Deterministic policies for Markov control processes with pathwise constraints *

Armando F. Mendoza-Pérez[1] and Onésimo Hernández-Lerma

August 30, 2010

Abstract

This paper deals with discrete-time Markov control processes in Borel spaces and unbounded rewards. Under suitable hypotheses, we show that a randomized stationary policy is optimal for a certain *expected constrained problem* (ECP) if and only if it is optimal for the corresponding *pathwise constrained problem* (pathwise CP). Moreover, we show that a parametric family of unconstrained optimality equations posseses compactness and convergence properties that lead to an *approximation scheme* which allows us to obtain constrained optimal policies as the limit of unconstrained deterministic optimal policies. In addition, we give sufficient conditions for the existence of *deterministic* policies that solve these constrained problems.

2000 Mathematics Subject Classification: 93E20, 90C40
Keywords and phrases: (discrete-time) Markov control processes, average reward criteria, pathwise average reward, constrained control problems

1 Introduction

This paper is about discrete-time Markov control processes (MCPs) in Borel spaces. Our problem is to maximize a pathwise long-run average reward subject to a constraint on a similar pathwise *cost*. To this end, we consider a corresponding *expected* average reward and average cost, and show that a stationary policy (either randomized or deterministic) is optimal for the *expected constrained problem* (ECP) if and only if it is optimal for the *pathwise constrained problem* (pathwise CP). Moreover, we show that a parametric family of unconstrained optimality equations posseses compactness and convergence properties that lead to an *approximation scheme* which allows us to obtain constrained optimal policies as the limit of unconstrained deterministic optimal policies. Furthermore, we give sufficient conditions for the existence of *deterministic* stationary policies that yield practical ways to solve our constrained problem. These results are clearly illustred with a linear system-quadratic reward/cost (also known as an LQ system).

Constrained MCPs form an important class of stochastic control problems with applications in many areas, including mathematical economics, signal processing, queueing systems, epidemic processes, etc.; see, for instance, [2, 3, 4, 5, 6, 7, 10, 11, 16, 19, 22, 23, 24] as well as the books [1] and [20] for MCPs with *expected* average rewards/costs and/or *countable* (possibly finite) state space.

*Research partially supported by CONACyT grant 104001.
[1]Corresponding author.

Among the few exceptions dealing with pathwise constraints we can mention the papers [22, 23, 10, 24] and our work [18].

In [18], we obtain the existence of optimal policies for a long-run pathwise (that is, sample-path) average reward subject to constraints on a long-run pathwise average cost. To do this, we give conditions for the existence of optimal policies for an average reward MCP with *expected constraints*, and then, these results are extended to the problem with *pathwise constraints*. The present paper is a sequel to [18].

We present here three main results. First, Theorem 4.3 proves that the ECP is "essentially" equivalent to the pathwise CP. Second, Theorem 4.8 gives several *characterizations* for a deterministic stationary policy to be optimal for the pathwise CP. Third, both Theorems 4.8 and 4.9 give *approximation schemes* to obtain randomized constrained optimal policies. To obtain these results we essentially follow the outline presented by Beutler and Ross [3] for finite-state finite-action MCPs. In short, we extend the results in [3] to MCPs with *uncountable* Borel spaces.

The remainder of the paper is organized as follows. In Section 2 we recall the basic components of a Markov control model, and state some of our main assumptions. Section 3 summarizes some facts on the *expected constrained problem* (ECP). In Section 4 we consider the *pathwise constrained problem* (pathwise CP) and introduce our main results, Theorems 4.3, 4.8, and 4.9. The proof of these results is presented in Section 5. Finally, an LQ system in Section 6 illustrates our results.

2 The control model

Let $(\mathbf{X}, \mathbf{A}, \{A(x) : x \in \mathbf{X}\}, Q, r, c)$ be a discrete time Markov control model with state space $\mathbf{X}$ and control (or action) set $\mathbf{A}$, both assumed to be separable metric spaces with Borel σ-algebras $\mathcal{B}(\mathbf{X})$ and $\mathcal{B}(\mathbf{A})$, respectively. For each $x \in \mathbf{X}$ there is a nonempty set $A(x)$ in $\mathcal{B}(\mathbf{A})$ which represents the set of feasible actions in the state x. The set

$$\mathbf{K} := \{(x, a) : x \in \mathbf{X}, a \in A(x)\} \tag{1}$$

is assumed to be a Borel subset of $\mathbf{X} \times \mathbf{A}$. The transition law Q is a stochastic kernel on $\mathbf{X}$ given $\mathbf{K}$. The one-stage reward r and the one-stage cost c are real-valued measurable functions on $\mathbf{K}$. We interpret r as a reward to be maximized with the restriction that the cost c does not exceed (in a suitably defined sense) a given value.

The class of measurable functions $f : \mathbf{X} \to \mathbf{A}$ such that $f(x)$ is in $A(x)$ for every $x \in \mathbf{X}$ is denoted by $\mathbf{F}$, and we suppose that it is nonempty. Let Φ be the set of stochastic kernels φ on $\mathbf{A}$ given $\mathbf{X}$ for which $\varphi(A(x)|x) = 1$ for all $x \in \mathbf{X}$.

Control policies. For every $n = 0, 1, \ldots$, let $\mathbf{H}_n$ be the family of admissible histories up to time n; that is, $\mathbf{H}_0 := \mathbf{X}$, and $\mathbf{H}_n := \mathbf{K}^n \times \mathbf{X}$ if $n \geq 1$. A *control policy* is a sequence $\pi = \{\pi_n\}$ of stochastic kernels π_n on $\mathbf{A}$ given $\mathbf{H}_n$ such that $\pi_n(A(x_n)|h_n) = 1$ for every n-history $h_n = (x_0, a_0, \cdots, x_{n-1}, a_{n-1}, x_n)$ in $\mathbf{H}_n$. The class of all policies is denoted by Π. Moreover, a policy $\pi = \{\pi_n\}$ is said to be a

(a) *randomized stationary policy* if there exists a stochastic kernel $\varphi \in \Phi$ such that $\pi_n(\cdot|h_n) = \varphi(\cdot|x_n)$ for all $h_n \in H_n$ and $n = 0, 1, \ldots$;

(b) *deterministic stationary policy* if there exists $f \in \mathbf{F}$ such that $\pi_n(\cdot|h_n)$ is the Dirac measure at $f(x_n) \in A(x_n)$ for all $h_n \in \mathbf{H}_n$ and $n = 0, 1, \ldots$.

Therefore, we have

$$\mathbf{F} \subset \Phi \subset \Pi.$$

Following a standard convention, we identify Φ with the class of randomized stationary policies and $\mathbf{F}$ with the class of deterministic stationary policies.

Given $\varphi \in \Phi$, we will use the following notation:

$$r_\varphi(x) := \int_{\mathbf{A}} r(x,a)\varphi(da|x), \quad c_\varphi(x) := \int_{\mathbf{A}} c(x,a)\varphi(da|x), \tag{2}$$

$$Q_\varphi(\cdot|x) := \int_{\mathbf{A}} Q(\cdot|x,a)\varphi(da|x) \tag{3}$$

for all $x \in \mathbf{X}$. Moreover, the n-step transition probabilities are denoted by Q^n_φ, with $Q^1_\varphi(\cdot|x) := Q_\varphi(\cdot|x)$ and $Q^0_\varphi(\cdot|x) := \delta_x$, the Dirac measure concentrated at the initial state x. We can write Q^n_φ recursively as

$$Q^n_\varphi(\cdot|x) = \int_{\mathbf{X}} Q_\varphi(\cdot|y) Q^{n-1}_\varphi(dy|x), \ n \geq 1. \tag{4}$$

In particular, for a deterministic policy $f \in \mathbf{F}$, (2)-(3) become

$$r_f(x) = c(x, f(x)), \quad c_f(x) = c(x, f(x)),$$

$$Q_f(\cdot|x) = Q(\cdot|x, f(x)).$$

Let $(\Omega, \mathcal{F})$ be the (canonical) measurable space consisting of the sample space $\Omega := (\mathbf{X} \times \mathbf{A})^\infty$ and its product σ-algebra $\mathcal{F}$. Then, for each policy $\pi \in \Pi$ and initial state $x \in \mathbf{X}$, a stochastic process $\{(x_n, a_n)\}$ and a probability measure P^π_x is defined on $(\Omega, \mathcal{F})$ in a canonical way, where x_n and a_n represent the state and control at time n, $n = 0, 1, \ldots$. The expectation operator with respect to P^π_x is denoted by E^π_x.

Given $\pi \in \Pi$, $x \in \mathbf{X}$, and $n = 1, 2, \ldots$, we define the n-stage pathwise reward and the n-stage expected reward as

$$S_n(\pi, x) := \sum_{k=0}^{n-1} r(x_k, a_k) \quad \text{and} \quad J_n(\pi, x) := E^\pi_x[S_n(\pi, x)],$$

respectively. Replacing the reward function r with the cost c we obtain the definition of $S_{c,n}(\pi, x)$ and $J_{c,n}(\pi, x)$.

Definition 2.1 *The (long-run) pathwise average reward and the (long-run) expected average reward are given by*

$$S(\pi, x) := \liminf_{n\to\infty} \frac{1}{n} S_n(\pi, x) \quad \text{and} \quad J(\pi, x) := \liminf_{n\to\infty} \frac{1}{n} J_n(\pi, x),$$

respectively. Similarly, the pathwise average cost and the expected average cost are respectively defined as

$$S_c(\pi, x) := \limsup_{n\to\infty} \frac{1}{n} S_{c,n}(\pi, x) \quad \text{and} \quad J_c(\pi, x) := \limsup_{n\to\infty} \frac{1}{n} J_{c,n}(\pi, x).$$

Observe that $J(\pi, x)$ (and $S(\pi, x)$) is defined as a "lim inf", whereas $J_c(\pi, x)$ (and $S_c(\pi, x)$) is a "lim sup". This is because according to standard conventions, the function r is interpreted as a *reward*-per-stage function, whereas the function c is a *cost*-per-stage.

We will introduce four sets of hypotheses. The first one, Assumption 2.2, consists of standard continuity-compactness conditions (see, for instance, [9, 13, 14, 21]) together with a growth condition (b1) on the one-step reward r and the cost c, and the Lyapunov-like condition (b3).

Assumption 2.2 *For every state $x \in \mathbf{X}$:*

(a) *$A(x)$ is a compact subset of $\mathbf{A}$;*

(b) *there exists a measurable function $W \geq 1$ on $\mathbf{X}$, a bounded measurable function $b \geq 0$, and nonnegative constants r_1, c_1, and β, with $\beta < 1$, such that*

(b1) $|r(x,a)| \leq r_1 W(x), \quad |c(x,a)| \leq c_1 W(x) \quad \forall (x,a) \in \mathbf{K}$;

(b2) $\int_{\mathbf{X}} W(y) Q(dy|x,a)$ *is continuous in $a \in A(x)$; and*

(b3) $\int_{\mathbf{X}} W(y) Q(dy|x,a) \leq \beta W(x) + b(x)$ *for every $x \in \mathbf{X}$.*

To state our second set of hypotheses, we will use the following notation, where W is the function in Assumption 2.2(b): $B_W(\mathbf{X})$ denotes the normed linear space of measurable functions u on $\mathbf{X}$ with a finite W-norm $\|u\|_W$, which is defined as

$$\|u\|_W := \sup_{x \in \mathbf{X}} |u(x)|/W(x). \tag{5}$$

In this case we say that u is W-bounded. Similarly, we say that a function $v : \mathbf{K} \to \mathbf{R}$ belongs to $B_W(\mathbf{K})$ if $x \mapsto \sup_{a \in A(x)} |v(x,a)|$ is in $B_W(\mathbf{X})$. In particular, by Assumption 2.2(b1), $r(x,a)$ and $c(x,a)$ are both in $B_W(\mathbf{K})$.

We write

$$\mu(u) := \int_{\mathbf{X}} u(y) \mu(dy),$$

whenever the integral is well defined.

The next set of hypotheses guarantees that the MCP has a nice "stable" behavior uniformly in Φ.

Assumption 2.3 *For each randomized stationary policy $\varphi \in \Phi$:*

(a) *(W-geometric ergodicity) There exists a (necessarily unique) probability measure μ_φ on $\mathbf{X}$ such that (with Q^t_φ as in (3)-(4))*

$$\left| \int_{\mathbf{X}} u(y) Q^t_\varphi(dy|x) - \mu_\varphi(u) \right| \leq \|u\|_W R \rho^t W(x), \tag{6}$$

for every $t = 0, 1, \ldots$, $u \in B_W(\mathbf{X})$, and $x \in \mathbf{X}$, where $R > 0$ and $0 < \rho < 1$ are constants independent of φ.

(b) *(Irreducibility) There exists a σ-finite measure ν on $\mathcal{B}(\mathbf{X})$ with respect to which Q_φ is ν-irreducible, which means that if $B \in \mathcal{B}(\mathbf{X})$ is such that $\nu(B) > 0$, then for every $x \in \mathbf{X}$ there exists $t > 0$ for which $Q^t_\varphi(B|x) > 0$.*

Remark 2.4 *For a discussion of Assumption 2.3, see Remark 2.4 in [18]. In particular, by Assumptions 2.3(a) and 2.2(b3), we have that*

$$\mu_\varphi(W) \le b/(1-\beta) < \infty \quad \forall \varphi \in \Phi, \tag{7}$$

with $b = \sup_{x\in\mathbf{X}} b(x)$. *Moreover, by (6),* $J(\varphi, x)$ *and* $J_c(\varphi, x)$ *in Definition 2.1 are constant (that is, do not depend on the initial state* x*), and verify that*

$$J(\varphi, x) = \lim_{n\to\infty} \frac{1}{n} E_x^\varphi \sum_{k=0}^{n-1} r(x_k, a_k) = \mu_\varphi(r_\varphi) =: g(\varphi),$$

where the letter g *is an abbreviation for "gain", which is another standard name for "average reward" [19], [21], and*

$$J_c(\varphi, x) = \lim_{n\to\infty} \frac{1}{n} E_x^\varphi \sum_{k=0}^{n-1} c(x_k, a_k) = \mu_\varphi(c_\varphi) =: g_c(\varphi).$$

In the following assumption we strengthen the growth condition on the reward function r and the cost function c in Assumption 2.2(b1).

Assumption 2.5 *There exist positive constants* r_2 *and* c_2 *such that*

$$r(x,a)^2 \le r_2 W(x) \quad \textit{and} \quad c(x,a)^2 \le c_2 W(x) \quad \forall (x,a) \in \mathbf{K}.$$

Note that, since $W \ge 1$, Assumption 2.5 implies Assumption 2.2(b1).

In the remainder of this paper we consider the function

$$w(x) := \sqrt{W(x)} \quad \forall x \in \mathbf{X}.$$

We also require the following assumption.

Assumption 2.6 (a) *The transition law* Q *is strongly continuous on* $\mathbf{K}$*, that is, the mapping*

$$(x,a) \mapsto \int_{\mathbf{X}} v(y) Q(dy|x,a)$$

is continuous on $\mathbf{K}$ *for every measurable bounded function* v *on* $\mathbf{X}$.

(b) *The cost function* c *is lower semicontinuous (l.s.c.) on* $\mathbf{K}$.

(c) *The reward function* r *is upper semicontinuous (u.s.c.) on* $\mathbf{K}$.

(d) *The function* w*, seen as a function* $(x,a) \mapsto w(x)$ *on* $\mathbf{K}$*, is continuous. Moreover,* w *is a so-called moment function on* $\mathbf{K}$*, that is, there exists a nondecreasing sequence of compact sets* $K_n \uparrow \mathbf{K}$ *such that*

$$\lim_{n\to\infty} \inf\{w(x) : (x,a) \notin K_n\} = \infty.$$

Remark 2.7 *In Assumption 2.6(b), we omit the restrictive condition on the cost fuction c imposed in [18, Assumption 3.3(b)], which establishes that c is nonnegative. Nonnegativity of c was crucial to prove that the set* $\Gamma(\theta)$ *in (26) below is compact in the w-weak topology (see, for instance, [18, Section 5] and [17, Lemma 5.2.2]). Here, if we assume the l.s.c. of c in addition to Assumptions 2.5 and 2.6(d) above, we can get the same results obtained in [18]. A moment function, such as w in Assumption 2.6(d), is also known as a Lyapunov (or Lyapunov-like) function.*

3 MCPs with expected constraints

In this section we summarize some facts from [17, 18] on MCPs with *expected* constraints. These results are used in Section 4 to state our main results.

By the Assumption 2.6(b) and the Remark 2.4, we can define

$$\theta_{min} := \min_{\varphi\in\Phi} \int_{\mathbf{X}} c_\varphi(y)\mu_\varphi(dy) \quad \text{and} \quad \theta_{max} := \sup_{\varphi\in\Phi} \int_{\mathbf{X}} c_\varphi(y)\mu_\varphi(dy), \tag{8}$$

which are finite numbers. To avoid trivial situations, we will consider a constraint constant θ such that

$$\theta_{min} < \theta < \theta_{max}. \tag{9}$$

Let $J(\pi, x)$ and $J_c(\pi, x)$ be the long-run expected averages in Definition 2.1, and let θ be a constant as in (9). Then the *expected constrained problem* (ECP) is:

$$\text{maximize } J(\pi, x) \tag{10}$$

$$\text{subject to:} \quad \pi \in \Pi \quad \text{and} \quad J_c(\pi, x) \le \theta \quad \forall x \in \mathbf{X}. \tag{11}$$

Definition 3.1 *A policy* $\pi \in \Pi$ *is said to be* feasible *for the ECP if it satisfies the constraints in (11), that is,* $J_c(\pi, x) \le \theta$ *for all x in* $\mathbf{X}$*. Moreover, a feasible policy* π^* *is called* optimal *for the ECP (10)-(11) if* $J(\pi, x) \le J(\pi^*, x)$ *for every feasible* π*.*

The following proposition states the existence of an optimal policy for the ECP (10)-(11). Furthermore, it establishes the existence of a solution to the *average reward optimality equation* (AROE) (12) below. For a proof of the proposition see [18, Theorem 5.2] or [17, Theorem 5.3.1].

Proposition 3.2 *Suppose that Assumptions 2.2, 2.3, 2.5, and 2.6 are satisfied. Then:*

(i) *There exists a number* $\Lambda_0 \le 0$*, a constant* $V(\theta)$ *which depends on* θ*, and* $h \in B_w(\mathbf{X})$ *such that the AROE*

$$V(\theta) + h(x) = \max_{a\in A(x)} \left[r(x,a) + \Big(c(x,a) - \theta\Big) \cdot \Lambda_0 + \int_{\mathbf{X}} h(y) Q(dy|x,a) \right] \tag{12}$$

holds for every $x \in \mathbf{X}$*.*

(ii) *There exists a randomized stationary policy $\varphi^* \in \Phi$ that attains the maximum in the right-side of (12), i.e.,*

$$V(\theta) + h(x) = r_{\varphi^*}(x) + \Big(c_{\varphi^*}(x) - \theta\Big) \cdot \Lambda_0 + \int_{\mathbf{X}} h(y) Q_{\varphi^*}(dy|x) \tag{13}$$

for all $x \in \mathbf{X}$. Moreover, φ^* is optimal for the ECP (10)-(11). In addition, the following "orthogonality" property (using the notation in the Remark 2.4) is satisfied

$$\Big(g_c(\varphi^*) - \theta\Big) \cdot \Lambda_0 = 0, \tag{14}$$

which together with (13) gives

$$V(\theta) = \mu_{\varphi^*}(r_{\varphi^*}) = g(\varphi^*), \tag{15}$$

that is, $V(\theta)$ is the optimal value for the ECP (10)-(11).

An optimal policy $\varphi^* \in \Phi$ for the ECP satisfying the AROE (13) is called *a canonical policy for the ECP*.

Proposition 3.2 shows that the ECP (10)-(11) induces an unconstrained problem depending on a real number $\Lambda_0 \leq 0$, which is unknown. The next result states that the ECP can be solved by means of a parametric family of AROEs (see, for instance, [18, Theorem 5.3]) or [17, Theorem 5.4.1]).

Proposition 3.3 *Suppose that the hypotheses of Proposition 3.2 are satisfied, and consider the ECP (10)-(11). For each real number $\Lambda \leq 0$, let $(\rho(\Lambda), h_\Lambda) \in \mathbf{R} \times B_W(\mathbf{X})$ be a solution to the AROE*

$$\rho(\Lambda) + h_\Lambda(x) = \max_{a \in A(x)} \Big[r(x,a) + (c(x,a) - \theta) \cdot \Lambda + \int_{\mathbf{X}} h_\Lambda(y) Q(dy|x,a) \Big] \tag{16}$$

for every $x \in \mathbf{X}$. Then

$$V(\theta) = \min_{\Lambda \leq 0} \rho(\Lambda). \tag{17}$$

4 MCPs with pathwise constraints: main results

Let $\theta \in \mathbf{R}$ be as in (9). With the notation in Definition 2.1 we want to maximize the pathwise average reward $S(\pi, x)$ over the set of all policies $\pi \in \Pi$ satisfying, for every initial state $x \in \mathbf{X}$, the following constraint on the pathwise average cost

$$S_c(\pi, x) \leq \theta \quad P_x^\pi - a.s.$$

Hence, we can explicitly state our pathwise CP as follows:

$$\text{maximize } S(\pi, x) \tag{18}$$

(19) $$\text{subject to:} \quad \pi \in \Pi \quad \text{and} \quad S_c(\pi, x) \leq \theta \;\; P_x^\pi - \text{a.s. } \forall x \in \mathbf{X}.$$

A policy $\pi \in \Pi$ is said to be *feasible* for the pathwise CP if it satisfies (19).

Let $\varphi \in \Phi$ be an arbitrary randomized *stationary* policy, and let $g(\varphi)$ and $g_c(\varphi)$ be as in Remark 2.4. Using the strong law of large numbers for Markov chains it can be shown that, for every $x \in \mathbf{X}$,

$$S(\varphi, x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} r_\varphi(x_k) = g(\varphi) \quad \text{and} \quad S_c(\varphi, x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} c_\varphi(x_k) = g_c(\varphi)$$

$P_x^\varphi - a.s.$ This fact is used in the following definition.

Definition 4.1 *Let $\varphi^* \in \Phi$ be a feasible policy for the pathwise CP, i.e., $g_c(\varphi^*) \leq \theta$. Then φ^* is said to be optimal for the pathwise CP (18)-(19) if for each feasible $\pi \in \Pi$ we have*

$$S(\pi, x) \leq g(\varphi^*) \quad P_x^\pi - a.s.$$

If φ^* is an optimal policy for the problem (18)-(19), then we define the optimal value of the pathwise CP as $V^*(\theta) := g(\varphi^*)$.

The following result establishes the existence of optimal policies for the pathwise CP (18)-(19) (see, for instance, [18, Theorem 3.4]) or [17, Theorem 5.5.2]).

Proposition 4.2 *Suppose that Assumptions 2.2, 2.3, 2.5, and 2.6 hold. Then:*

(i) *There exists an optimal policy $\varphi^* \in \Phi$ for the pathwise CP (18)-(19). In particular, $g_c(\varphi^*) \leq \theta$ and $g(\varphi^*) = V^*(\theta)$, with $V^*(\theta)$ as in Definition 4.1.*

(ii) *There exist $\Lambda_0 \leq 0$ and $h \in B_w(\mathbf{X})$ such that the average reward optimality equation (AROE)*

$$V^*(\theta) + h(x) = \max_{a\in A(x)} \Big[r(x,a) + \Big(c(x,a) - \theta \Big) \cdot \Lambda_0 + \int_{\mathbf{X}} h(y) Q(dy|x,a) \Big]$$

$$= r_{\varphi^*}(x) + \Big(c_{\varphi^*}(x) - \theta \Big) \cdot \Lambda_0 + \int_{\mathbf{X}} h(y) Q_{\varphi^*}(dy|x) \tag{20}$$

holds for every $x \in \mathbf{X}$. Furthermore, we have the "orthogonality" property

$$(g_c(\varphi^*) - \theta) \cdot \Lambda_0 = 0. \tag{21}$$

(iii) *For each $\Lambda \leq 0$, let $(\rho(\Lambda), h_\Lambda) \in \mathbf{R} \times B_W(\mathbf{X})$ be a solution to the AROE*

$$\rho(\Lambda) + h_\Lambda(x) = \max_{a\in A(x)} \Big[r(x,a) + (c(x,a) - \theta) \cdot \Lambda + \int_{\mathbf{X}} h_\Lambda(y) Q(dy|x,a) \Big]$$

for every $x \in \mathbf{X}$. Then $V^*(\theta) = V(\theta) = \min_{\Lambda\leq 0} \rho(\Lambda)$, with $V(\theta)$ as in Proposition 3.2(i) and Proposition 3.3.

We can now state our first main result, which is proved in Section 5. In this result we establish that a (randomized) stationary policy is optimal for the pathwise CP (18)-(19) if and only if it is optimal for the ECP (10)-(11), i.e., the pathwise CP is, under our assumptions, "essentially" equivalent to the ECP.

Notation. Let $\Phi_{ecp} \subset \Phi$ be the class of randomized stationary optimal policies for the ECP (10)-(11), and Φ_{cecp} the subclass of Φ_{ecp} of canonical policies for the ECP.

Moreover, let $\mathbf{F}_{ecp} \subset \Phi_{ecp}$ be the class of deterministic stationary optimal policies for the ECP, and $\mathbf{F}_{cecp} \subset \mathbf{F}$ the subclass of Φ_{cecp} of deterministic stationary canonical policies for the ECP.

Theorem 4.3 *Suppose that Assumptions 2.2, 2.3, 2.5, and 2.6 are satisfied.*

(a) *Let $V(\theta)$ be as in Proposition 3.2. Then, for each feasible policy $\pi \in \Pi$ for the pathwise CP (18)-(19), and for each initial state $x \in \mathbf{X}$*

$$V(\theta) \geq S(\pi, x) \quad P_x^\pi - a.s. \tag{22}$$

Moreover, $V(\theta)$ is the optimal value for the pathwise CP (18)-(19), i.e., $V(\theta) = V^*(\theta)$. Furthermore, if $\widehat{\varphi} \in \Phi_{ecp}$ is an optimal policy for the ECP (10)-(11), then it is an optimal policy for the pathwise CP (18)-(19).

(b) *Conversely, let $\widehat{\varphi} \in \Phi$ be an optimal policy for the pathwise CP (18)-(19). Then $\widehat{\varphi}$ is an optimal policy for the ECP (10)-(11) satisfying*

$$[g_c(\widehat{\varphi}) - \theta] \cdot \Lambda_0 = 0. \tag{23}$$

In addition, there exists an optimal policy $\varphi^* \in \Phi_{cecp}$ for the ECP (10)-(11) satisfying Proposition 3.2(ii) and such that

$$\widehat{\varphi}(\cdot|x) = \varphi^*(\cdot|x) \quad \mu_{\widehat{\varphi}} - a.s.,$$

and so $\mu_{\widehat{\varphi}} = \mu_{\varphi^*}$.

(c) *Suppose that there exists a deterministic stationary optimal policy $\widehat{f} \in \mathbf{F}$ for the pathwise CP (18)-(19). Then $\widehat{f}$ is an optimal policy for the ECP (10)-(11) satisfying*

$$[g_c(\widehat{f}) - \theta] \cdot \Lambda_0 = 0. \tag{24}$$

Furthermore, there exists a deterministic stationary optimal policy $f^* \in \mathbf{F}_{cecp}$ for the ECP (10)-(11) satisfying Proposition 3.2(ii) and such that

$$\widehat{f}(x) = f^*(x) \quad \mu_{\widehat{f}} - a.s.,$$

and so $\mu_{\widehat{f}} = \mu_{f^*}$.

Remark 4.4 *Denoting by $\Phi_{scp} \subset \Phi$ the class of randomized stationary optimal policies for the pathwise (sample-path) CP (18)-(19), we may rewrite the statements in Theorem 4.3(a), (b) as*

$$\Phi_{ecp} = \Phi_{scp}.$$

Similarly, if we denote by $\mathbf{F}_{scp} \subset \Phi_{scp}$ the subclass of deterministic stationary optimal policies for the pathwise CP (18)-(19), then

$$\mathbf{F}_{ecp} = \mathbf{F}_{scp}.$$

Theorems 4.8 and 4.9 below give conditions to guarantee that $\mathbf{F}_{ecp}$ is a nonempty set.

Finally, thanks to Theorem 4.3, we can identify the ECP and the pathwise CP. Hence, we will refer to these equivalent problems as *the constrained problem* (CP).

To state our second main result, we will use the following notation.

Let W be as in Assumption 2.2, $w := \sqrt{W}$, and $\mathcal{B}(\mathbf{K})$ the Borel σ-algebra on $\mathbf{K}$; see (1). We denote by $\mathcal{P}_w(\mathbf{K})$ the set of probability measures $\widehat{\mu}$ on $\mathcal{B}(\mathbf{K})$ such that

$$\int_{\mathbf{K}} w(x)\widehat{\mu}(d(x,a)) < \infty.$$

This set is supposed to be endowed with the w-weak topology [8, Appendix A.5], i.e., the smallest topology for which the mapping

$$\widehat{\mu} \mapsto \int_{\mathbf{K}} v \; d\widehat{\mu}$$

on $\mathcal{P}_w(\mathbf{K})$ is continuous, for every $v \in C_w(\mathbf{K})$, where $C_w(\mathbf{K})$ is the subspace of continuous-functions in $B_w(\mathbf{K})$. With this topology $\mathcal{P}_w(\mathbf{K})$ is separable and metrizable.

For every $\varphi \in \Phi$, let μ_φ be as in Assumption 2.3(a), and define $\hat{\mu}_\varphi \in \mathcal{P}_w(\mathbf{K})$ as

$$\hat{\mu}_\varphi(B \times C) := \int_B \varphi(C|x)\mu_\varphi(dx) \quad \forall B \in \mathcal{B}(\mathbf{X}), C \in \mathcal{B}(\mathbf{A}).$$

The set of all of these measures is denoted by Γ, i.e.,

$$\Gamma := \{\hat{\mu}_\varphi : \varphi \in \Phi\} \subset \mathcal{P}_w(\mathbf{K}) \tag{25}$$

Moreover, for each $\theta \in (\theta_{min}, \theta_{max})$, with θ_{min} and θ_{max} as in (9), let

$$\Gamma(\theta) := \{\hat{\mu} \in \Gamma : \int_{\mathbf{K}} c \; d\hat{\mu} \leq \theta\}. \tag{26}$$

It can be verified that Γ and $\Gamma(\theta)$ are both convex sets. Furthermore, after some calculations (see [17, Lemma 5.2.2] for details) and using Prohorov's theorem [8, Appendix A.5] it follows that Γ and $\Gamma(\theta)$ are both compact sets in the w-weak topology.

For each $\Lambda \leq 0$ let $r_\Lambda(x,a) := r(x,a) + (c(x,a) - \theta) \cdot \Lambda$. Then, given a stationary policy $\varphi \in \Phi$, define

$$G_\Lambda(\varphi) := \widehat{\mu}_\varphi(r_\Lambda). \tag{27}$$

On the other hand, by our continuity and compactness conditions in Assumptions 2.2 and 2.6, well-known measurable selection theorems (see [12, Appendix D], for instance) give the existence of a stationary deterministic policy $f_\Lambda \in \mathbf{F}$ (not necessarily unique) such that for every $x \in \mathbf{X}$, the action $f_\Lambda(x) \in A(x)$ attains the maximum in the right-hand side of (16). By the Axiom of Choice, for each $\Lambda \leq 0$, we take one from those f_Λ.

Remark 4.5 *By standard dynamic programming results (see, for instance, [13, Section 10.3]), the function $\Lambda \mapsto \rho(\Lambda) = G_\Lambda(f_\Lambda)$ is well defined and it does not depend on the particular choice of f_Λ. Furthermore, Let $\varphi \in \Phi$ be arbitrary, then (16) implies*

$$\rho(\Lambda) + h_\Lambda(x) \geq r_\varphi(x) + (c_\varphi(x) - \theta) \cdot \Lambda + \int_{\mathbf{X}} h_\Lambda(y) Q_\varphi(dy|x)$$

for all $x \in \mathbf{X}$. Integrating both sides of this inequality with respect to μ_φ, we have

$$\rho(\Lambda) \geq G_\Lambda(\varphi) \quad \forall \varphi \in \Phi. \tag{28}$$

Next, we introduce

$$\gamma := \sup\{\Lambda \leq 0 : g_c(f_\Lambda) \leq \theta\}. \tag{29}$$

According to Lemma 5.2 below, γ defined in (29) is finite. Notice that $-\infty < \gamma \leq 0$.

Proposition 3.2 establishes the existence of an optimal policy for our CP. Our purpose now is to use the parametric family of unconstrained optimization problems (16) to obtain this optimal policy as a function of the parameter Λ (see Theorem 4.8 and Theorem 4.9 below).

We state the following assumptions.

Assumption 4.6 *The cost function c is continuous on $\mathbf{K}$.*

Assumption 4.7 *Let γ be defined in (29). We assume*

$$-\infty < \gamma < 0.$$

Theorem 4.8 *Suppose that Assumptions 2.2, 2.3, 2.5, and 2.6 are satisfied.*

(a) *Suppose that there exists $\Lambda \leq 0$ and $\widehat{\varphi} \in \Phi$ satisfying*

$$g_c(\widehat{\varphi}) = \theta \quad \text{and} \quad G_\Lambda(\widehat{\varphi}) = \rho(\Lambda). \tag{30}$$

Then $\widehat{\varphi}$ is an optimal policy for the CP. Hence,

$$\rho(\Lambda) = \min_{\lambda \leq 0} \rho(\lambda) = V(\theta). \tag{31}$$

Moreover, if $g_c(f_\Lambda) = \theta$ (with f_Λ as in Remark 4.5), then f_Λ solves the CP.

(b) *Assume that $\Lambda \mapsto \rho(\Lambda)$ is differentiable at some point $\Lambda < 0$. Then*

$$\frac{d\rho}{d\Lambda}(\Lambda) = g_c(f_\Lambda) - \theta. \tag{32}$$

In particular, if $\Lambda < 0$ is a critical point of $\rho(\cdot)$, then f_Λ is an optimal policy for the CP, and $\rho(\cdot)$ attains a minimum in Λ satisfying (31).

(c) *Let us suppose that there exists $\Lambda < 0$ such that $\rho(\cdot)$ is differentiable at Λ. Then the following statements are equivalent:*

1) *f_Λ solves the CP;*

2) *Λ is a critical point of $\rho(\cdot)$;*

3) *$g_c(f_\Lambda) = \theta$.*

(d) *In addition, assume that the mapping $\Lambda \mapsto g_c(f_\Lambda)$ is continuous on the interval $(-\infty, 0)$. Then the function $\rho(\cdot)$ is differentiable on the interval $(-\infty, 0)$.*

Recall the definition (29) of γ, which is used again in the following theorem.

Theorem 4.9 *Suppose that Assumptions 2.2, 2.3, 2.5, 2.6, 4.6, and 4.7 hold. Then there exist two sequences of negative numbers $\{\Lambda_n\}$, $\{\Lambda_\nu\}$ such that $\Lambda_n \uparrow \gamma$, and $\Lambda_\nu \downarrow \gamma$ satisfying:*

(i) *The corresponding sequences of measures $\{\widehat{\mu}_{f_{\Lambda_n}}\}$ and $\{\widehat{\mu}_{f_{\Lambda_\nu}}\}$ converge on $\mathcal{P}_w(\mathbf{K})$, with respect to the w-weak topology, toward measures $\widehat{\mu}_{\varphi_1}$ and $\widehat{\mu}_{\varphi_2}$ in Γ, with $\varphi_1, \varphi_2 \in \Phi$ such that*

$$g_c(\varphi_1) \leq \theta \quad \textit{and} \quad g_c(\varphi_2) \geq \theta, \tag{33}$$

and

$$G_\gamma(\varphi_1) = G_\gamma(\varphi_2) = \rho(\gamma). \tag{34}$$

(ii) *There exist a randomized stationary policy $\varphi^* \in \Phi$, and a number $q_0 \in [0, 1]$ such that*

$$\widehat{\mu}_{\varphi^*} = q_0 \widehat{\mu}_{\varphi_1} + (1 - q_0)\widehat{\mu}_{\varphi_2} \quad \textit{and} \quad g_c(\varphi^*) = \theta.$$

Hence, the policy $\varphi^* \in \Phi$ is optimal for the CP. Moreover, the function $\Lambda \mapsto \rho(\Lambda)$ attain a minimum in γ, i.e.,

$$\rho(\gamma) = \min_{\Lambda \leq 0} \rho(\Lambda) = V(\theta).$$

(iii) *In addition, suppose that $\rho(\cdot)$ is differentiable at γ. Then f_γ solves the CP. In particular, γ is a critical point of $\rho(\cdot)$, and $g_c(f_\gamma) = \theta$. In this case, we can identify Λ_0 in Proposition 3.2 with $\gamma < 0$.*

(iv) *If we suppose that Assumption 4.7 does not hold, then $\varphi_1 \in \Phi$ satisfying (33) and (34) for $\gamma = 0$, is an optimal policy for the CP. In particular, if $g_c(f_0) \leq \theta$, then f_0 is an optimal policy for the CP.*

5 Proof of Theorems 4.3, 4.8, 4.9

Suppose that Assumptions 2.2, 2.3, 2.5, and 2.6 hold throughout this section.

Proof of Theorem 4.3. *Proof of (a).* The inequality in (22) follows from the proof of Theorem 3.4(i) in [18].

Now, suppose that $\widehat{\varphi}$ is an optimal policy for the ECP (10)-(11). By (15) and Remark 2.4, we have

$$g(\widehat{\varphi}) = V(\theta) \quad \text{and} \quad g_c(\widehat{\varphi}) \leq \theta. \tag{35}$$

From Remark 2.4(iv) in [18], together with (22) and (35), we have that $\widehat{\varphi}$ is optimal for the pathwise CP (18)-(19), and $V(\theta)$ is the optimal value, that is, $V(\theta) = V^*(\theta)$.

Proof of (b). Let $\widehat{\varphi}$ be an optimal policy for the pathwise CP (18)-(19). By part (a) of this theorem, $V(\theta)$ is the optimal value for the pathwise CP. Thus

$$g(\widehat{\varphi}) = V(\theta) \quad \text{and} \quad g_c(\widehat{\varphi}) \leq \theta. \tag{36}$$

Furthermore, by Remark 2.4 again,

$$J(\widehat{\varphi}, x) = V(\theta) \quad \text{and} \quad J_c(\widehat{\varphi}, x) \leq \theta \quad \forall x \in \mathbf{X}.$$

So, $\widehat{\varphi}$ is also an optimal policy for the ECP (10)-(11). Hence, the rest of the proof of part (b) is the same as the proof of Theorem 5.2(ii) in [18].

Finally, the proof of part (c) is very similar as the proof of (b) above. ■

To prove Theorems 4.8 and 4.9, we need the following lemmas.

Lemma 5.1 *For each $\Lambda \leq 0$ and every real number η such that $\Lambda + \eta \leq 0$, the following inequalities hold*

$$\begin{aligned} \eta \cdot [g_c(f_\Lambda) - \theta] &= G_{\Lambda+\eta}(f_\Lambda) - \rho(\Lambda) \\ &\leq \rho(\Lambda + \eta) - \rho(\Lambda) \\ &\leq \rho(\Lambda + \eta) - G_\Lambda(f_{\Lambda+\eta}) \\ &= \eta \cdot [g_c(f_{\Lambda+\eta}) - \theta]. \end{aligned} \tag{37}$$

As a consequence of these inequalities we have the following facts:

(i) *$g_c(f_\Lambda)$ and $g(f_\Lambda)$ are monotone nondecreasing functions in the parameter Λ.*

(ii) *If $g_c(f_\Lambda) \leq \theta$, then $\rho(\cdot)$ is monotone nonincreasing on the interval $(-\infty, \Lambda]$. If $g_c(f_\Lambda) \geq \theta$, then $\rho(\cdot)$ is monotone nondecreasing on the interval $[\Lambda, 0]$.*

(iii) *$\rho(\cdot)$ is continuous in $\Lambda \leq 0$.*

Proof. Consider $\Lambda \leq 0$ and a real number η such that $\Lambda + \eta \leq 0$. By the AROE (16), with $\Lambda + \eta$ in lieu of Λ, we obtain

$$\rho(\Lambda+\eta) + h_{\Lambda+\eta}(x) \geq r_\Lambda(x,a) + (c(x,a) - \theta)\cdot\eta + \int_X h_{\Lambda+\eta}(y)Q(dy|x,a)$$

for all $(x,a) \in \mathbf{K}$. Hence,

$$\rho(\Lambda+\eta) + h_{\Lambda+\eta}(x) \geq r_\Lambda(x,f_\Lambda(x)) + (c(x,f_\Lambda(x)) - \theta)\cdot\eta + \int_X h_{\Lambda+\eta}(y)Q(dy|x,f_\Lambda(x))$$

for all $x \in \mathbf{X}$. Integrating both sides of this inequality with respect to the measure μ_{f_Λ}, we obtain

$$\rho(\Lambda+\eta) \geq \rho(\Lambda) + (g_c(f_\Lambda) - \theta)\cdot\eta = G_{\Lambda+\eta}(f_\Lambda). \tag{38}$$

Now, from (28) in Remark 4.5, we have

$$\rho(\Lambda) \geq G_\Lambda(f_{\Lambda+\eta}). \tag{39}$$

Moreover

$$\rho(\Lambda+\eta) - G_\Lambda(f_{\Lambda+\eta}) = G_{\Lambda+\eta}(f_{\Lambda+\eta}) - G_\Lambda(f_{\Lambda+\eta}) = (g_c(f_{\Lambda+\eta}) - \theta)\cdot\eta. \tag{40}$$

Combining (38), (39) and (40), we obtain the inequalities in (37).

Proof of (i)-(iii). From (37), we have that $g_c(f_\Lambda)$ is nondecreasing in the parameter Λ.

On the other hand, from the first inequality in (37), we have that if $g_c(f_\Lambda) \leq \theta$ and $\eta < 0$, then $0 \leq \eta\cdot[g_c(f_\Lambda) - \theta] \leq \rho(\Lambda+\eta) - \rho(\Lambda)$. This implies that $\rho(\cdot)$ is nonincreasing on $(-\infty, \Lambda)$. Similarly, if $g_c(f_\Lambda) \geq \theta$ and $\eta > 0$, by the same inequality we have that $\rho(\Lambda) \leq \rho(\Lambda+\eta)$ with $\eta > 0$, i.e., $\rho(\cdot)$ is nondecreasing on $[\Lambda, 0]$. Thus, we have proved the assertions in (ii).

Next, we prove that $g(f_\Lambda)$ is nondecreasing. Arguing by contradiction, suppose that $g(f_\Lambda)$ is not monotone nondecreasing. Hence, there exist $\Lambda \leq 0$ and $\eta < 0$ such that $g(f_\Lambda) < g(f_{\Lambda+\eta})$. By the first part of (i), $g_c(f_\Lambda)$ is nondecreasing. So, $g_c(f_{\Lambda+\eta}) \leq g_c(f_\Lambda)$. Thus, we have the contradiction (see (39) above)

$$\rho(\Lambda) = g(f_\Lambda) + (g_c(f_\Lambda) - \theta)\cdot\Lambda < g(f_{\Lambda+\eta}) + (g_c(f_{\Lambda+\eta}) - \theta)\cdot\Lambda = G_\Lambda(f_{\Lambda+\eta}).$$

Finally, the statement in (iii) is a direct consequence of (37). ■

The following lemma proves that γ defined in (29) is a finite number.

Lemma 5.2 *There exists $\Lambda \leq 0$ such that $g_c(f_\Lambda) \leq \theta$. Moreover, we have*

(a) *γ is a finite number such that $-\infty < \gamma \leq 0$.*

(b) *Consider $\Lambda < 0$. If $\Lambda < \gamma$, then $g_c(f_\Lambda) \leq \theta$. If $\Lambda > \gamma$, then $g_c(f_\Lambda) > \theta$.*

(c) $\rho(\gamma) = \min_{\Lambda\leq 0}\rho(\Lambda) = V(\theta)$.

Proof. By contradiction, assume that $g_c(f_\Lambda) > \theta$ for all $\Lambda \leq 0$. From Lemma 5.1(i), $g(f_\Lambda)$ is nondecreasing. Thus

$$\rho(\Lambda) = g(f_\Lambda) + (g_c(f_\Lambda) - \theta) \cdot \Lambda \leq \rho(0) \quad \forall \Lambda \leq 0. \tag{41}$$

On the other hand, from the definition of θ_{min} in (8) and θ in (9), there exists $\varphi \in \Phi$ such that $g_c(\varphi) < \theta$. Defining $\delta := \theta - g_c(\varphi) > 0$, we have that

$$G_\Lambda(\varphi) = g(\varphi) + (g_c(\varphi) - \theta) \cdot \Lambda = g(\varphi) - \delta \cdot \Lambda \quad \forall \Lambda \leq 0.$$

Hence, $\lim_{\Lambda \to -\infty} G_\Lambda(\varphi) = \infty$. This limit and (28) in Remark 4.5 imply the existence of $\Lambda < 0$ such that

$$\rho(0) < G_\Lambda(\varphi) \leq \rho(\Lambda),$$

which contradicts (41).

Proof of (a). From the first part of this proof and the definition of γ in (29), we have that $-\infty < \gamma \leq 0$.

Proof of (b). This part follows from the definition of γ in (29), and the fact that $g_c(f_\Lambda)$ is nondecreasing in the parameter $\Lambda \leq 0$ (see Lemma 5.1(i)).

Proof of (c). From part (b) of this lemma, and Lemma 5.1(ii)-(iii), we have that

$$\rho(\Lambda) \geq \rho(\gamma) \quad \forall \Lambda < \gamma,$$

and

$$\rho(\Lambda) \geq \rho(\gamma) \quad \forall \Lambda > \gamma.$$

These inequalities imply that $\rho(\gamma) = \min_{\Lambda \leq 0} \rho(\Lambda)$. Furthermore, from Proposition 3.3, $V(\theta) = \rho(\gamma)$. ■

Proof of Theorem 4.8. *Proof of (a).* Let $\Lambda \leq 0$ and $\widehat{\varphi} \in \Phi$ satisfy (30). In particular, $\widehat{\varphi}$ is a feasible policy for the ECP (10)-(11), and by (28) it follows that

$$g(\widehat{\varphi}) = G_\Lambda(\widehat{\varphi}) = \rho(\Lambda) \geq G_\Lambda(\varphi) \quad \forall \varphi \in \Phi.$$

Since $G_\Lambda(\varphi) \geq g(\varphi)$ for each feasible policy $\varphi \in \Phi$ for the CP (10)-(11), Proposition 3.2 and the latter inequality imply that $\widehat{\varphi}$ is an optimal policy for the CP. Now, from (17) in Proposition 3.3, $V(\theta) = \rho(\Lambda) = \min_{\lambda \leq 0} \rho(\lambda)$.

In particular, if $g_c(f_\Lambda) = \theta$, since $\rho(\Lambda) = G_\Lambda(f_\Lambda)$, then f_Λ is an optimal policy for the CP.

Proof of (b). Assuming that $\rho(\cdot)$ is differentiable at $\Lambda < 0$, from the first inequality in (37) we obtain, for each $\eta > 0$,

$$g_c(f_\Lambda) - \theta \leq \frac{\rho(\Lambda + \eta) - \rho(\Lambda)}{\eta},$$

and

$$g_c(f_\Lambda) - \theta \geq \frac{\rho(\Lambda - \eta) - \rho(\Lambda)}{-\eta}.$$

Taking the limit as $\eta \to 0$, we obtain (32).

On the other hand, if $\Lambda < 0$ is a critical point of $\rho(\cdot)$, then, from (32), we have that $g_c(f_\Lambda) = \theta$. Hence, from part (a) above, f_Λ solves the CP.

Proof of (c). This part is a direct consequence of parts (a) and (b) of this theorem.

Proof of (d). Suppose that the function $\Lambda \mapsto g_c(f_\Lambda)$ is continuous on the interval $(-\infty, 0)$. From (37) in Lemma 5.1, we obtain that the continuous function $\Lambda \mapsto \rho(\Lambda)$ is differentiable with continuous derivative

$$\frac{d\rho}{d\Lambda}(\Lambda) = g_c(f_\Lambda) - \theta, \quad \forall \Lambda < 0. \qquad \blacksquare$$

Proof of Theorem 4.9. *Proof of (i).* From Assumption 4.7, we can consider two sequences of negative numbers $\{\Lambda_n\}$ and $\{\Lambda_\nu\}$ satisfying $\Lambda_n \uparrow \gamma$ and $\Lambda_\nu \downarrow \gamma$. Now, since Γ is a compact (separable) metric space with respect to the w-weak topology [8, Appendix 5], each sequence in Γ has a subsequence which converges in Γ. Thus, we can assume that the sequences $\{\widehat{\mu}_{f_{\Lambda_n}}\}$ and $\{\widehat{\mu}_{f_{\Lambda_\nu}}\}$ converge in $\mathcal{P}_w(\mathbf{K})$ with respect to the w-weak topology, to some measures $\widehat{\mu}_{\varphi_1}$ and $\widehat{\mu}_{\varphi_2}$ in Γ, with $\varphi_1, \varphi_2 \in \Phi$.

From Lemma 5.2(b), we have that $g_c(f_{\Lambda_n}) \leq \theta$ for all n, and $g_c(f_{\Lambda_\nu}) > \theta$ for all ν. By Assumption 4.6, the cost function c is continuous on $\mathbf{K}$, and so $g_c(\varphi_1) = \lim_{n\to\infty} g_c(f_{\Lambda_n}) \leq \theta$ and $g_c(\varphi_2) = \lim_{\nu\to\infty} g_c(f_{\Lambda_\nu}) \geq \theta$, obtaining (33).

Next we prove (34). From the upper semicontinuity of r (see Assumption 2.6(c)), and the continuity of c, we have that $r_\gamma := r + (c - \theta) \cdot \gamma$ is upper semicontinuous on $\mathbf{K}$. Thus, the mapping $\widehat{\mu} \mapsto \int_{\mathbf{K}} r_\gamma \, d\widehat{\mu} \in \mathbf{R}$ on $\mathcal{P}_w(\mathbf{K})$ is u.s.c. on $\mathcal{P}_w(\mathbf{K})$ with respect to the w-weak topology (see, for instance, [17, Lemma 5.2.5]). Now, since $\{\widehat{\mu}_{f_{\Lambda_n}}\}$ converges to the measure $\widehat{\mu}_{\varphi_1}$,

$$\limsup_{n\to\infty} G_\gamma(f_{\Lambda_n}) = \limsup_{n\to\infty} \widehat{\mu}_{f_{\Lambda_n}}(r_\gamma) \leq \widehat{\mu}_{\varphi_1}(r_\gamma) = G_\gamma(\varphi_1). \tag{42}$$

By Assumption 2.2(b1), combined with (7) in Remark 2.4, the definition of θ_{min} and θ_{max} in (8), and (37) in Lemma 5.1, we see that

$$(\Lambda_n - \gamma) \cdot [g_c(f_\gamma) - \theta] \leq \rho(\Lambda_n) - G_\gamma(f_{\Lambda_n}) \leq (\Lambda_n - \gamma) \cdot [\theta_{min} - \theta].$$

Thus,

$$\lim_{n\to\infty} [\rho(\Lambda_n) - G_\gamma(f_{\Lambda_n})] = 0. \tag{43}$$

Since $\rho(\cdot)$ is continuous, (43) implies the limit $\rho(\gamma) = \lim_{n\to\infty} G_\gamma(f_{\Lambda_n})$. Hence, (42) yields that $\rho(\gamma) \leq G_\gamma(\varphi_1)$. On the other hand, the inequality in (28) gives $G_\gamma(\varphi_1) \leq \rho(\gamma)$. Therefore $\rho(\gamma) = G_\gamma(\varphi_1)$. In a similar way we can prove that $\rho(\gamma) = G_\gamma(\varphi_2)$.

Proof of (ii). The function

$$q \mapsto q g_c(\varphi_1) + (1 - q) g_c(\varphi_2) = (q\widehat{\mu}_{\varphi_1} + (1 - q)\widehat{\mu}_{\varphi_2})(c) \quad \forall q \in \mathbf{R}$$

is continuous on $\mathbf{R}$. By (33), there exists $q_0 \in [0,1]$ such that $q_0 g_c(\varphi_1) + (1-q_0) g_c(\varphi_2) = \theta$. On the other hand, since Γ is a convex set we have that $q_0 \widehat{\mu}_{\varphi_1} + (1-q_0)\widehat{\mu}_{\varphi_2} \in \Gamma$. Hence, there exist φ^* such that

$$\widehat{\mu}_{\varphi^*} = q_0 \widehat{\mu}_{\varphi_1} + (1-q_0)\widehat{\mu}_{\varphi_2}. \tag{44}$$

Thus,

$$g_c(\varphi^*) = \widehat{\mu}_{\varphi^*}(c) = \theta. \tag{45}$$

From (34) we have

$$G_\gamma(\varphi^*) = \widehat{\mu}_{\varphi^*}(r_\gamma) = q_0 G_\gamma(\varphi_1) + (1-q_0) G_\gamma(\varphi_2) = \rho(\gamma). \tag{46}$$

Hence, by (45) and (46), it follows that φ^* satisfies (30) in Theorem 4.8. Therefore, φ^* is an optimal policy for the CP. Furthermore, from Lemma 5.2(c) or by Theorem 4.8, we obtain that $V(\theta) = \rho(\gamma) = \min_{\Lambda \le 0} \rho(\Lambda)$.

Proof of (iii). Assume that $\rho(\cdot)$ is differentiable at γ. From part (ii) of this theorem, $\rho(\cdot)$ attains a minimum in $\gamma < 0$. Hence, γ is a critical point of $\rho(\cdot)$. From Theorem 4.8(b), $g_c(f_\gamma) = \theta$ and f_γ solves the CP.

Proof of (iv). If Assumption 4.7 fails to hold then from Lemma 5.2(a) we obtain that $\gamma = 0$. By Lemma 5.1(ii), $\rho(\cdot)$ is nonincreasing on the interval $(-\infty, 0]$, thus

$$\rho(0) = \min_{\Lambda \le 0} \rho(\Lambda) = V(\theta). \tag{47}$$

In a similar way as in the proof of part (i) above, there exists $\varphi_1 \in \Phi$ such that

$$g_c(\varphi_1) \le \theta \quad \text{and} \quad G_0(\varphi_1) = \rho(0). \tag{48}$$

Hence, noting that $g(\varphi_1) = G_0(\varphi_1)$, from (47) and (48), we have that

$$g_c(\varphi_1) \le \theta \quad \text{and} \quad g(\varphi_1) = V(\theta).$$

Then, φ_1 is an optimal policy for the CP.

Finally, if $g_c(f_0) \le \theta$, f_0 is an admisible policy for the ECP (10)-(11). From (47), $g(f_0) = \rho(0) = \min_{\Lambda \le 0} \rho(\Lambda) = V(\theta)$, and so f_0 is an optimal policy for the CP. ■

6 A LQ system

In this section we present a Linear-Quadratic system that satisfies all the hypotheses of Theorems 4.8 and 4.9.

Consider the linear system

$$x_{t+1} = k_1 x_t + k_2 a_t + z_t, \quad t = 0, 1, \cdots, \tag{49}$$

with state space $\mathbf{X} := \mathbf{R}$ and positive coefficients k_1, k_2. The control set is $A := \mathbf{R}$, and the set of admisible controls in each state x is the interval

$$A(x) := [-k_1|x|/k_2, k_1|x|/k_2]. \tag{50}$$

The disturbances z_t in (49) are i.i.d. random variables with values in $Z := \mathbf{R}$, and have zero mean and finite variance, that is,

$$E(z_t) = 0 \quad \text{and} \quad \sigma^2 := E(z_t^2) < \infty. \tag{51}$$

To complete the description of our constrained control model we introduce the quadratic reward-per-stage function

$$r(x, a) := e - (r_1 x^2 + r_2 a^2) \quad \forall (x, a) \in \mathbf{K}, \tag{52}$$

with positive coefficients e, r_1, and r_2, and the cost-per-stage function

$$c(x, a) := c_1 x^2 + c_2 a^2 \quad \forall (x, a) \in \mathbf{K}, \tag{53}$$

with positive coefficients c_1, c_2. We also define

$$W(x) := \exp[\zeta|x|] \quad \text{for all} \quad x \in \mathbf{X}, \tag{54}$$

with $\zeta \geq 2$. Moreover, let $\hat{s} > 0$ be such that

$$\zeta \hat{s} < \log(\zeta/2 + 1)$$

which implies

$$\beta := \frac{2}{\zeta}(\exp[\zeta \hat{s}] - 1) < 1.$$

With this β, we have that Assumption 2.2(b3) holds. On the other hand, observe that r^2, c^2 are functions in $B_W(\mathbf{K})$, and $W \geq 1$. Moreover, $w := \sqrt{W}$ is continuous on $\mathbf{K}$ and it is a moment function on $\mathbf{K}$. Hence, Assumptions 2.2, 2.5 and 2.6 hold.

As in [15, Section 5], we will suppose the following.

Assumption 6.1 $0 < k_1 < 1/2$.

Assumption 6.2 *The i.i.d. disturbances z_t have a common density d, which is a continuous bounded function supported on the interval $S := [-\hat{s}, \hat{s}]$. Moreover, there exists a positive number ε such that $d(s) \geq \varepsilon$ for all $s \in S$.*

Let $S_0 := [0, \hat{s}]$, and let Υ be the Lebesgue measure on $\mathbf{X} = \mathbf{R}$. We define

$$l(x, a) := 1_{S_0}(x) \ \forall (x, a) \in \mathbf{K}, \text{ and } \nu(B) := \varepsilon \Upsilon(B \cap S_0) \ \forall B \in \mathcal{B}(\mathbf{X}). \tag{55}$$

Then, we have that the LQ system (49)-(53) satisfies Lemmas 4.4, 4.5, 4.6, 4.7, 4.8, and 4.9 in [18]. This yields the following.

Proposition 6.3 *Under the Assumptions 6.1 and 6.2, the LQ system (49)-(53) satisfies the Assumptions 2.2, 2.3, 2.5, and 2.6.*

Proposition 6.4 *Suppose that Assumptions 6.1 and 6.2 hold. Then:*

(i) *The LQ system (49)-(53) has a constrained optimal policy. Moreover, for each $\Lambda \leq 0$ let $(\rho(\Lambda), h_\Lambda) \in \mathbf{R} \times B_W(\mathbf{X})$ be a solution to the AROE*

$$h_\Lambda(x) + \rho(\Lambda) = \sup_{a \in A(x)} \Big[r_\Lambda(x,a) + \int_{\mathbf{X}} h_\Lambda(y) Q(dy|x,a) \Big], \tag{56}$$

with $r_\Lambda(x,a) := r_1(\Lambda)x^2 + r_2(\Lambda)a^2 + b$, where $r_i(\Lambda) := \Lambda \cdot c_i - r_i < 0$, $i = 1, 2$, and $b := e - \Lambda \cdot \theta$, then the constrained optimal value $V(\theta)$ satisfies

$$V(\theta) = \min_{\Lambda \leq 0} \rho(\Lambda). \tag{57}$$

(ii) *The function $\Lambda \mapsto \rho(\Lambda)$ is differentiable on the interval $(-\infty, 0)$ with*

$$\frac{d\rho}{d\Lambda}(\Lambda) = g_c(f_\Lambda) - \theta, \quad \forall \Lambda < 0.$$

Furthermore, if $\Lambda < 0$, the following conditions are equivalents:

1) *f_Λ solves the CP;*

2) *Λ is a critical point of $\rho(\cdot)$;*

3) *$g_c(f_\Lambda) = \theta$.*

Thus, if $\Lambda < 0$ satisfies some of the contions 1), 2) or 3), $\rho(\cdot)$ attains a minimum in Λ such that $\rho(\Lambda) = V(\theta) = \min_{\lambda \leq 0} \rho(\lambda)$.

(iii) *Assume that $\gamma := \sup\{\Lambda \leq 0 : g_c(f_\Lambda) \leq \theta\} < 0$, then $\rho(\cdot)$ attains a minimun in γ, and so γ is a critical point of $\rho(\cdot)$. In this case, f_γ satisfies $g_c(f_\gamma) = \theta$ and solves the CP.*

(iv) *If $g_c(f_0) \leq \theta$, then f_0 is an optimal policy for the CP.*

To prove Proposition 6.4 we need the following result which is a slight variation of Lemma 6.5 in [11].

Lemma 6.5 *Let $\widehat{f}$ be a constant, and let $f \in \mathbf{F}$ be a deterministic policy given by $f(x) := -\widehat{f}x$ for all $x \in \mathbf{X}$. Furthermore, let $\widehat{k} := k_1 - k_2\widehat{f}$, where k_1, k_2 are the coefficients in (49). Suppose that $|\widehat{k}| < 1$. Then, for all $x \in \mathbf{X}$*

$$g(f) = \liminf_{n \to \infty} \frac{1}{n} E_x^f \sum_{k=0}^{n-1} r_f(x_k) = e - (r_1 + r_2\widehat{f}^2)\sigma^2/(1 - \widehat{k}^2), \tag{58}$$

and

$$g_c(f) = \limsup_{n \to \infty} \frac{1}{n} E_x^f \sum_{k=0}^{n-1} c_f(x_k) = (c_1 + c_2\widehat{f}^2)\sigma^2/(1 - \widehat{k}^2). \tag{59}$$

with r and c as defined in (52) and (53), respectively.

Proof. Replacing a_t in (49) with $a_t := f(x_t) = -\widehat{f}x_t$, we obtain

$$x_t = (k_1 - k_2\widehat{f})x_{t-1} + z_{t-1} = \widehat{k}x_{t-1} + z_{t-1} \quad \forall t = 1, 2, \cdots.$$

By an induction procedure, for all $t = 1, 2, \cdots$,

$$x_t = \widehat{k}^t x_0 + \sum_{j=0}^{t-1} \widehat{k}^j z_{t-1-j}.$$

From this relation, we obtain

$$E_x^f(x_t^2) = \widehat{k}^{2t}x^2 + \sigma^2(1 - \widehat{k}^{2t})/(1 - \widehat{k}^2).$$

This yields that

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} E_x^f(x_t^2) = \liminf_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} E_x^f(x_t^2) = \sigma^2/(1 - \widehat{k}^2). \tag{60}$$

Since $a = f(x) = -\widehat{f}x$, we obtain

$$r_f(x) = e - (r_1 + r_2\widehat{f}^2)x^2 \quad \text{and} \quad c_f(x) = (c_1 + c_2\widehat{f}^2)x^2 \tag{61}$$

for all $x \in \mathbf{X}$. Finally, inserting (60) in (61) we obtain (58) and (59). ■

Proof of Proposition 6.4. *Proof of (i)* From Proposition 6.3, the assumptions in Propositions 3.2, 3.3, and Theorem 4.3 are satisfied. Hence, the stated result in (i) follows from these results.

Proof of (ii). In [15, Section 5] it is proved, under the Assumptions 6.1 and 6.2, that $\rho(\Lambda)$ in the AROE (56) has the form

$$\rho(\Lambda) = b - v_0(\Lambda)\sigma^2, \tag{62}$$

with σ as in (51), and $v_0(\Lambda)$ is the unique positive solution to the quadratic (so-called Riccati) equation

$$k_2^2 v_0(\Lambda)^2 + [k_2^2 r_1(\Lambda) + k_1^2 r_2(\Lambda) - r_2(\Lambda)]v_0(\Lambda) - r_1(\Lambda)r_2(\Lambda) = 0. \tag{63}$$

Hence, from the fact that $r_i(\Lambda) < 0$, for $i = 1, 2$, we have that $v_0(\Lambda)$ is strictly positive, and depends continuously on Λ. Moreover, we define, for all $x \in \mathbf{X}$

$$f_\Lambda(x) := -\widehat{f_0}(\Lambda)x, \quad \text{with} \quad \widehat{f_0}(\Lambda) := (k_2^2 v_0(\Lambda) - r_2(\Lambda))^{-1} k_1 k_2 v_0(\Lambda). \tag{64}$$

and

$$h_\Lambda(x) := -v_0(\Lambda)x^2. \tag{65}$$

Notice that $\widehat{f_0}(\Lambda)$ depends continuously on the parameter Λ. Since $r_2(\Lambda) < 0$, we have $|f_\Lambda(x)| \le k_1/k_2|x|$. Therefore $f_\Lambda(x) \in A(x)$ for all $x \in \mathbf{X}$, that is, f_Λ is in $\mathbf{F}$. Then, by

a direct calculation we can show that $(h_\Lambda, f_\Lambda, \rho(\Lambda))$ is a canonical triplet that satisfies the AROE (56).

On the other hand, by (59) in Lemma 6.5, we obtain that

$$g_c(f_\Lambda) = (c_1 + c_2\widehat{f_0}(\Lambda)^2)\sigma^2/(1 - \widehat{k}(\Lambda)^2).$$

with $\widehat{k}(\Lambda) := k_1 - k_2\widehat{f_0}(\Lambda)$. From Assumption 6.1 it follows that $|\widehat{k}(\Lambda)| < 1$. Thus, $g_c(f_\Lambda)$ is continuous on the parameter Λ on the interval $(-\infty, 0)$. By Theorem 4.8(b), (d), $\rho(\cdot)$ is differentiable on the interval $(-\infty, 0)$ with continuous derivative

$$\frac{d\rho}{d\Lambda}(\Lambda) = g_c(f_\Lambda) - \theta, \quad \forall \Lambda < 0.$$

The rest of the statements in part (ii) are direct consequences of Theorem 4.8(a), (c).

Proof of (iii). This part follows from Theorem 4.9(iii).

Proof of (iv). This part follows from Theorem 4.9(iv). ■

Case 1. Now we analyse a particular case in which the reward-per-stage function (52) and the cost-per-stage function (53) satisfy $r_1 = r_2$ and $c_1 = c_2$, respectively, and $k_2 = 1$ in (49). For this case, we will find the optimal value and the optimal policy for the LQ model above, with expected and pathwise constraints.

Note that

$$r_1(\Lambda) = r_2(\Lambda) \quad \forall \Lambda \leq 0. \tag{66}$$

By (66), the positive solution of (63) is

$$v_0(\Lambda) = -kr_1(\Lambda) \quad \text{with} \quad k = \frac{k_1^2 + \sqrt{k_1^4 + 4}}{2}. \tag{67}$$

Inserting these values in (62) and using the definition of the constant b, we obtain the explicit form of $\rho(\Lambda)$

$$\rho(\Lambda) = e - (\sigma^2 k) \cdot r_1 + [(\sigma^2 k) \cdot c_1 - \theta]\Lambda \tag{68}$$

which is the equation of a straigth line with slope $(\sigma^2 k) \cdot c_1 - \theta$. Because we need to choose θ satisfying the relation (57), then we will impose the following assumption:

$$(\sigma^2 k) \cdot c_1 < \theta. \tag{69}$$

Under this condition, we have that

$$\begin{aligned} V(\theta) &= \min_{\Lambda \leq 0} \rho(\Lambda) \\ &= \min_{\Lambda \leq 0} \Big(e - (\sigma^2 k) \cdot r_1 + [(\sigma^2 k) \cdot c_1 - \theta]\Lambda\Big) \\ &= e - (\sigma^2 k) \cdot r_1 = \rho(0). \end{aligned} \tag{70}$$

Thus, the minimun is attained at $\Lambda = 0$, and $V(\theta) = \rho(0)$. Furthermore, inserting $\Lambda = 0$ in (64) and (65), we obtain

$$f_0(x) = -\widehat{f_0}x \quad \text{with} \quad \widehat{f_0} := \frac{kk_1}{1+k}, \tag{71}$$

for all $x \in \mathbf{X}$.

Recalling that $r_1 = r_2$ and $c_1 = c_2$, $k_2 = 1$, we have that $|\widehat{k}| = k_1/(1+k) < 1$, with $\widehat{k} := k_1 - \widehat{f_0}$ and k as in (67). By (59) in Lemma 6.5, a direct calculation yields that $g_c(f_0) = (\sigma^2 k)c_1$. Hence, from (69) and by Proposition 6.4(iv), we have that f_0 is and optimal policy for the CP. Finally, by (58) in Lemma 6.5, we obtain that $g(f_0) = e - (\sigma^2 k)r_1$, which coincides with the value of $V(\theta)$ in (70).

Case 2. Consider the LQ system (49)-(53) with the following numerical special case. Suppose that the reward-per-stage function (52) and the cost-per-stage function (53) satisfy $r_1 = 1, r_2 = 2$, $e = 10$, and $c_1 = c_2 = 1$, respectively. Moreover, assume that $k_1 = 1/3$, $k_2 = 1$ in (49), $\theta := 191/180$ and $\sigma^2 = 1$ in (51).

In this particular case, solving the Riccati equation (63), and inserting the corresponding value in (62), we obtain

$$\rho(\Lambda) = \Big(187 - 18.1\Lambda - \sqrt{325\Lambda^2 - 958\Lambda + 697}\Big)/18 \quad \forall \Lambda \leq 0. \tag{72}$$

We consider the critical points of $\rho(\cdot)$. Then, we obtain the unique negative critical point

$$\Lambda_0 = -0.38767819\cdots.$$

By Proposition 6.4(ii), f_{Λ_0} solves the CP. Moreover, $\rho(\cdot)$ attaints its minimum value, which is also the optimal value for the constrained problem, that is

$$V(\theta) = \rho(\Lambda_0) = 8.921767464\cdots, \quad \text{with} \quad \theta = 191/180.$$

In addition

$$v_0 \equiv v_0(\Lambda_0) = 1.48960217\cdots.$$

By (64) and (65), we have that

$$f_{\Lambda_0}(x) = -\widehat{f_0}x \quad \forall x \in \mathbf{R}, \quad \text{with} \quad \widehat{f_0} = 0.12806246\cdots.$$

and

$$h(x) \equiv h_{\Lambda_0}(x) = -v_0 x^2.$$

By a straigthforward calculation, we can check that $(V(\theta), f_{\Lambda_0}, h)$ is a canonical triplet that satisfies the AROE (12) in Proposition 3.2. On the other hand, Proposition 6.4(ii) establishes that $g(f_{\Lambda_0}) = V(\theta)$ and $g_c(f_{\Lambda_0}) = \theta$. We can verify the latter equalities from Lemma 6.5. Indeed, by a direct calculation, we obtain

$$g(f_{\Lambda_0}) = 8.9217674\cdots \quad \text{and} \quad g_c(f_{\Lambda_0}) = 1.061111\cdots = 191/180.$$

So, the constrained problem is solved.

Remark 6.6 *Proposition 6.4(ii)-(iii), give us different methods to obtain f_Λ that solves the constrained problem. For example, we can find Λ_0 in the case 2 above, as the root of the equation*

$$g_c(f_\Lambda) = \theta,$$

which can be easily verified.

Another way is calculating the constant $\gamma = \sup\{\Lambda \leq 0 : g_c(f_\Lambda) \leq \theta\} \leq 0$. If $\gamma < 0$, then f_γ solves the CP.

References

[1] Altman E., *Constrained Markov Decision Processes*, Chapman & Hall/CRC, Boca Raton, FL, 1999.

[2] Belkina T.A. and Rotar V.I., *On optimality in probability and almost surely for processes with a commucation property. I. The discrete time case*, Theory Probab. Appl. 50 (2006), 16-33.

[3] Beutler F. J. and Ross K. W., *Optimal policies for controlled Markov chains with a constraint*, J. Math. Anal. Appl. 112 (1985), 236-252.

[4] Borkar V.S., *Ergodic control of Markov chains with constraints—the general case*, SIAM J. Control Optim. 32 (1994), 176-186.

[5] Ding Y., Jia R. and Tang S., *Dynamical principal agent model based on CMCP*, Math. Methods Oper. Res. 58 (2003), 149-157.

[6] Djonin D.V. and Krishnamurthy V., *MIMO transmission control in fading channels—a constrained Markov decision process formulation with monotone randomized policies*, IEEE Trans. Signal Process. **55** (2007), 5069-5083.

[7] Feinberg E. and Shwartz A., *Constrained discounted dynamic programming*, Math. Oper. Res. 21 (1996), 922-945.

[8] Föllmer, H., and Schied A., *Stochastic Finance. An Introduction in Discrete Time*, Walter de Gruyter & Co, Berlin, 2002.

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

[10] Haviv M., *On constrained Markov decision processes,* Oper. Res. Lett. 19 (1996), 25-28.

[11] Hernández-Lerma O., González-Hernández J. and López-Martínez, R.R., *Constrained average cost Markov control processes in Borel spaces*, SIAM J. Control Optim. 42 (2003), 442-468.

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

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

[14] Hernández-Lerma O., Vega-Amaya O. and Carrasco G., *Sample-path optimality and variance-minimization of average cost Markov control processes*, SIAM J. Control Optim. 38 (1999), 79-93.

[15] Hilgert N. and Hernández-Lerma O., *Bias optimality versus strong 0-discount optimality in Markov control processes with unbounded costs*, Acta Appl. Math. 77 (2003), 215-235.

[16] Krishnamurthy V., Vázquez Abad, F. and Martin K., *Implementation of gradient estimation to a constrained Markov decision problem*, 42nd IEEE Conference on Decision and Control, 2003, pp. 4841-4846.

[17] Mendoza-Pérez A.F., *Pathwise average reward Markov control processes*, Doctoral thesis, CINVESTAV-IPN, México, 2008. Available at http://www.math.cinvestav.mx/ohernand_students

[18] Mendoza-Pérez A.F. and Hernández-Lerma O., *Markov control processes with pathwise constraints*, Math. Methods Oper. Res. 71(2010), 477-502.

[19] Prieto-Rumeau T. and Hernández-Lerma O., *Ergodic control of continuous-time Markov chains with pathwise constraints*, SIAM J. Control Optim. 47 (2008), 1888-1908.

[20] Piunovskiy A.B., *Optimal Control of Random Sequences in Problems with Constraints*, Kluwer, Boston, 1997.

[21] Puterman M.L., *Markov Decision Process*, Wiley, New York, 1994.

[22] Ross K.W. and Varadarajan R., *Markov decision processes with sample path constraints*, Oper. Res. 37 (1989), 780-790.

[23] Ross K.W. and Varadarajan R., *Multichain Markov decision processes with a sample path constraint*, Math. Oper. Res. 16 (1991), 195-207.

[24] Vega-Amaya O., *Expected and sample-path constrained average Markov decision processes*, Internal Report no. 35, Departamento de Matemáticas, Universidad de Sonora. (Submitted.)

Armando F. Mendoza-Pérez
Universidad Politécnica de Chiapas,
Calle Eduardo J.Selvas S/N,
Tuxtla Gutiérrez, Chiapas, MEXICO.
mepa680127@hotmail.com

Onésimo Hernández-Lerma
Mathematics Department
CINVESTAV-IPN
A. Postal 14-740,
México D.F. 07000, MEXICO
ohernand@math.cinvestav.mx