'_{m+1}$. Moreover, we can deduce the next lemma.

Lemma 4 *The set of extreme points of* $\nu^* \otimes \Phi'$ *is* $\nu^* \otimes \Phi'_{m+1}$. *Moreover,* $\nu^* \otimes \Phi'$ *is convex*, (*weakly*) *closed*, *and sequentially compact*.

Proof By Lemma 3, the set $\nu^* \otimes \Phi'$ is convex, and $\nu^* \otimes \Phi'_{m+1}$ is the set of its extreme points.

Let $\mu_k = \nu^* \otimes \varphi_k$ with $\varphi_k \in \Phi'$ such that (μ_k) converges weakly to a measure $\mu \in \mathcal{M}(\mathbb{X} \times \mathbb{A})$. Observe that $\nu^* = \hat{\mu}_k$, so $\hat{\mu} = \nu^*$ (recall that $\hat{\rho} = \rho(\cdot \times \mathbb{A})$ for $\rho \in \mathcal{M}(\mathbb{X} \times \mathbb{A})$). Hence, we get $\mu = \nu^* \otimes \varphi$ for some $\varphi \in \Phi$. We need to prove that $\varphi \in \Phi'$.

Let $u \in C_b(\mathbb{X})$. By Hypothesis 1(d) together with weak convergence and Definition 4,

$$\int u(x)\nu^*(\mathrm{d}x) = \int\int \left(\int u(y)Q(\mathrm{d}y|x,a)\right)\varphi_k(\mathrm{d}a|x)\nu^*(\mathrm{d}x)$$
$$\longrightarrow \int\int \left(\int u(y)Q(\mathrm{d}y|x,a)\right)\varphi(\mathrm{d}a|x)\nu^*(\mathrm{d}x);$$

therefore, $\nu^* = \mu \otimes Q$, that is,

$$\nu^*(B) = \int Q(B|x,\varphi)\nu^*(\mathrm{d}x),$$

which means $\varphi \in \Phi'$.

Finally, by Hypothesis 1(b) and Prokhorov's Theorem (Bourbaki 1969, No. 5.5), $\nu^* \otimes \Phi'$ is sequentially compact. □

End of the proof of Theorem 2 Let μ be an extreme point of $\nu^* \otimes \Phi'$. From a theorem given by Winkler (1988, Theorem 2.1), there are $\mu_1, \ldots, \mu_{m+1} \in \nu^* \otimes \Phi_s$ and $\alpha_1, \ldots, \alpha_{m+1} \in [0,1]$ such that $\sum_{k=1}^{m+1} \alpha_k = 1$ and $\mu = \sum_{k=1}^{m+1} \alpha_k \mu_k$. Following the proof of a theorem given by Piunovskiy (1997, 1993, Theorem 10, Sect. 2.2.3), for each $k \in \{1, \ldots, m+1\}$, there is a $\varphi_k \in \Phi_{1,s}$ such that $\mu_k = \nu^* \otimes \varphi_k$. Let $\varphi = \sum_{k=1}^{m+1} \alpha_k \varphi_k$. Then, by Lemma 3, φ is an extreme point of Φ'.

We have that $\nu^* \otimes \Phi'$ is a subset of a metric linear space with the Prokhorov metric (Billingsley 1999, p. 72). Since $\nu^* \otimes \Phi'$ is closed and sequentially compact, it is compact.

By the Krein–Milman Theorem (Phelps 1966, p. 59), the minimum of the CP is attained in an extreme point $\nu^* \otimes \varphi^*$ of $\nu^* \otimes \Phi'$. □

4 Examples

An example with optimal stable policy Let us consider the following inventory problem. Assume that we have a store with finite capacity M and n different cereals. Let

x_t^i be the observed stock volume of the ith kind of cereal at the beginning of the stage t when it is not negative and minus the nonsatisfied demand to spurt in the following period, which will be given to the client directly if it is negative. The manager then orders a quantity a_t^i of the ith kind of cereal. The state and control variables are $x_t = (x_t^1, \ldots, x_t^n)$ and $a_t = (a_t^1, \ldots, a_t^n)$, respectively, and are such that $x_t^i + a_t^i \geq 0$ and $\sum_{j=1}^n (x_t^j + a_t^j) \leq M$ for all $i \in \{1, \ldots, n\}$ and $t \in \{0, 1, 2, \ldots\}$.

Let D_t^i be the demand of the ith kind of cereal through the period t. We assume that $(D_t^i)_{t=0}^\infty$ is a sequence of continuous identically distributed random variables and $0 \leq D_t^i \leq R$ for each $i \in \{1, \ldots, n\}$ and some constant $R > M$. The dynamics of the process is given by

$$x_{t+1}^i = x_t^i + a_t^i - D_t^i.$$

In this example, the state space is $\mathbb{X} = \{(y_1, \ldots, y_n) \in [-R, M]^n : \sum_{i=1}^n y_i \leq M\}$, the action space is $\mathbb{A} = [0, M + R]^n$, and

$$A(y_1, \ldots, y_n) = \left\{ (b_1, \ldots, b_n) \in \mathbb{A} : \sum_{i=1}^n (b_i + y_i) \leq M \text{ and } b_j + y_j \geq 0 \text{ for all } j \in \{1, \ldots, n\} \right\}.$$

The objective of the control problem is to maximize

$$K(\pi, \nu) := \liminf_{N\to\infty} \frac{1}{N+1} \mathrm{E}_\nu^\pi \left(\sum_{t=0}^N \sum_{i=1}^n \left(P_i\left(x_t^i + a_t^i - x_{t+1}^i\right) - C_i a_t^i - S\left(x_t^i + a_t^i\right)\right)\right),$$

subject to

$$J_1(\pi, \nu) := \limsup_{N\to\infty} \frac{1}{N+1} \mathrm{E}_\nu^\pi \left(\sum_{t=0}^N \sum_{i=1}^n P_i \max\left\{-x_t^i, 0\right\}\right) \leq k_1,$$

where for each i, P_i is the price, C_i is the cost, S is the storage cost, and k_1 is a given positive number. Note that the expression $\sum_{i=1}^n (P_i(x_t^i + a_t^i - x_{t+1}^i) - C_i a_t^i - S(x_t^i + a_t^i))$ depends of the three variables x_t, a_t, and x_{t+1}. Under the same constraint, the problem is equivalent to minimize

$$J(\pi, \nu) := \limsup_{N\to\infty} \frac{1}{N+1} \mathrm{E}_\nu^\pi \left(\sum_{t=0}^N c(x_t, a_t)\right),$$

where $c(x_t, a_t) := \sum_{i=1}^n (C_i a_t^i + S(x_t^i + a_t^i) - P_i a_t^i + R P_i)$. Effectively, to maximize

$$\liminf_{N\to\infty} \frac{1}{N+1} \mathrm{E}_\nu^\pi \left(\sum_{t=0}^N \sum_{i=1}^n \left(P_i\left(x_t^i + a_t^i - x_{t+1}^i\right) - C_i a_t^i - S\left(x_t^i + a_t^i\right)\right)\right)$$

is equivalent to minimize

$$\limsup_{N\to\infty}\frac{-1}{N+1}\mathrm{E}_{\nu}^{\pi}\left(\sum_{t=0}^{N}\sum_{i=1}^{n}\left(P_i\left(x_t^i+a_t^i-x_{t+1}^i\right)-C_i a_t^i-S\left(x_t^i+a_t^i\right)\right)\right),$$

but

$$\begin{aligned}
&\limsup_{N\to\infty}\frac{-1}{N+1}\mathrm{E}_{\nu}^{\pi}\left(\sum_{t=0}^{N}\sum_{i=1}^{n}\left(P_i\left(x_t^i+a_t^i-x_{t+1}^i\right)-C_i a_t^i-S\left(x_t^i+a_t^i\right)\right)\right)\\
&\quad=\limsup_{N\to\infty}\frac{1}{N+1}\mathrm{E}_{\nu}^{\pi}\left(\sum_{t=0}^{N}c(x_t,a_t)+\sum_{t=0}^{N}\sum_{i=1}^{n}P_i\left(x_{t+1}^i-x_t^i-R\right)\right)\\
&\quad=\limsup_{N\to\infty}\frac{1}{N+1}\mathrm{E}_{\nu}^{\pi}\left(\sum_{t=0}^{N}c(x_t,a_t)+\sum_{i=1}^{n}P_i\left(x_{N+1}^i-x_0^i\right)-(N+1)R\sum_{i=1}^{n}P_i\right)\\
&\quad=\limsup_{N\to\infty}\left(\frac{1}{N+1}\mathrm{E}_{\nu}^{\pi}\sum_{t=0}^{N}c(x_t,a_t)+\frac{1}{N+1}\mathrm{E}_{\nu}^{\pi}\sum_{i=1}^{n}P_i\left(x_{N+1}^i-x_0^i\right)-R\sum_{i=1}^{n}P_i\right),
\end{aligned}$$

and since $\sum_{i=1}^{n}P_i(x_{N+1}^i-x_0^i)$ is bounded, we have

$$\begin{aligned}
&\limsup_{N\to\infty}\left(\frac{1}{N+1}\mathrm{E}_{\nu}^{\pi}\sum_{t=0}^{N}c(x_t,a_t)+\frac{1}{N+1}\mathrm{E}_{\nu}^{\pi}\sum_{i=1}^{n}P_i\left(x_{N+1}^i-x_0^i\right)-R\sum_{i=1}^{n}P_i\right)\\
&\quad=\limsup_{N\to\infty}\left(\frac{1}{N+1}\mathrm{E}_{\nu}^{\pi}\sum_{t=0}^{N}c(x_t,a_t)-R\sum_{i=1}^{n}P_i\right).
\end{aligned}$$

Now, $R\sum_{i=1}^{n}P_i$ is constant with respect to N and (π,ν), therefore to maximize $K(\pi,\nu)$ is equivalent to minimize $J(\pi,\nu)=\limsup_{N\to\infty}\frac{1}{N+1}\mathrm{E}_{\nu}^{\pi}(\sum_{t=0}^{N}c(x_t,a_t))$.

We have that the new equivalent problem fulfills Hypotheses 1 and 2.

To illustrate Theorem 2 in this example, assume that we have only a kind of cereal, $D_t:=D_t^1$ is uniformly distributed in $[0,R]$, and $P_1>S+C_1$. For each $\gamma\in[0,1]$, let us define the measurable selector f_γ as

$$f_\gamma(x):=\begin{cases}\gamma M-x & \text{if } x\le\gamma M,\\ 0 & \text{if } x>\gamma M,\end{cases}$$

let φ_γ be the deterministic control such that $\varphi_\gamma(\{f_\gamma(x)\}|x)=1$, and $\pi_\gamma:=\pi^{\varphi_\gamma}$. The policy π_γ represents the action to start every period with a inventory level γ if possible. If λ is the Lebesgue measure in $\mathbb{R}$ and $\nu_\gamma(\cdot):=\frac{\lambda(\cdot\cap[\gamma M-R,\gamma M])}{R}$, then $\nu_\gamma\otimes\varphi_\gamma$ is a stable measure, so the deterministic control φ_γ is stable. Moreover,

$$\lim_{t\to\infty}\mathrm{E}_{\nu_\gamma}^{\pi_\gamma}\left(P_1\max\{-x_t,0\}\right)=\int_{\gamma M-R}^{0}\frac{-P_1x}{R}\,\mathrm{d}x=\frac{P_1(R-\gamma M)^2}{2R};$$

thus, π_γ fulfills the constraint if and only if $\frac{P_1(R-\gamma M)^2}{2R} \leq k_1$, that is, π_γ fulfills the constraint if and only if $\gamma \geq \frac{R-\sqrt{\frac{2k_1R}{P_1}}}{M}$. Now, if $\frac{R-\sqrt{\frac{2k_1R}{P_1}}}{M} \leq 1$, an optimal solution for CP is given by $(\pi_{\gamma^*}, \nu_{\gamma^*})$, where $\gamma^* := \max\{\frac{R-\sqrt{\frac{2k_1R}{P_1}}}{M}, 0\}$.

An example without optimal stable policies Let $\mathbb{X} = \mathbb{Q} \cap (-1, 1)$ be the state space, and let $\mathbb{A} = \{-1, 1\}$ be the action space. The dynamics of the system is given by the conditional probability $Q(\{\alpha x + (1-\alpha)a\}|x, a) = 1$ for every $x \in \mathbb{X}$, where α is a fixed number in $(0, \frac{1}{2}] \cap \mathbb{Q}$. The cost function is given by $c(x, a) := 1 + x$.

In this case the definitions of stable measures $\mu = \hat{\mu} \otimes \varphi$ and stable policies φ are the following:

$$\text{(a)} \quad J(\pi^\varphi, \hat{\mu}) = \sum_{x \in \mathbb{X}} \sum_{a \in \mathbb{A}} c(x, a)\varphi\big(\{a\} \mid x\big)\hat{\mu}\big(\{x\}\big) < +\infty$$

and

$$\text{(b)} \quad \hat{\mu}(B) = \sum_{y \in \mathbb{X}} \sum_{a \in \mathbb{A}} Q(B|y, a)\varphi\big(\{a\}|y\big)\hat{\mu}\big(\{y\}\big), \quad \text{for every } B \subset \mathbb{X}.$$

If we take two points with the same image $x = \alpha x_0 + (1-\alpha)a_0 = \alpha x_1 + (1-\alpha)a_1$, then the only solution in $\mathbb{X}$ fulfills $a_0 = a_1$ and $x_0 = x_1$. Note that applying the former definition (b) in the case that B is a singleton $\{x\}$, the first sum of this definition has at most a nonzero term. Hence, the stable policies are deterministic policies. Let π^{φ_f} any of these policies, where φ_f is the control corresponding to the measurable selector f. We have, for any singleton $\{x_1\}$ with positive measure $\hat{\mu}(\{x_1\})$, that (b) becomes

$$\hat{\mu}\big(\{x_1\}\big) = Q\big(\big\{\alpha x_0 + (1-\alpha) f(x_0)\big\}|x_0, f(x_0)\big)\hat{\mu}\big(\{x_0\}\big) = \hat{\mu}\big(\{x_0\}\big).$$

Now we apply the same argue to x_1 to obtain x_2, and so on. In this way we obtain a sequence $(x_i)_{i=0}^{\infty}$ such that $\hat{\mu}(\{x_0\}) = \hat{\mu}(\{x_1\}) = \hat{\mu}(\{x_2\}) = \cdots$. We can not have an infinite number of equiprobable points; therefore, there are two natural numbers n, m with $m < n$ such that $x_m = x_n$. We can suppose that $x_0 = x_m$. In this way we have

$$a_0\alpha^{n-1}(1-\alpha) + a_1\alpha^{n-2}(1-\alpha) + \cdots + a_{n-1}(1-\alpha) = x_0\big(1-\alpha^n\big), \tag{10}$$

where $a_i \in \{-1, 1\}$ for $i \in \{0, 1, \ldots, n-1\}$. For any stable measure, the points in the support of this measure satisfies (10). Then the points x_i in a "cycle" satisfy

$$x_i = a_0\alpha^{i-1}(1-\alpha) + a_1\alpha^{i-2}(1-\alpha) + \cdots + a_{i-1}(1-\alpha) \quad \text{for } i \in \{1, \ldots, n\}, \tag{11}$$

where $a_i \in \{-1, 1\}$ for $i \in \{0, 1, \ldots, n-1\}$.

The average expected cost for one cycle of (11) is

$$J(\pi^{\varphi_f}, \nu) = \frac{1}{n}\left(n + x_0\frac{1-\alpha^n}{1-\alpha} + a_0(1-\alpha)^{n-1} + \cdots + a_{n-2}(1-\alpha)\right),$$

where the initial distribution is given by $\hat{\mu}(x_i) = \frac{1}{n}$ for the points x_i satisfying (11). Note that $J(\pi_{\varphi_f}, \nu)$ is positive because c is strictly positive.

On the other hand, let f^* be the function given by $f^*(x) := -1$, let x_0 be any element in $\mathbb{X}$, and let us define x_i again as in (11) but now not in a cycle. By a straightforward calculation we get $J(\pi^{\varphi_{f^*}}, \nu) = 0$. Hence, the optimal policy is not a stable policy.

An example without stable policies Now let us consider the same model as in the former example but with the space

$$\mathbb{Y} := \left\{ x \in \mathbb{R}\backslash\mathbb{Q} : x = \alpha^i x_0 + (1-\alpha) \sum_{k=0}^{i-1} a_k \alpha^{i-k-1} \right.$$
$$\left. \text{for } a_i \in \{-1, 1\} \text{ and } i \in \{0, 1, 2, \dots\} \right\},$$

where x_0 is any fixed irrational number in $(-1, 1)$. Then the same formula (10) holds for the stable policies, but its solution are rational numbers. Hence, there are no solutions in $\mathbb{Y}$. Therefore, there are no stable policies. On the other hand, the deterministic stationary policy $\pi^{\varphi_{f^*}}$ given by $f^*(x) = -1$ still is an optimal policy.

Acknowledgements This work was partially sponsored by CONACYT grant SEP-2003-C02-45448/A-1 and PAICYT-UANL grant CA826-04. The authors are thankful to the referees for their valuable comments.

References

Altman E (1999) Constraint Markov decision processes. Chapman & Hall/CRC, Boca Raton

Ash RB (1972) Real analysis and probability. Academic Press, London

Beutler FJ, Ross KW (1985) Optimal policies for controlled Markov chains with a constraint. J Math Anal Appl 112:236–252

Billingsley P (1995) Probability and measure, 3rd edn. Wiley–Interscience, New York

Billingsley P (1999) Convergence of probability measures, 2nd edn. Wiley–Interscience, New York

Borkar VS (1994) Ergodic control of Markov chains with constraints—the general case. SIAM J Control Optim 32:176–186

Bourbaki N (1969) Éléments de mathématique, chapitre IX. Hermann, Paris

Collins EJ, McNamara JM (1998) Finite-horizon dynamic optimization when the terminal reward is a concave function of the distribution of the final state. Adv Appl Probab 30:122–136

Feimberg EA, Shwartz A (1996) Constrained discounted dynamic programming. Math Oper Res 21:922–945

Frid EB (1972) On optimal strategies in control problems with constraints. Theory Probab Appl 17:188–192

González-Hernández J, Hernández-Lerma O (2005) Extreme points of sets randomized strategies in constrained optimization and control problems. SIAM J Optim 15(4):1085–1104

Haviv M (1996) On constrained Markov decision processes. Oper Res Lett 19(1):25–28

Hernández-Lerma O, González-Hernández J (2000) Constrained Markov control processes in Borel spaces: the discounted case. Math Methods Oper Res 52:271–285

Hernández-Lerma O, Lasserre JB (1996) Discrete-time Markov control processes: basic optimality criteria. Springer, New York

Hernández-Lerma O, González-Hernández J, López-Martínez RR (2003) Constrained average cost Markov control processes in Borel space. SIAM J Control Optim 42(2):442–468

Hinderer K (1970) Foundations of non-stationary dynamic programming discrete-time parameter. Lecture notes oper res math syst, vol 33. Springer, Berlin
Hu Q, Yue W (2008) Markov decision processes with their applications. Springer, New York
Karr AF (1983) Extreme points of certain sets of probability measures with applications. Math Oper Res 8:74–85
Kurano M, Nakagami J, Huang Y (2000a) Constrained Markov decision processes with compact state and action spaces: the average case. Optimization 48(2):255–269
Kurano M, Yasuda M, Nakagami J, Yoshida Y (2000b) A fuzzy treatment of uncertain Markov decision processes: average case. In Proceedings of ASSM2000 International Conference on Applied Stochastic System Modeling, Kyoto, pp 148–157
Loève M (1977) Probability theory, vol 1, 4th edn. Springer, New York
Munkres JR (1975) Topology: a first course. Prentice–Hall, New Jersey
Phelps RR (1966) Lectures on Choquet's theorem. Van Nostrand, New York
Piunovskiy AB (1993) Control of random sequences in problems with constraints. Theory Probab Appl 38(4):751–762
Piunovskiy AB (1997) Optimal control of random sequences in problems with constraints. Kluwer Academic, Dordrecht
Piunovskiy AB, Khametov VM (1991) Optimal control by random sequences with constraints. Math Notes 49:654–656
Sennott LI (1991) Constrained discounted Markov decision chains. Probab Eng Inf Sci 5:463–475
Sennott LI (1993) Constrained average cost Markov decision chains. Probab Eng Inf Sci 7:69–83
Tanaka K (1991) On discounted dynamic programming with constraints. J Math Anal Appl 155:264–277
Winkler G (1988) Extreme points of moment sets. Math Oper Res 13:581–587
Yoshida K (1978) Functional analysis, 5th edn. Springer, Berlin
Yushkevich AA (1997) The compactness of a policy space in dynamic programming via an extension theorem for Carathéodory functions. Math Oper Res 22:458–467