A convex programming approach for discrete-time Markov decision processes under the expected total reward criterion

François Dufour, Alexandre Genadot

A convex programming approach for discrete-time Markov decision processes under the expected total reward criterion

F. Dufour
Institut Polytechnique de Bordeaux
INRIA Bordeaux Sud Ouest, Team: CQFD
IMB, Institut de Mathématiques de Bordeaux, Université de Bordeaux, France
e-mail: francois.dufour@math.u-bordeaux.fr

A. Genadot
IMB, Institut de Mathématiques de Bordeaux, Université de Bordeaux, France
INRIA Bordeaux Sud Ouest, Team: CQFD
e-mail: alexandre.genadot@math.u-bordeaux.fr

Abstract

In this work, we study discrete-time Markov decision processes (MDPs) under constraints with Borel state and action spaces and where all the performance functions have the same form of the expected total reward (ETR) criterion over the infinite time horizon. One of our objective is to propose a convex programming formulation for this type of MDPs. It will be shown that the values of the constrained control problem and the associated convex program coincide and that if there exists an optimal solution to the convex program then there exists a stationary randomized policy which is optimal for the MDP. It will be also shown that in the framework of constrained control problems, the supremum of the expected total rewards over the set of randomized policies is equal to the supremum of the expected total rewards over the set of stationary randomized policies. We consider standard hypotheses such as the so-called continuity-compactness conditions and a Slater-type condition. Our assumptions are quite weak to deal with cases that have not yet been addressed in the literature. An example is presented to illustrate our results with respect to those of the literature.

Keywords: Markov decision process, expected total reward criterion, occupation measure, constraints, convex program.
AMS 2010 Subject Classification: 90C40, 60J10, 90C90.

1 Introduction

We consider a discrete-time Markov decision process with constraints when all the objectives have the same form of the expected total reward over the infinite time horizon. Markov decision processes are a general family of controlled stochastic processes, which are suitable for the modeling of sequential decision-making problems under uncertainty. They arise in many applications, such as engineering, medicine, biology, operations research, management science, economics, among others.

Markov decision processes (MDPs) under the expected total reward (ETR) criterion have been extensively studied by using mainly different approaches, see *e.g.* [9] for a complete and exhaustive survey on that subject and also [15, Chapter 2] for an analysis of that topic through examples.

When dealing with constraints, the linear/convex programming approach (also called the convex analytic method, see, e.g. [4, 5]) has proved to be a very powerful technique for solving MDPs. It

has been extensively studied in the literature and we refer the interested reader to the following works [2, 4, 5, 10, 14] and the references therein to get an overview of this technique. The convex programming approach can be applied to a large class of control problems including for example, the finite-horizon and the infinite-horizon discounted-reward problems; see, e.g., [5] for further examples of performance functions. For such criteria, the key idea is to reformulate the original dynamic control problem as an infinite dimensional static optimization problem over a space of finite measures given by the occupation measures of the controlled process. However, it must be emphasized that the expected total reward criterion is an exception where the convex programming formulation may not be suitable except for very specific models. As mentioned in [5, p. 357-358] and [12, p. 92-93], the ETR criterion is very demanding from a technical point of view and yields some important technical difficulties which are basically of two types:

a) The first issue is directly related to the question of how to properly formulate a convex program associated with an MDP under the ETR criterion. Indeed, as described in [5], the classical and natural approach to formulate a convex program associated to a MDP is to consider as underlying vector space the set of signed finite measures and as variables the occupation measures of the process. However, in the context of the ETR criterion, this approach fails since the occupation measures are not necessarily finite and may take the value infinity. Therefore, the space of finite signed measures is not the appropriate vector space to define the convex program.

b) An important issue is related to the so-called characteristic equation satisfied by the occupation measures of the process which is of the form:

$$\mu_X(\cdot) = \nu(\cdot) + \int_{X\times A} Q(\cdot|x,a)\mu(dx,da)$$

where X and A are respectively the state and action spaces; Q is the transition probability function of the MDP and μ_X is the marginal of the measure μ on X. Indeed, a solution μ to this equation may not correspond to any occupation measures of the controlled process. This difficulty makes the analysis of the ETR criterion very involved by using the convex programming approach.

The objective of the current paper is to propose a suitable convex program for MDPs under the ETR criterion. Our purpose is also to show that the value of the constrained control problem corresponds to the value of an associated convex program and that if there exists an optimal solution to the associated convex program then there exists a stationary randomized policy which is optimal for the MDP. We consider standard assumptions, the so-called continuity-compactness conditions introduced by Schäl in [16, 17]. These assumptions are of two types, namely conditions (S) and (W). Roughly speaking condition (S) requires the transition kernel to be *strongly continuous* whereas condition (W) refers to the case where the transition kernel is *weakly continuous*, see, e.g., [17, p. 367-368] for a precise statement of these assumptions. We also suppose the existence of a policy in the *interior* of the set of admissible policies. This is the so-called Slater condition. Conditions (W) and (S) do not play the same role in the sense that when working with condition (W) instead of condition (S) we have to consider an additional hypothesis requiring the transition kernel of the model to be absolutely continuous with respect to a Markov kernel uniformly in the action variables. Our approach differs from that classically considered in the literature in the sense that the variables of the convex program are not given by the occupation measures of the controlled process but defined on the positive cone of the vector space given by the pair of finite signed stochastic kernels on the action space given the state space.

When compared to the literature, our results appear complementary and our assumptions are rather weak. The references dealing with the ETR criterion by using the convex programming formulation are very scarce in the literature. As for our work, the results in [6, 8] are concerned with general Borel state and action spaces. However, it is important to observe that the approach proposed in [6, 8] does not correspond to a linear/convex programming formulation of an MDP under the ETR criterion. Indeed, the underlying variables of the optimization problem under consideration are given by measures that may take the value infinity and therefore, this set does not enjoy the structure of a standard vector space. This technical issue aside, the results of the current paper differ significantly from those obtained in [6, 8]. The approach developed in [6] deals with models satisfying condition (W) and strongly relies on the positiveness of the cost functions. It must be emphasized that the general framework of signed cost functions cannot be addressed with the technique presented in [6]. In [8], the model under consideration satisfies condition (S) and it was assumed that the transition kernel is absolutely continuous with respect to a reference probability measure uniformly in the state and action variables. In the present work, we show that this assumption is not needed under condition (S). It must be also observed that the approach developed in [8] for signed cost function cannot be applied under condition (W). In [2, Chapter 8], the model is transient or absorbing and is restricted to discrete state and action spaces. Here, we do not impose the MDP to be transient or absorbing. Another advantage of our approach is to propose a convex programming formulation for constrained MDPs under the ETR criterion with signed reward functions and satisfying condition (W). In this context, such formulation has not been so far investigated in the literature. It should be also mentioned that in our work we imposed the so-called Slater condition which is not required in [2, 6, 8]. However, this condition is rather weak and it is a standard assumption in convex optimization problems with constraints, see e.g. [3].

The rest of the paper is organized as follows. In Section 2, we present the control problem that will be considered throughout this work. The assumptions and the convex programming formulation of a constrained discrete-time MDP under the ETR criterion is introduced in Section 3. Important properties of the convex program as well as the constrained control problem are established in Section 4. Our main results are presented in Section 5 showing that the original control problem is equivalent to the convex program. Section 6 is dedicated to the presentation of an example illustrating our results. Finally, a technical result used in Section 4 is derived in an appendix.

2 Description of the control problem

The main goal of this section is to introduce the notation, the parameters defining the model, and to present the construction of the controlled process.

2.1 Notation and terminology

The following basic notation will be used in the forthcoming.

The set of integers is denoted by $\mathbb{Z}$ and $\mathbb{N}$ corresponds to the non-negative integers, that is, $\mathbb{N} = \{0, 1, 2, \ldots\}$. The set of real numbers is given by $\mathbb{R}$. For any subset $\mathbb{D}$ of $\mathbb{R}$, $\mathbb{D}^*$ denotes $\mathbb{D} \setminus \{0\}$ and $\mathbb{D}_+ = \{d \in \mathbb{D} : d \geq 0\}$. We write $\mathbb{N}_p$ for $\{1, \ldots, p\}$ with $p \in \mathbb{N}^*$, $\overline{\mathbb{R}}$ is the set of extended real numbers, that is, $\mathbb{R} \cup \{-\infty, +\infty\}$ and $\overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{+\infty\}$. Given x and y in the Euclidean space $\mathbb{R}^n$, let $\langle x, y \rangle$ be the usual inner product of x and y. By $|x| = \langle x, x \rangle^{1/2}$ we will denote the norm of $x \in \mathbb{R}^n$. Let $\mathbf{0}_n$ be the element of $\mathbb{R}^n$ with all components equal to zero. If θ_1 and θ_2 are in $\mathbb{R}^n$, we

shall write $\theta_1 \geq \theta_2$ when all the components of θ_1 are greater than or equal to the corresponding components of θ_2.

Let X be a metric space and denote by $\mathfrak{B}(X)$ its associated Borel σ-algebra. We use the symbol f^+ (respectively f^-) to denote the positive part (respectively, negative part) of a function $f : X \to \overline{\mathbb{R}}$. The function $\mathcal{I}_\infty$ is the function whose values are constant and equal to $+\infty$. If X is a metric space, $\mathfrak{M}(X)$ denotes the set of real-valued measurable functions defined on X. Furthermore, $\mathcal{C}(X)$ is the space of real-valued bounded continuous functions defined on X. The term measure will always refer to a countably additive, $\overline{\mathbb{R}}_+$-valued set function. The set of measures defined on $(X, \mathfrak{B}(X))$ is denoted by $\boldsymbol{\mathcal{M}}(X)$ and the set of probability measures on $(X, \mathfrak{B}(X))$ by $\boldsymbol{\mathcal{P}}(X)$. For $\mu \in \boldsymbol{\mathcal{M}}(X)$ and a positive function h in $\mathfrak{M}(X)$, $\mu(h) = \int_X h(x)\mu(dx)$ and for $g \in \mathfrak{M}(X)$, $\mu(g)$ is defined by $\mu(g^+) - \mu(g^-)$ where by convention $(+\infty) - (+\infty) = -\infty$. Consider two metric spaces X and Y. If μ is a measure on $X \times Y$ then μ_X denotes the marginal of the measure μ on X. A kernel K on X given Y is a $\overline{\mathbb{R}}_+$-valued mapping defined on $\mathfrak{B}(X) \times Y$ such that for any $y \in Y$, $K(\cdot|y) \in \boldsymbol{\mathcal{M}}(X)$ and for any $\Lambda \in \mathfrak{B}(X)$, $K(\Lambda|\cdot)$ is a measurable function defined on Y. A kernel K on X given Y is said to be finite if $K(X|y) \in \mathbb{R}_+$ for any $y \in Y$. The set of finite kernels on X given Y is denoted $\boldsymbol{\mathcal{K}}(X|Y)$. A stochastic (or Markov) kernel K on X given Y is a kernel in $\boldsymbol{\mathcal{K}}(X|Y)$ satisfying $K(X|y) = 1$ for any $y \in Y$. The set of stochastic kernels on X given Y will be denoted by $\boldsymbol{\mathcal{P}}(X|Y)$. Let Q be a stochastic kernel on X given Y, then, for a function $v : X \to \overline{\mathbb{R}}$, we define $Qv : Y \to \overline{\mathbb{R}}$ as

$$Qv(y) := \int_X v^+(x)Q(dx|y) - \int_X v^-(x)Q(dx|y),$$

provided that v is quasi-integrable with respect to the probability measure $Q(\cdot|y)$ for any $y \in Y$. For a measure μ on Y, we denote by μQ the measure $\displaystyle\int_Y Q(\cdot|y)\mu(dy)$ on X.

2.2 The control model.

Let us consider the stationary model

$$\big(\mathbf{X}, \mathbf{A}, \{\mathbf{A}(x) : x \in \mathbf{X}\}, Q, r, c, \theta_*, \nu\big) \tag{1}$$

consisting of:

(a) A Borel space $\mathbf{X}$ (that is, a Borel subset of a complete and separable metric space), which is the state space.

(b) A Borel space $\mathbf{A}$, representing the control or action set.

(c) A family $\{\mathbf{A}(x) : x \in \mathbf{X}\}$ of non-empty measurable subsets of $\mathbf{A}$, where $\mathbf{A}(x)$ is the set of feasible controls or actions when the system is in state $x \in \mathbf{X}$. We suppose that

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

is a measurable subset of $\mathbf{X} \times \mathbf{A}$. There exists a measurable map $\vartheta : \mathbf{X} \to \mathbf{A}$ with $\vartheta(x) \in \mathbf{A}(x)$. For notational convenience, we introduce recursively the set $\mathbf{H}_t$ of histories up to time $t \in \mathbb{N}^*$ by defining $\mathbf{H}_1 = \mathbf{X}$ and $\mathbf{H}_t = \mathbf{K}^{t-1} \times \mathbf{X}$ for $t \geq 2$.

(d) A stochastic kernel Q on $\mathbf{X}$ given $\mathbf{K}$, which stands for the transition probability function.

(e) The one-step reward function is given by a measurable function $r : \mathbf{K} \to \mathbb{R}$.

(f) For $i \in \mathbb{N}_q$, the measurable mappings $c_i : \mathbf{K} \to \mathbb{R}$ are the one-step constraint functions.

(g) The constraint limits are real numbers given by $\theta^* = \{\theta_i^*\}_{i\in\mathbb{N}_q}$.

(h) Finally, the initial distribution is $\nu \in \boldsymbol{\mathcal{P}}(\mathbf{X})$.

A control policy (a policy, for short) is a sequence $\pi = \{\pi_t\}_{t\in\mathbb{N}^*}$ of stochastic kernels π_t on $\mathbf{A}$ given $\mathbf{H}_t$ such that $\pi_t(\mathbf{A}(x_t)|h_t) = 1$ for any $h_t = (x_1, a_1, \ldots, x_t) \in \mathbf{H}_t$. Let Π be the set of all policies. A policy $\pi = \{\pi_t\}_{t\in\mathbb{N}^*} \in \Pi$ is called a stationary randomized policy if there exists a stochastic kernel $\boldsymbol{\varphi}$ on $\mathbf{A}$ given $\mathbf{X}$ satisfying $\boldsymbol{\varphi}(\mathbf{A}(x)|x) = 1$ for any $x \in \mathbf{X}$ and $\pi_t(\cdot|h_t) = \boldsymbol{\varphi}(\cdot|x_t)$ for any $h_t = (x_1, a_1, \ldots, x_t) \in \mathbf{H}_t$ and $t \in \mathbb{N}^*$. In such as case, we will write $\boldsymbol{\varphi}$ instead of π to emphasize that the corresponding stationary randomized policy π is generated by $\boldsymbol{\varphi}$. Let Π_s be the set of all stationary randomized policies.

To state the optimal control problem we are concerned with, we introduce the canonical space $(\Omega, \mathcal{F})$ consisting of the set of sample paths $\Omega = (\mathbf{X} \times \mathbf{A})^\infty$ and the associated product σ-algebra $\mathcal{F}$. The projection from Ω to the state space and the action space at time t are denoted by X_t and A_t. That is, for

$$\omega = (y_1, b_1, \ldots, y_t, b_t \ldots) \in \Omega \quad \text{we have} \quad X_t(\omega) = y_t \ \text{ and } \ A_t(\omega) = b_t$$

for $t \in \mathbb{N}^*$. Consequently, $\{X_t\}_{t\in\mathbb{N}^*}$ is the state process and $\{A_t\}_{t\in\mathbb{N}^*}$ is the control process. It is a well known result that for every policy $\pi \in \Pi$ and any initial probability measure ν on $(\mathbf{X}, \mathfrak{B}(\mathbf{X}))$ there exists a unique probability measure $\mathbb{P}_\nu^\pi$ on $(\Omega, \mathcal{F})$ such that $\mathbb{P}_\nu^\pi(\mathbf{K}^\infty) = 1$ and

$$\mathbb{P}_\nu^\pi(X_1 \in B) = \nu(B), \quad \text{for } B \in \mathfrak{B}(\mathbf{X}),$$

$$\mathbb{P}_\nu^\pi(X_{t+1} \in C|\sigma\{X_1, \ldots, X_t, A_t\}) = Q(C|X_t, A_t) \quad \text{for } C \in \mathfrak{B}(\mathbf{X}),$$

$$\mathbb{P}_\nu^\pi(A_t \in D|\sigma\{X_1, \ldots, X_{t-1}, A_{t-1}, X_t\}) = \pi_t(D|X_1, \ldots, X_{t-1}, A_{t-1}, X_t) \quad \text{for } D \in \mathfrak{B}(\mathbf{A}),$$

$\mathbb{P}_\nu^\pi - a.s.$, for any $t \in \mathbb{N}^*$.

The expectation with respect to $\mathbb{P}_\nu^\pi$ is denoted by $\mathbb{E}_\nu^\pi$. The so-called *occupation measure* generated by a policy $\pi \in \Pi$, denoted by μ^π, is defined by

$$\mu^\pi(\Gamma) = \sum_{t=1}^{\infty} \mathbb{P}_\nu^\pi\big((X_t, A_t) \in \Gamma\big)$$

for any $\Gamma \in \mathfrak{B}(\mathbf{X} \times \mathbf{A})$. Denote by $\boldsymbol{\mathcal{O}}$ (respectively, $\boldsymbol{\mathcal{O}}_s$) the set of occupation measures generated by randomized (respectively, stationary) policies.

Statement of the control problem.
For $h \in \mathfrak{M}(\mathbf{K})$ and $\pi \in \Pi$, define $\mathcal{J}_\nu(h, \pi)$ by

$$\mathcal{J}_\nu(h, \pi) = \sum_{t=1}^{\infty} \mathbb{E}_\nu^\pi\big[h^+(X_t, A_t)\big] - \sum_{t=1}^{\infty} \mathbb{E}_\nu^\pi\big[h^-(X_t, A_t)\big]$$

where by convention $(+\infty) - (+\infty) = -\infty$. In fact, assumptions will be introduced in the next section to avoid dealing with such cases. Observe that $\mathcal{J}_\nu(h, \pi)$ can be written equivalently in terms of the occupation measure generated by the policy $\pi \in \Pi$ as follows

$$\mathcal{J}_\nu(h, \pi) = \mu^\pi(h).$$

In this paper, we will repeatedly use this equality without mentioning it.

Definition 2.1 *A policy $\pi \in \Pi$ is said to be admissible if $\mathcal{J}_\nu(c_i, \pi) \geq \theta_i^*$ for $i \in \mathbb{N}_q$. The set of admissible policies will be denoted by Π_{θ^*}. The optimal control problem we consider consists in maximizing the expected reward $\mathcal{J}_\nu(r, \pi)$ over the set of admissible policies $\pi \in \Pi_{\theta^*}$. The value associated to this constrained control problem is given by $\sup\left\{\mathcal{J}_\nu(r, \pi) : \pi \in \Pi_{\theta^*}\right\}$. A policy $\hat{\pi} \in \Pi$ is optimal if $\hat{\pi} \in \Pi_{\theta^*}$ and $\mathcal{J}_\nu(r, \hat{\pi}) = \sup\left\{\mathcal{J}_\nu(r, \pi) : \pi \in \Pi_{\theta^*}\right\}$.*

3 Assumptions and the convex programming formulation

The objective of this section is both to list the assumptions we will use in this work and to introduce the convex program associated with the control problem we presented in the previous section. In this work, we deal with MDPs satisfying the so-called Conditions (W) or (S) which are standard hypotheses of the literature, see for example [16].

Condition (W)

(W1) For any $x \in \mathbf{X}$, the action set $\mathbf{A}(x)$ is compact and the multifunction from $\mathbf{X}$ to $\mathbf{A}$ defined by $x \to \mathbf{A}(x)$ is upper-semicontinuous.

(W2) For any $f \in \boldsymbol{\mathcal{C}}(\mathbf{X})$, Qf is continuous on $\mathbf{K}$.

(W3) The reward r and the constraint c_i for $i \in \mathbb{N}_q$ are upper-semicontinuous on $\mathbf{K}$.

Condition (S)

(S1) For any $x \in \mathbf{X}$, $\mathbf{A}(x)$ is compact.

(S2) For any $x \in \mathbf{X}$ and $\Lambda \in \mathfrak{B}(\mathbf{X})$, $Q(\Lambda|x, \cdot)$ is continuous on $\mathbf{A}(x)$.

(S3) For any $x \in \mathbf{X}$, the reward $r(x, \cdot)$ and the constraint $c_i(x, \cdot)$ for $i \in \mathbb{N}_q$ are upper-semicontinuous on $\mathbf{A}(x)$.

In order to introduce the convex program associated to an MDP under the ETR criterion, we need to make some hypotheses. First, it is assumed that the transition kernel Q of the MDP under consideration is absolutely continuous with respect to a Markov kernel P (see Assumption A). This hypothesis is rather weak and is satisfied in a large number of practical cases as discussed in the remark below.

Assumption A. There exists $P \in \boldsymbol{\mathcal{P}}(\mathbf{X}|\mathbf{X})$ satisfying $Q(\cdot|x, a) \ll P(\cdot|x)$ for any $(x, a) \in \mathbf{K}$. Associated to the kernel P, p will denote the probability measure on $\mathbf{X}$ defined by

$$p(dx) = \sum_{k \in \mathbb{N}} \frac{1}{2^{k+1}} \nu P^k(dx). \tag{2}$$

Remark 3.1 *1. In Lemma 3.2 below, it is shown that under Conditions (S1) and (S2), Assumption A is satisfied.*

2. If the sets of feasible actions are countable, that is $\mathbf{A}(x) = \{a_k(x)\}_{k \in \mathbb{N}^*}$ where for any $k \in \mathbb{N}^*$ a_k is a measurable function from $\mathbf{X}$ to $\mathbf{A}$ then Assumption A is satisfied for P defined by

$$P(dy|x) = \sum_{k \in \mathbb{N}^*} \frac{1}{2^k} Q(dy|x, a_k(x)),$$

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

3. *If $Q(\cdot|x,a) \ll q(\cdot)$ for any $(x,a) \in \mathbf{K}$ then clearly Assumption A is satisfied. This condition corresponds to the main hypothesis used in [8]. It is of course less general than Assumption A but it is naturally satisfied for a large class of practical systems. Indeed, in many applications, the evolution of an MDP is specified by a discrete-time equation of the form $x_{t+1} = F(x_t, a_t) + \xi_t$ where F is an $\mathbb{R}^n$-valued measurable mapping defined on $\mathbb{R}^n \times A$ and $(\xi_t)_{t\in\mathbb{N}^*}$ is an independent and identically distributed sequence of random variables with density α with respect to the Lebesgue measure on $\mathfrak{B}(\mathbb{R}^n)$. By using the change of variable formula, we obtain that $Q(A|x,a) = \int_A \alpha(y - F(x,a))dy$ showing that $Q(\cdot|x,a) \ll q(\cdot)$ for any $(x,a) \in \mathbf{K}$ is satisfied for q defined for example by the standard normal distribution on $\mathfrak{B}(\mathbb{R}^n)$.*
Observe also that when $\mathbf{X}$ is finite or countable, $Q(\cdot|x,a) \ll q(\cdot)$ for any $(x,a) \in \mathbf{K}$ is satisfied when q is given for example by a geometric distribution.

Lemma 3.2 *Conditions (S1) and (S2) imply Assumption A, that is, $Q \ll P$ with $P \in \boldsymbol{\mathcal{P}}(\mathbf{X}|\mathbf{X})$ given by*

$$P(dy|x) = \sum_{k\in\mathbb{N}^*} \frac{1}{2^k} Q(dy|x, \xi_k(x)) \tag{3}$$

where $\{\xi_k\}_{k\in\mathbb{N}^*}$ is a sequence of measurable selectors from the multifunction defined from $\mathbf{X}$ to $\mathbf{A}$ by $x \to \mathbf{A}(x)$ and satisfying $\mathbf{A}(x) = \overline{\{\xi_n(x) : n \in \mathbb{N}^*\}}$ for any $x \in \mathbf{X}$.

Proof: The multifunction $\mathfrak{A}$ from $\mathbf{X}$ to $\mathbf{A}$ defined by $x \to \mathbf{A}(x)$ is by assumption Borel measurable and so, weakly measurable. From (S1), Corollary 18.15 in [1] gives the existence of a sequence $\{\xi_n\}_{n\in\mathbb{N}^*}$ of measurable selectors from the multifunction $\mathfrak{A}$ satisfying $\mathbf{A}(x) = \overline{\{\xi_n(x) : n \in \mathbb{N}^*\}}$ for any $x \in \mathbf{X}$. Now by using (S2), we obtain that $Q(dy|x,a) \ll P(dy|x)$ for any $(x,a) \in \mathbf{K}$ for the Markov kernel P defined by (3). □

Remark 3.3 *The previous proof is an extension of an argument used in the proof of Theorem 1 in [13, p. 183].*

In the next definition, we introduce the set of feasible variables. It will be shown below that it is a convex subset of the vector space of finite signed kernels on $\mathbf{A}$ given $\mathbf{X}$.

Definition 3.4 *Suppose Assumption A holds and let p be the measure introduced in (2).*

- *For $\Phi = (\varphi^\infty, \varphi^*) \in \boldsymbol{\mathcal{K}}(\mathbf{A}|\mathbf{X})^2$, η^Φ will denote the measure in $\boldsymbol{\mathcal{M}}(\mathbf{X} \times \mathbf{A})$ given by*

$$\eta^\Phi(dx, da) = \mathcal{I}_\infty(x)\varphi^\infty(da|x)p(dx) + \varphi^*(da|x)p(dx), \tag{4}$$

  recalling that $\mathcal{I}_\infty$ is constant function equal to infinity.

- *Consider $\boldsymbol{\mathcal{K}}_p$ as the set of $\Phi = (\varphi^\infty, \varphi^*) \in \boldsymbol{\mathcal{K}}(\mathbf{A}|\mathbf{X})^2$ satisfying*

$$\varphi^\infty(\mathbf{A}|x) + \varphi^*(\mathbf{A}|x) > 0,$$

$$\varphi^\infty(\mathbf{A}(x)^c|x) + \varphi^*(\mathbf{A}(x)^c|x) = 0,$$

  and

$$\eta^\Phi_{\mathbf{X}} = \nu + \eta^\Phi Q.$$

  Any $\Phi \in \boldsymbol{\mathcal{K}}_p$ induces a measure η^Φ that will be called the $\boldsymbol{\mathcal{K}}_p$-measure generated by Φ. $\boldsymbol{\mathcal{K}}_p$ is called the set of feasible variables.

Remark 3.5 *Observe first that* $\alpha\Phi_1 + (1-\alpha)\Phi_2 \in \boldsymbol{\mathcal{K}}_p$ *and in particular,*

$$\eta^{\alpha\Phi_1+(1-\alpha)\Phi_2} = \alpha\eta^{\Phi_1} + (1-\alpha)\eta^{\Phi_2}, \tag{5}$$

for any $\alpha \in [0,1]$ *and* $(\Phi_1, \Phi_2) \in \boldsymbol{\mathcal{K}}_p^2$. *Therefore,* $\boldsymbol{\mathcal{K}}_p$ *is a convex subset of the vector space of signed finite kernel on* $\mathbf{A}$ *given* $\mathbf{X}$.

Definition 3.6 *Let* $\Phi = (\varphi^\infty, \varphi^*) \in \boldsymbol{\mathcal{K}}_p$. *Introduce the kernel* φ_Φ *on* $\mathbf{A}$ *given* $\mathbf{X}$ *defined by*

$$\varphi_\Phi(da|x) = I_{\boldsymbol{\mathcal{E}}_\Phi^c}(x)\frac{\varphi^\infty(da|x)}{\varphi^\infty(\mathbf{A}|x)} + I_{\boldsymbol{\mathcal{E}}_\Phi}(x)\frac{\varphi^*(da|x)}{\varphi^*(\mathbf{A}|x)}. \tag{6}$$

where

$$\boldsymbol{\mathcal{E}}_\Phi = \{x \in \mathbf{X} : \varphi^\infty(\mathbf{A}|x) = 0\}. \tag{7}$$

Observe that φ_Φ *is a stochastic kernel satisfying* $\varphi_\Phi(\mathbf{A}(x)|x) = 1$ *for any* $x \in \mathbf{X}$. *The stationary randomized policy* φ_Φ *will be called the policy induced by* Φ.

We will also need the following technical hypothesis:

Assumption B.

(B.1) $\sup\{\eta^\Phi(r^+) : \Phi \in \boldsymbol{\mathcal{K}}_p\}$ and $\sup\{\eta^\Phi(c_i^+) : \Phi \in \boldsymbol{\mathcal{K}}_p\} < +\infty$ for any $i \in \mathbb{N}_q$.

(B.2) $\mu(r^-) < +\infty$ and $\mu(c_i^-) < +\infty$ for any $\mu \in \boldsymbol{\mathcal{O}}$, $i \in \mathbb{N}_q$.

This hypothesis is comparable to Assumption (A2) introduced in [8, p. 847]. Assumption (B.1) essentially imposes that the values of the unconstrained convex programs associated to a reward function given by either r or c_i for $i \in \mathbb{N}_q$ are different from $+\infty$ while Assumption (B.2) ensure that the performance criteria associated to the reward r and the constraints c_i for $i \in \mathbb{N}_q$ are not equal $-\infty$. In particular, Assumption (B.1) will be used to introduce the linear program.

Definition 3.7 *Suppose Assumptions A and (B.1) hold. The convex program, denoted by* $\boldsymbol{\mathcal{KP}}_p$, *consists in maximizing* $\eta^\Phi(r)$ *over* $\Phi \in \boldsymbol{\mathcal{K}}_p$ *subject to* $\eta^\Phi(c_i) \geq \theta_i^*$ *for any* $i \in \mathbb{N}_q$. *The value of the convex program is given by*

$$\sup\{\eta^\Phi(r) : \Phi \in \boldsymbol{\mathcal{K}}_p \text{ and } \eta^\Phi(c_i) \geq \theta_i^* \text{ for } i \in \mathbb{N}_q\}. \tag{8}$$

A variable $\hat{\Phi} \in \boldsymbol{\mathcal{K}}_p$ *is said to be an optimal solution to the convex program* $\boldsymbol{\mathcal{KP}}_p$ *if*

$$\eta^{\hat{\Phi}}(r) = \sup\{\eta^\Phi(r) : \Phi \in \boldsymbol{\mathcal{K}}_p \text{ and } \eta^\Phi(c_i) \geq \theta_i^* \text{ for } i \in \mathbb{N}_q\}$$

and $\eta^{\hat{\Phi}}(c_i) \geq \theta_i^*$ *for any* $i \in \mathbb{N}_q$.

Remark 3.8 *Let* h *be a function given by either* r *or* c_i *for* $i \in \mathbb{N}_q$. *From Assumption (B.1), it follows that* $\alpha\eta^{\Phi_1}(h) + (1-\alpha)\eta^{\Phi_2}(h)$ *is well defined for any* $\alpha \in [0,1]$ *and* $(\Phi_1, \Phi_2) \in \boldsymbol{\mathcal{K}}_p^2$. *Therefore, we obtain from equation* (5) *that*

$$\eta^{\alpha\Phi_1+(1-\alpha)\Phi_2}(h) = \alpha\eta^{\Phi_1}(h) + (1-\alpha)\eta^{\Phi_2}(h)$$

for any $\alpha \in [0,1]$ *and* $(\Phi_1, \Phi_2) \in \boldsymbol{\mathcal{K}}_p^2$. *This implies that the mathematical program defined in* (8) *is indeed a convex program. In [3, p. 153], a convex program is written in terms of an infimum. The* $\boldsymbol{\mathcal{KP}}_p$ *program introduced in Definition 3.7 can be equivalently written in terms of an infimum by changing the sign of the reward function. We prefer to keep this setting to deal with an MDP under a reward optimization criterion.*

Finally, we introduce an additional standard hypothesis:

The Slater condition

There exists $\mu^* \in \boldsymbol{\mathcal{O}}$ such that $\theta_i^* < \mu^*(c_i)$ for any $i \in \mathbb{N}_q$.

4 Preliminary results

The main goal of this section is to establish several properties of the constrained control problem as well as properties of the convex program.

4.1 Properties of the convex program

In this subsection, we will show in Lemma 4.2 that for any stationary randomized policy $\pi \in \Pi_s$ there exists $\Phi \in \boldsymbol{\mathcal{K}}_p$ such that the $\boldsymbol{\mathcal{K}}_p$-measure generated by Φ is equal to the occupation measure generated by the stationary randomized policy π. An important result which is a cornerstone of the paper is presented at the end of this subsection. It can be roughly stated as follows: for any feasible variable $\Phi \in \boldsymbol{\mathcal{K}}_p$ of the convex program, the reward $\mathcal{J}_\nu(h, \varphi_\Phi)$ associated to the stationary randomized policy $\varphi_\Phi \in \Pi_s$ is greater than $\eta^\Phi(h)$ for specific functions h that will be discussed in Theorem 4.3. To get these results, we first need to establish that the occupation measures of the controlled process have a special structure, that is, the marginal on $\mathbf{X}$ of any occupation measure is absolutely continuous with respect to the probability measure p introduced in Assumption A.

Lemma 4.1 *Suppose Assumption A holds. Then for any $\mu \in \boldsymbol{\mathcal{O}}$,*

$$\mu_{\mathbf{X}}(dx) \ll p(dx) \tag{9}$$

where $p \in \boldsymbol{\mathcal{P}}(\mathbf{X})$ is defined in (2).

Proof: For any $\mu \in \boldsymbol{\mathcal{O}}$, it can be easily shown from Lemma 9.4.3 in [11] the existence of an increasing sequence of finite measures $\{\mu_k\}_{k\in\mathbb{N}^*}$ on $\mathbf{X}$ and a sequence of stochastic kernels $\{\varphi_k\}_{k\in\mathbb{N}^*}$ on $\mathbf{A}$ given $\mathbf{X}$ satisfying $\varphi_k(\mathbf{A}(x)|x) = 1$ and

$$\lim_{k\to\infty} \mu_k(\Lambda) = \mu_{\mathbf{X}}(\Lambda) \tag{10}$$

and

$$\mu_{k+1}(\Lambda) = \nu(\Lambda) + \int_{\mathbf{X}} \int_{\mathbf{A}} Q(\Lambda|x,a)\varphi_k(da|x)\mu_k(dx) \tag{11}$$

for $\Lambda \in \mathfrak{B}(\mathbf{X})$, $k \in \mathbb{N}^*$ and $\mu_1 = \nu$. Let us show by induction that $\mu_k \ll p$ for any $k \in \mathbb{N}^*$. We have clearly $\mu_1 \ll p$. Assume that $\mu_k \ll p$. Observe that $\displaystyle\int_{\mathbf{A}} Q(\cdot|x,a)\varphi_k(da|x) \ll P(\cdot|x)$ for any $x \in \mathbf{X}$ implying that

$$\int_{\mathbf{X}} \int_{\mathbf{A}} Q(\cdot|x,a)\varphi_k(da|x)\mu_k(dx) \ll \int_{\mathbf{X}} P(\cdot|x)p(dx)$$

and so, combining (2) and (11) we have $\mu_{k+1} \ll p$. We obtain the result by using (10). □

As a consequence, we can show that the set of the $\boldsymbol{\mathcal{K}}_p$-measures contains the occupation mesures generated by the stationary randomized policies.

Lemma 4.2 *Suppose Assumption A holds. For any $\pi \in \Pi_s$, there exists $\Phi \in \boldsymbol{\mathcal{K}}_p$ such that*

$$\mu^{\pi} = \eta^{\Phi}.$$

Proof: Let $\pi \in \Pi_s$. Clearly, the increasing sequence $\{\mu_t^{\pi}\}_{t\in\mathbb{N}^*}$ of finite measures defined on $\mathbf{X}\times\mathbf{A}$ by

$$\mu_t^{\pi}(\Gamma) = \sum_{k=1}^{t} \mathbb{P}_{\nu}^{\pi}\big((X_k, A_k) \in \Gamma\big)$$

for any $\Gamma \in \mathfrak{B}(\mathbf{X}\times\mathbf{A})$ converges to μ^{π}. From Lemma 4.1, there exists a sequence of increasing measurable $\mathbb{R}_+$-valued functions $\{\mathcal{D}_t\}_{t\in\mathbb{N}^*}$ defined on $\mathbf{X}$ such that $\sum_{k=1}^{t} \mathbb{P}_{\nu}^{\pi}(X_k \in \Lambda) = \int_{\Lambda} \mathcal{D}_t(x)p(dx)$ for $\Lambda \in \mathfrak{B}(\mathbf{X})$ and so, $\mu_t^{\pi}(dx,da) = \mathcal{D}_t(x)\pi(da|x)p(dx)$. Therefore,

$$\begin{aligned}\mu^{\pi}(dx,da) &= \mathcal{D}(x)\pi(da|x)p(dx)\\ &= \mathcal{I}_{\infty}(x)I_{\{\mathcal{D}(x)=\infty\}}\pi(da|x)p(dx) + \mathcal{D}(x)I_{\{\mathcal{D}(x)<\infty\}}\pi(da|x)p(dx)\end{aligned}$$

where $\mathcal{D}(x) = \lim_{t\to\infty}\mathcal{D}_t(x)$. Consequently, $\Phi = (\varphi^{\infty}, \varphi^*)$ defined by $\varphi^{\infty}(da|x) = I_{\{\mathcal{D}(x)=\infty\}}\pi(da|x)$ and $\varphi^*(da|x) = \mathcal{D}(x)I_{\{\mathcal{D}(x)<\infty\}}\pi(da|x)$ belongs to $\boldsymbol{\mathcal{K}}_p$ since $\mu_{\mathbf{X}}^{\pi} = \nu + \mu^{\pi}Q$. □

The following result is in a way a converse of the previous one. It is a key result in our work. Roughly speaking, it states that for any feasible variable $\Phi \in \boldsymbol{\mathcal{K}}_p$ of the convex program, the reward $\mathcal{J}_{\nu}(h,\varphi_{\Phi})$ associated to the stationary randomized policy $\varphi_{\Phi} \in \Pi_s$ is greater than $\eta^{\Phi}(h)$ for specific functions h described below.

Theorem 4.3 *Suppose that Assumption A holds. For any $\Phi \in \boldsymbol{\mathcal{K}}_p$, there exists $\varphi_{\Phi} \in \Pi_s$ such that*

$$\mathcal{J}_{\nu}(h,\varphi_{\Phi}) \geq \eta^{\Phi}(h),$$

for any $h \in \mathfrak{M}(\mathbf{K})$ satisfying $\sup\{\eta^{\Phi}(h^+) : \Phi \in \boldsymbol{\mathcal{K}}_p\} < +\infty$.

Proof: For $h \in \mathfrak{M}(\mathbf{K})$ satisfying $\sup\{\eta^{\Phi}(h^+) : \Phi \in \boldsymbol{\mathcal{K}}_p\} < +\infty$, let us prove the result by showing that

$$\mu^{\varphi_{\Phi}}(h) \geq \eta^{\Phi}(h) \tag{12}$$

where φ_{Φ} is the stationary randomized policy induced by Φ (see (6)). There is no loss of generality to assume that $\eta^{\Phi}(h) > -\infty$ and so we have $\eta^{\Phi}(|h|) < \infty$. We are going to proceed by contradiction to get (12). More precisely, if $\mu^{\varphi_{\Phi}}(h) < \eta^{\Phi}(h)$ then we will introduce a sequence $\{\Psi_k\}_{k\in\mathbb{N}}$ in $\boldsymbol{\mathcal{K}}_p$ satisfying $\lim_{k\to\infty}\eta^{\Psi_k}(h) = +\infty$ contradicting the hypothesis. The proof is divided into two steps. We will first introduce $\{\Psi_k\}_{k\in\mathbb{N}}$ and show that $\Psi_k \in \boldsymbol{\mathcal{K}}_p$ for any $k \in \mathbb{N}$. In a second step, it will be proven that $\lim_{k\to\infty}\eta^{\Psi_k}(h) = +\infty$ showing the result.

First step: construction of a sequence $\{\Psi_k\}_{k\in\mathbb{N}}$ in $\boldsymbol{\mathcal{K}}_p$.

Let $\mu^{\varphi_{\Phi}}$ be the occupation measure induced by the stationary randomized policy φ_{Φ}. As in the proof of Lemma 4.2, there exists a measurable $\overline{\mathbb{R}}_+$-valued function $\mathcal{D}_{\varphi_{\Phi}}$ defined on $\mathbf{X}$ satisfying

$$\mu^{\varphi_{\Phi}}(dx,da) = \mathcal{D}_{\varphi_{\Phi}}(x)\varphi_{\Phi}(da|x)p(dx). \tag{13}$$

For $k \in \mathbb{N}$, consider $\Psi_k = (\psi^{\infty}, \psi_k^*)$ where $\psi^{\infty} \in \boldsymbol{\mathcal{K}}(\mathbf{A}|\mathbf{X})$ is given by

$$\psi^{\infty}(da|x) = I_{\boldsymbol{\mathcal{E}}_{\Phi}^{c}}(x)\varphi_{\Phi}(da|x) \tag{14}$$

and ψ_k^* is a signed kernel on $\mathbf{A}$ given $\mathbf{X}$ defined by

$$\psi_k^*(da|x) = I_{\boldsymbol{\mathcal{E}}_\Phi}(x)\Big[\varphi^*(\mathbf{A}|x) + k\big[\varphi^*(\mathbf{A}|x) - \mathcal{D}_{\varphi_\Phi}(x)\big]\Big]\varphi_\Phi(da|x) + (k+1)I_{\boldsymbol{\mathcal{E}}_\Phi^c}(x)\varphi^*(da|x). \quad (15)$$

Observe that in the previous definition, $\varphi^*(\mathbf{A}|x) - \mathcal{D}_{\varphi_\Phi}(x)$ is well defined since $\varphi^* \in \boldsymbol{\mathcal{K}}(\mathbf{A}|\mathbf{X})$. To get the result, we will proceed in two steps. First we will show that $\varphi^*(\mathbf{A}|\cdot) \geq \mathcal{D}_{\varphi_\Phi}(\cdot)$ on $\boldsymbol{\mathcal{E}}_\Phi$ implying that $\psi_k^* \in \boldsymbol{\mathcal{K}}(\mathbf{A}|\mathbf{X})$ and so, $\Psi_k \in \boldsymbol{\mathcal{K}}(\mathbf{A}|\mathbf{X})^2$ for any $k \in \mathbb{N}$. In a second step, we will prove that $\Psi_k \in \boldsymbol{\mathcal{K}}_p$.

• Let us show that $\Psi_k \in \boldsymbol{\mathcal{K}}(\mathbf{A}|\mathbf{X})^2$.
From (4), $\eta^\Phi(dx, da) = \mathcal{I}_\infty(x)\varphi^\infty(da|x)p(dx) + \varphi^*(da|x)p(dx)$ and so, by using (7)

$$\eta^\Phi(dx, da) \;=\; I_{\boldsymbol{\mathcal{E}}_\Phi}(x)\varphi^*(da|x)p(dx) + I_{\boldsymbol{\mathcal{E}}_\Phi^c}(x)\mathcal{I}_\infty(x)\varphi^\infty(da|x)p(dx) + I_{\boldsymbol{\mathcal{E}}_\Phi^c}(x)\varphi^*(da|x)p(dx),$$

where by convention $0 \times \infty = 0$. Recalling the Definition of φ_Φ (see equation (6)), we easily obtain $I_{\boldsymbol{\mathcal{E}}_\Phi}(x)\varphi^*(da|x) = I_{\boldsymbol{\mathcal{E}}_\Phi}(x)\varphi^*(\mathbf{A}|x)\varphi_\Phi(da|x)$ and $I_{\boldsymbol{\mathcal{E}}_\Phi^c}(x)\mathcal{I}_\infty(x)\varphi^\infty(da|x) = I_{\boldsymbol{\mathcal{E}}_\Phi^c}(x)\mathcal{I}_\infty(x)\varphi_\Phi(da|x)$ and so, we get

$$\begin{aligned}\eta^\Phi(dx, da) =& I_{\boldsymbol{\mathcal{E}}_\Phi}(x)\varphi^*(\mathbf{A}|x)\varphi_\Phi(da|x)p(dx) + I_{\boldsymbol{\mathcal{E}}_\Phi^c}(x)\mathcal{I}_\infty(x)\varphi_\Phi(da|x)p(dx) \\ &+ I_{\boldsymbol{\mathcal{E}}_\Phi^c}(x)\varphi^*(da|x)p(dx). \end{aligned} \quad (16)$$

Therefore,

$$\begin{aligned}\eta_{\mathbf{X}}^\Phi(dx) &= \Big[I_{\boldsymbol{\mathcal{E}}_\Phi}(x)\varphi^*(\mathbf{A}|x) + I_{\boldsymbol{\mathcal{E}}_\Phi^c}(x)\big[\mathcal{I}_\infty(x) + \varphi^*(\mathbf{A}|x)\big]\Big]p(dx) \\ &= \Big[I_{\boldsymbol{\mathcal{E}}_\Phi}(x)\varphi^*(\mathbf{A}|x) + I_{\boldsymbol{\mathcal{E}}_\Phi^c}(x)\mathcal{I}_\infty(x)\Big]p(dx). \end{aligned} \quad (17)$$

Since $\eta_{\mathbf{X}}^\Phi = \nu + \eta^\Phi Q$, we have by using (16)

$$\begin{aligned}\eta_{\mathbf{X}}^\Phi(\Lambda) =& \nu(\Lambda) + \int_{\boldsymbol{\mathcal{E}}_\Phi} Q^{\varphi_\Phi}(\Lambda|x)\varphi^*(\mathbf{A}|x)p(dx) + \int_{\boldsymbol{\mathcal{E}}_\Phi^c} Q^{\varphi_\Phi}(\Lambda|x)\mathcal{I}_\infty(x)p(dx) \\ &+ \int_{\boldsymbol{\mathcal{E}}_\Phi^c} Q^{\varphi^*}(\Lambda|x)p(dx), \end{aligned} \quad (18)$$

and with (17) it follows

$$\eta_{\mathbf{X}}^\Phi(\Lambda) = \nu(\Lambda) + \eta_{\mathbf{X}}^\Phi Q^{\varphi_\Phi}(\Lambda) + \int_{\boldsymbol{\mathcal{E}}_\Phi^c} Q^{\varphi^*}(\Lambda|x)p(dx).$$

However, $\mu_{\mathbf{X}}^{\varphi_\Phi}$ is the minimal solution to the equation $\beta = \nu + \beta Q^{\varphi_\Phi}$ and so, $\mu_{\mathbf{X}}^{\varphi_\Phi} \leq \eta_{\mathbf{X}}^\Phi$. Combining equations (13) and (17), we obtain $\Big[I_{\boldsymbol{\mathcal{E}}_\Phi}(\cdot)\varphi^*(\mathbf{A}|\cdot) + I_{\boldsymbol{\mathcal{E}}_\Phi^c}(\cdot)\mathcal{I}_\infty(\cdot)\Big] \geq \mathcal{D}_{\varphi_\Phi}(\cdot)\ p-a.s.$. Consequently, $\mathcal{D}_{\varphi_\Phi}(\cdot) \leq \varphi^*(\mathbf{A}|\cdot)\ p-a.s.$ on $\boldsymbol{\mathcal{E}}_\Phi$ and according to the definition of $\mathcal{D}_{\varphi_\Phi}(\cdot)$ (see equation (13)), there is no loss of generality to claim

$$\mathcal{D}_{\varphi_\Phi}(\cdot) \leq \varphi^*(\mathbf{A}|\cdot) \text{ on } \boldsymbol{\mathcal{E}}_\Phi. \quad (19)$$

Therefore, $\psi_k^* \in \boldsymbol{\mathcal{K}}(\mathbf{A}|\mathbf{X})$ and so, $\Psi_k \in \boldsymbol{\mathcal{K}}(\mathbf{A}|\mathbf{X})^2$ for any $k \in \mathbb{N}$.

• Let us show that $\Psi_k \in \boldsymbol{\mathcal{K}}_p$.
Recalling the definition Ψ_k (see equations (14)-(15)), we have $\psi^\infty(\mathbf{A}(x)^c|x) + \psi_k^*(\mathbf{A}(x)^c|x) = 0$ and

$\psi^{\infty}(\mathbf{A}|x)+\psi_k^*(\mathbf{A}|x)\geq I_{\boldsymbol{\mathcal{E}}_\Phi}(x)\varphi^*(\mathbf{A}|x)+I_{\boldsymbol{\mathcal{E}}_\Phi^c}(x)>0$ for any $x\in\mathbf{X}$. The only point which remains to prove is that $\eta^{\Psi_k}(dx,da)=\mathcal{I}_\infty(x)\psi^{\infty}(da|x)p(dx)+\psi_k^*(da|x)p(dx)$ satisfies

$$\eta_{\mathbf{X}}^{\Psi_k} \quad = \quad \nu+\eta^{\Psi_k}Q. \tag{20}$$

Combining the definition of Ψ_k (see equations (14)-(15)) and the expression of η^{Φ} (see equation (16)), we obtain

$$\eta^{\Psi_k} \quad = \quad \eta^{\Phi}+k\gamma \tag{21}$$

where $\gamma\in\boldsymbol{\mathcal{M}}(\mathbf{X}\times\mathbf{A})$ is given by

$$\gamma(dx,da)=I_{\boldsymbol{\mathcal{E}}_\Phi}(x)\big[\varphi^*(\mathbf{A}|x)-\mathcal{D}_{\varphi_\Phi}(x)\big]\varphi_\Phi(da|x)p(dx)+I_{\boldsymbol{\mathcal{E}}_\Phi^c}(x)\varphi^*(da|x)p(dx). \tag{22}$$

To show that (20) holds, we will consider two cases.
a) Firstly, we will show that equation (20) is satisfied on $\mathfrak{B}(\boldsymbol{\mathcal{E}}_\Phi)$. For that, let us consider $\Lambda\in\mathfrak{B}(\boldsymbol{\mathcal{E}}_\Phi)$. From (21), we have $\eta_{\mathbf{X}}^{\Psi_k}(\Lambda)=\eta_{\mathbf{X}}^{\Phi}(\Lambda)+k\gamma_{\mathbf{X}}(\Lambda)$. However, $\eta_{\mathbf{X}}^{\Phi}(\Lambda)=\nu(\Lambda)+\eta^{\Phi}Q(\Lambda)$ showing that $\eta_{\mathbf{X}}^{\Psi_k}(\Lambda)=\nu(\Lambda)+\eta^{\Phi}Q(\Lambda)+k\gamma_{\mathbf{X}}(\Lambda)$. If we show that $\gamma_{\mathbf{X}}(\Lambda)=\gamma Q(\Lambda)$ then $\eta_{\mathbf{X}}^{\Psi_k}(\Lambda)=\nu(\Lambda)+\eta^{\Psi_k}Q(\Lambda)$ implying that (20) holds on $\mathfrak{B}(\boldsymbol{\mathcal{E}}_\Phi)$. To see that $\gamma_{\mathbf{X}}(\Lambda)=\gamma Q(\Lambda)$, observe from (22) that

$$\gamma_{\mathbf{X}}(\Lambda) \quad = \quad \int_\Lambda\big[\varphi^*(\mathbf{A}|x)-\mathcal{D}_{\varphi_\Phi}(x)\big]p(dx).$$

Assuming that $\eta_{\mathbf{X}}^{\Phi}(\Lambda)<\infty$ and combining (13), (17) and the previous equation we have

$$\gamma_{\mathbf{X}}(\Lambda) \quad = \quad \int_\Lambda\big[\varphi^*(\mathbf{A}|x)-\mathcal{D}_{\varphi_\Phi}(x)\big]p(dx)=\eta_{\mathbf{X}}^{\Phi}(\Lambda)-\mu_{\mathbf{X}}^{\varphi_\Phi}(\Lambda). \tag{23}$$

Now, we obtain by using (18) and the fact that $\eta_{\mathbf{X}}^{\Phi}(\Lambda)<\infty$

$$\int_{\boldsymbol{\mathcal{E}}_\Phi^c}Q^{\varphi_\Phi}(\Lambda|x)\mathcal{I}_\infty(x)p(dx)=0 \tag{24}$$

implying also

$$\int_{\boldsymbol{\mathcal{E}}_\Phi^c}Q^{\varphi_\Phi}(\Lambda|x)\mathcal{D}_{\varphi_\Phi}(x)p(dx)=0. \tag{25}$$

Now, combining (18) and (24)

$$\eta_{\mathbf{X}}^{\Phi}(\Lambda)=\nu(\Lambda)+\int_{\boldsymbol{\mathcal{E}}_\Phi}Q^{\varphi_\Phi}(\Lambda|x)\varphi^*(\mathbf{A}|x)p(dx)+\int_{\boldsymbol{\mathcal{E}}_\Phi^c}Q^{\varphi^*}(\Lambda|x)p(dx).$$

Recalling that $\mu_{\mathbf{X}}^{\varphi_\Phi}=\nu+\mu_{\mathbf{X}}^{\varphi_\Phi}Q^{\varphi_\Phi}$, we have with (13) and (25)

$$\mu_{\mathbf{X}}^{\varphi_\Phi}(\Lambda) \quad = \quad \nu(\Lambda)+\int_{\boldsymbol{\mathcal{E}}_\Phi}Q^{\varphi_\Phi}(\Lambda|x)\mathcal{D}_{\varphi_\Phi}(x)p(dx).$$

The two previous equations gives

$$\eta_{\mathbf{X}}^{\Phi}(\Lambda)-\mu_{\mathbf{X}}^{\varphi_\Phi}(\Lambda) \quad = \quad \int_{\boldsymbol{\mathcal{E}}_\Phi}Q^{\varphi_\Phi}(\Lambda|x)\big[\varphi^*(\mathbf{A}|x)-\mathcal{D}_{\varphi_\Phi}(x)\big]p(dx)+\int_{\boldsymbol{\mathcal{E}}_\Phi^c}Q^{\varphi^*}(\Lambda|x)p(dx). \tag{26}$$

From (23) and (26)

$$\gamma_{\mathbf{X}}(\Lambda) \;=\; \int_{\boldsymbol{\mathcal{E}}_\Phi} Q^{\varphi_\Phi}(\Lambda|x)\big[\varphi^*(\mathbf{A}|x) - \mathcal{D}_{\varphi_\Phi}(x)\big]p(dx) + \int_{\boldsymbol{\mathcal{E}}_\Phi^c} Q^{\varphi^*}(\Lambda|x)p(dx)$$

Recalling the definition of γ (see (22)) we get $\gamma_{\mathbf{X}}(\Lambda) = \gamma Q(\Lambda)$ for $\Lambda \in \mathfrak{B}(\boldsymbol{\mathcal{E}}_\Phi)$ with $\eta_{\mathbf{X}}^\Phi(\Lambda) < \infty$. However, equation (17) implies that $\eta_{\mathbf{X}}^\Phi$ is σ-finite on $\boldsymbol{\mathcal{E}}_\Phi$ and combining (16) and (22), we have $\gamma_{\mathbf{X}} \leq \eta_{\mathbf{X}}^\Phi$. Therefore, it follows that $\gamma_{\mathbf{X}}(\Lambda) = \gamma Q(\Lambda)$ for any Λ in $\mathfrak{B}(\boldsymbol{\mathcal{E}}_\Phi)$, and so (20) holds on $\mathfrak{B}(\boldsymbol{\mathcal{E}}_\Phi)$.

b) Secondly, we will show that equation (20) is satisfied on $\mathfrak{B}(\boldsymbol{\mathcal{E}}_\Phi^c)$. For that, let $\Lambda \in \mathfrak{B}(\boldsymbol{\mathcal{E}}_\Phi^c)$. It is important to observe from (17) that in this case $\eta_{\mathbf{X}}^\Phi(\Lambda) = 0$ or $+\infty$. Therefore, we obtain on one hand $\eta_{\mathbf{X}}^{\Psi_k}(\Lambda) = \eta_{\mathbf{X}}^\Phi(\Lambda) + k\gamma_{\mathbf{X}}(\Lambda) = \eta_{\mathbf{X}}^\Phi(\Lambda)$ by recalling (21) and using the fact that $\gamma_{\mathbf{X}} \leq \eta_{\mathbf{X}}^\Phi$ and on the other hand $\eta_{\mathbf{X}}^\Phi(\Lambda) = \eta_{\mathbf{X}}^\Phi(\Lambda) + k\gamma Q(\Lambda)$ since by (22)

$$\gamma Q(\Lambda) \;=\; \int_{\boldsymbol{\mathcal{E}}_\Phi} Q^{\varphi_\Phi}(\Lambda|y)\big[\varphi^*(\mathbf{A}|y) - \mathcal{D}_{\varphi_\Phi}(y)\big]p(dy) + \int_{\boldsymbol{\mathcal{E}}_\Phi^c} Q^{\varphi^*}(\Lambda|y)p(dy) \leq \eta_{\mathbf{X}}^\Phi(\Lambda)$$

where the last inequality comes from (18). Therefore, $\eta_{\mathbf{X}}^{\Psi_k}(\Lambda) = \eta_{\mathbf{X}}^\Phi(\Lambda) + k\gamma Q(\Lambda) = \nu\Lambda) + \eta^{\Psi_k} Q(\Lambda)$ showing that (20) holds on $\mathfrak{B}(\boldsymbol{\mathcal{E}}_\Phi^c)$.

Finally, equation (20) is satisfied and as a consequence $\Psi_k \in \boldsymbol{\mathcal{K}}_p$ for any $k \in \mathbb{N}$.

Second step: $\lim_{k\to\infty} \eta^{\Psi_k}(h) = +\infty$.

Recalling that $\eta^\Phi(|h|) < \infty$, we get from (16)

$$\int_{\boldsymbol{\mathcal{E}}_\Phi^c} |h(x,a)|\mathcal{I}_\infty(x)\varphi_\Phi(da|x)p(dx) = 0$$

implying also

$$\int_{\boldsymbol{\mathcal{E}}_\Phi^c} |h(x,a)|\mathcal{D}_{\varphi_\Phi}(x)\varphi_\Phi(da|x)p(dx) = 0.$$

Therefore, combining (13), (16), (22) and the two previous equations we obtain easily that

$$\eta^\Phi(h) - \mu^{\varphi_\Phi}(h) = \gamma(h).$$

If $\eta^\Phi(h) > \mu^{\varphi_\Phi}(h)$ then $\gamma(h) > 0$ and $\lim_{k\to\infty} \eta^{\Psi_k}(h) = \eta^\Phi(h) + \lim_{k\to\infty} k\gamma(h) = +\infty$ giving the result. □

4.2 Properties of the constrained control problem

The main objective of this subsection is to show that in the framework of constrained control problems, the supremum of the expected total rewards over the set of randomized policies is equal to the supremum of the expected total rewards over the set of *stationary* randomized policies. Our results use Theorem A.1 presented in the Appendix which is a slight modification of Theorem 1 in Schäl [17] who has established a stronger version of this type of result but in the unconstrained case. To use Schäl's results, we need to impose Conditions (W) or (S) and in addition, to deal with the constrained case, we need to impose a Slater-type condition.

The next technical Lemma shows that, roughly speaking, under Assumption (B.1), the unconstrained control problems associated to a reward function given by either r or c_i for $i \in \mathbb{N}_q$ are different from $+\infty$.

Lemma 4.4 *Suppose Assumptions A and (B.1) and either Conditions (W) or (S) hold. Then,*

$$\sup\left\{\mu(r^+) : \mu \in \boldsymbol{\mathcal{O}}\right\} < +\infty \text{ and } \sup\left\{\mu(c_i^+) : \mu \in \boldsymbol{\mathcal{O}}\right\} < +\infty,$$

for $i \in \mathbb{N}_q$.

Proof: The idea is to apply Theorem A.1 to the unconstrained models associated to the reward functions given by one of the following mappings: r^+ and c_i^+ for $i \in \mathbb{N}_q$. Clearly, the Convergence Assumption and the Continuity and Compactness Assumptions in [17, p. 367] are satisfied. Therefore, we have by using Theorem A.1

$$\sup\{\mu(h) : \mu \in \boldsymbol{\mathcal{O}}\} = \sup\{\mu(h) : \mu \in \boldsymbol{\mathcal{O}}_s\}$$

for any function h given by either r^+ or c_i^+ for $i \in \mathbb{N}_q$. Now, from Assumption A we can apply Lemma 4.2 to have

$$\sup\{\mu(h) : \mu \in \boldsymbol{\mathcal{O}}_s\} \leq \sup\{\eta^{\Phi}(h) : \Phi \in \boldsymbol{\mathcal{K}}_p\}.$$

Recalling Assumption (B.1) we obtain the result. □

The next result shows that if the Slater condition is satisfied for an arbitrary policy then there exists a stationary randomized policy satisfying the same type of condition.

Proposition 4.5 *Suppose Assumptions A, B and either Conditions (W) or (S) hold. If the Slater condition is satisfied, then there exists* $\widetilde{\mu} \in \boldsymbol{\mathcal{O}}_s$ *satisfying* $\theta_i^* < \widetilde{\mu}(c_i)$ *for any* $i \in \mathbb{N}_q$.

Proof: The result is proved by induction. Applying Theorem A.1 for the unconstrained model associated to the reward function c_1, we have

$$\sup\{\mu(c_1) : \mu \in \boldsymbol{\mathcal{O}}\} = \sup\{\mu(c_1) : \mu \in \boldsymbol{\mathcal{O}}_s\}.$$

Since $\mu^*(c_1) > \theta_1^*$ (by recalling the Slater condition), we have $\sup\{\mu(c_1) : \mu \in \boldsymbol{\mathcal{O}}_s\} > \theta_1^*$ implying the existence of $\mu_1 \in \boldsymbol{\mathcal{O}}_s$ such that $\mu_1(c_1) > \theta_1^*$. For $n \in \mathbb{N}_{q-1}$, let us assume the existence of $\mu_n \in \boldsymbol{\mathcal{O}}_s$ such that $\mu_n(c_i) > \theta_i^*$ for $i \in \mathbb{N}_n$. Therefore, we can combine Lemma 4.4 and Proposition A.2 to obtain

$$\begin{aligned}\sup\left\{\mu(c_{n+1}) : \mu \in \boldsymbol{\mathcal{O}} \text{ and } \mu(c_i) > \theta_i^* \text{ for } i \in \mathbb{N}_n\right\} \\ = \sup\left\{\mu(c_{n+1}) : \mu \in \boldsymbol{\mathcal{O}}_s \text{ and } \mu(c_i) > \theta_i^* \text{ for } i \in \mathbb{N}_n\right\}.\end{aligned}$$

However,

$$\sup\left\{\mu(c_{n+1}) : \mu \in \boldsymbol{\mathcal{O}} \text{ and } \mu(c_i) > \theta_i^* \text{ for } i \in \mathbb{N}_n\right\} \geq \mu^*(c_{n+1}) > \theta_{n+1}^*$$

implying the existence of $\mu_{n+1} \in \boldsymbol{\mathcal{O}}_s$ such that $\mu_{n+1}(c_i) > \theta_i^*$ for $i \in \mathbb{N}_{n+1}$. This gives the result. □

Below is the main result of this subsection that states roughly speaking that in the framework of constrained control problems, the supremums of the expected total rewards over the set of randomized policies and over the set of stationary randomized policies coincide.

Theorem 4.6 *Suppose Assumptions A, B and either Conditions (W) or (S) hold. If the Slater condition is satisfied, then*

$$\begin{aligned}\sup\{\mu(r) : \mu \in \boldsymbol{\mathcal{O}} \text{ and } \mu(c_i) \geq \theta_i^* \text{ for } i \in \mathbb{N}_q\} \\ = \sup\{\mu(r) : \mu \in \boldsymbol{\mathcal{O}}_s \text{ and } \mu(c_i) \geq \theta_i^* \text{ for } i \in \mathbb{N}_q\}.\end{aligned}$$

Proof: Applying Proposition 4.5, there exists $\widetilde{\mu} \in \boldsymbol{\mathcal{O}}_s$ satisfying the Slater condition, that is, $\widetilde{\mu}(c_i) > \theta_i^*$ for $i \in \mathbb{N}_q$. Now, combining Lemma 4.4 and Proposition A.2, we obtain the result. □

5 Main results

In this section, we present the main results of this paper showing that the original control problem is equivalent to the convex program introduced in Definition 3.7 for a weakly or strongly continuous transition kernel.

The case of Condition (W)

Theorem 5.1 *Suppose Assumptions A, B and Condition (W) hold. If the Slater condition is satisfied, then*

$$\sup\left\{\mathcal{J}_\nu(r,\pi) : \pi \in \Pi_{\theta^*}\right\} = \sup\left\{\eta^\Phi(r) : \Phi \in \boldsymbol{\mathcal{K}}_p \text{ and } \eta^\Phi(c_i) \geq \theta_i^* \text{ for } i \in \mathbb{N}_q\right\} \tag{27}$$

where $p \in \boldsymbol{\mathcal{P}}(\mathbf{X})$ is defined in (2). Moreover, if $\hat{\Phi}$ is an optimal solution to the convex program $\boldsymbol{\mathcal{KP}}_p$ then the stationary randomized policy $\varphi_{\hat{\Phi}}$ induced by $\hat{\Phi}$ is optimal for the constrained control problem, that is,

$$\mathcal{J}_\nu(r,\varphi_{\hat{\Phi}}) = \sup\left\{\mathcal{J}_\nu(r,\pi) : \pi \in \Pi_{\theta^*}\right\}. \tag{28}$$

Proof: Theorem 4.6 states that

$$\sup\left\{\mathcal{J}_\nu(r,\pi) : \pi \in \Pi_{\theta^*}\right\} = \sup\left\{\mathcal{J}_\nu(r,\pi) : \pi \in \Pi_s \cap \Pi_{\theta^*}\right\}.$$

However, from Lemma 4.2, we have

$$\sup\left\{\mathcal{J}_\nu(r,\pi) : \pi \in \Pi_s \cap \Pi_{\theta^*}\right\} \leq \sup\left\{\eta^\Phi(r) : \Phi \in \boldsymbol{\mathcal{K}}_p \text{ and } \eta^\Phi(c_i) \geq \theta_i^* \text{ for } i \in \mathbb{N}_q\right\},$$

Now, consider $\Phi \in \boldsymbol{\mathcal{K}}_p$. By using Theorem 4.3, $\mathcal{J}_\nu(h,\varphi_\Phi) \geq \eta^\Phi(h)$ for h given either r or c_i for $i \in \mathbb{N}_q$ implying that $\varphi_\Phi \in \Pi_s \cap \Pi_{\theta^*}$ and also the reverse inequality

$$\sup\left\{\mathcal{J}_\nu(r,\pi) : \pi \in \Pi_s \cap \Pi_{\theta^*}\right\} \geq \sup\left\{\eta^\Phi(r) : \Phi \in \boldsymbol{\mathcal{K}}_p \text{ and } \eta^\Phi(c_i) \geq \theta_i^* \text{ for } i \in \mathbb{N}_q\right\}$$

showing the first part of the result.

Now if $\hat{\Phi} \in \boldsymbol{\mathcal{K}}_p$ is an optimal solution to the convex program $\boldsymbol{\mathcal{KP}}_p$ then $\eta^{\hat{\Phi}}(c_i) \geq \theta_i^*$ for any $i \in \mathbb{N}_q$ and $\eta^{\hat{\Phi}}(r) = \sup\left\{\eta^\Phi(r) : \Phi \in \boldsymbol{\mathcal{K}}_p \text{ and } \eta^\Phi(c_i) \geq \theta_i^* \text{ for } i \in \mathbb{N}_q\right\}$. Therefore, the stationary randomized policy $\varphi_{\hat{\Phi}} \in \Pi_{\theta^*}$ satisfies $\mathcal{J}_\nu(r,\varphi_{\hat{\Phi}}) \geq \eta^{\hat{\Phi}}(r)$ by using Theorem 4.3. Now, by using the first part of the result (see equation (27)) it follows that $\mathcal{J}_\nu(r,\varphi_{\hat{\Phi}}) \geq \sup\left\{\mathcal{J}_\nu(r,\pi) : \pi \in \Pi_{\theta^*}\right\}$ giving the last part of the result. □

Remark 5.2 *As mentioned in the introduction, the previous result has the advantage of proposing a convex programming formulation for constrained MDPs under the ETR criterion with signed reward functions and satisfying condition (W) which has not been so far addressed in the literature. In [6], the authors do not really analyse a convex program, but study a related optimization problem where the MPDs under consideration satisfy condition (W) but the proposed approach strongly relies on the positiveness of the cost functions and cannot be generalized to the framework of signed cost functions.*

The case of condition (S)

Theorem 5.3 *Suppose Assumptions B and Condition (S) hold. If the Slater condition is satisfied, then*

$$\sup\left\{\mathcal{J}_\nu(r,\pi) : \pi \in \Pi_{\theta^*}\right\} = \sup\left\{\eta^\Phi(r) : \Phi \in \boldsymbol{\mathcal{K}}_p \text{ and } \eta^\Phi(c_i) \geq \theta_i^* \text{ for } i \in \mathbb{N}_q\right\}$$

where p is defined in (2) *for P given by* (3)*. Moreover, if $\hat{\Phi}$ is an optimal solution to the convex program then the stationary randomized policy $\varphi_{\hat{\Phi}}$ induced by $\hat{\Phi}$ is optimal for the constrained control problem introduced in Definition 2.1.*

Proof: Up to the definition of p whose existence is established in Lemma 3.2, the proof of this result is identical to that of Theorem 5.1. □

Remark 5.4 *In [8], the authors do not really analyse a convex program but study a related optimization problem where the MPDs under consideration satisfy condition (S) by assuming that the transition kernel is absolutely continuous with respect to a reference probability measure uniformly in the state and action variables. In the previous result, we show that this assumption is not needed under condition (S) if this hypothesis is replaced by a Slater-type condition.*

6 Example

In this section, we provide an example with one constraint to illustrate our results and compare them with reference [8]. The results obtained in [6] cannot be used for this model because the contraint function takes positive and negative values. We will show that one of the conditions of [8] is not satisfied while the approach developed in the present paper can be applied. This example shows that there is a gap between the initial optimization problem and the mathematical program associated to the measures satisfying the characteristic equation, that is,

$$\begin{aligned}\sup\left\{\mathcal{J}_\nu(r,\pi) : \pi \in \Pi \text{ and } \mathcal{J}_\nu(c_1,\pi) \geq \theta_1^*\right\}& \\ < \sup\left\{\mu(r) : \mu \in \boldsymbol{\mathcal{M}}(\mathbf{X}),\ \mu_{\mathbf{X}} = \nu + \mu Q \text{ and } \mu(c_1) \geq \theta_1^*\right\}&.\end{aligned}$$

It means that the characteristic equation $\mu_{\mathbf{X}} = \nu + \mu Q$ generates measures that do not correspond to any occupation measures of the process. This type of measures has been called in [7] phantom solutions of the characteristic equation. The interesting point is that at the same time, we may have

$$\sup\left\{\mathcal{J}_\nu(r,\pi) : \pi \in \Pi \text{ and } \mathcal{J}_\nu(c_1,\pi) \geq \theta_1^*\right\} = \sup\left\{\eta^\Phi(r) : \Phi \in \boldsymbol{\mathcal{K}}_p \text{ and } \eta^\Phi(c_1) \geq \theta_1^*\right\}.$$

This means that the set $\left\{\eta^\Phi : \Phi \in \boldsymbol{\mathcal{K}}_p\right\}$ which is by the way a subset of $\left\{\mu \in \boldsymbol{\mathcal{M}}(\mathbf{X}) : \mu_{\mathbf{X}} = \nu + \mu Q\right\}$ may generate less of phantom solutions.

Two different values of the constraint limit θ_1^* will be studied. For the first value of the constraint limit, it will be shown that the approach proposed in the present paper can be applied implying that the value of the original control problem coincides with the value of the convex program $\boldsymbol{\mathcal{KP}}_p$. When changing the value of the constraint limit, the Slater condition will not be satisfied. However, it is interesting to observe that in this latter case, the values of the original control problem and its associated convex program $\boldsymbol{\mathcal{KP}}_p$ still coincide although the Slater condition is not fulfilled. It appears that the Slater condition is not a necessary condition to establish the correspondance between the constrained control problem and its associated convex program $\boldsymbol{\mathcal{KP}}_p$.

We consider the control model

$$\left(\mathbf{X}, \mathbf{A}, Q, r, c_1, \theta_1^*, \nu\right)$$

where $\mathbf{X} = \mathbb{Z} \cup \{\Delta\}$ and the action set is given by $\mathbf{A} = \{a, b\}$. For $x \neq 1$, $\mathbf{A}(x) = \{a\}$; $\mathbf{A}(1) = \{a, b\}$ and $\mathbf{A}(\Delta) = \{a\}$. The stochastic kernel Q is given by $Q(x+1|x, a) = 1$ for $x \leq 0$ and $Q(y|x, a) = (1/2)I_{\{x+1\}}(y) + (1/2)I_{\{x+2\}}(y)$, for $x \geq 1$ and finally, $Q(\Delta|1, b) = Q(\Delta|\Delta, a) = 1$. The one-step reward function is given by $r(x, a) = (1/2)^{|x|}$ for $x \neq 1$; $r(1, a) = r(1, b) = 1/2$ and $r(\Delta, a) = 0$. The one-step constraint function is given by $c_1(x) = (-1/2)^{|x|}$ for $x \neq 1$; $c_1(1, a) = -1/18$ and $c_1(1, b) = 1$. The initial distribution ν satisfies $\nu(\{1\}) = \nu(\{\Delta\}) = 1/2$. The constraint limit is given by θ_1^*. Two cases are studied: $\theta_1^* = 1/4$ and $\theta_1^* = 1/2$.

Let $\mu \in \boldsymbol{\mathcal{M}}(\mathbf{X})$ satisfying the characteristic equation $\mu_{\mathbf{X}} = \nu + \mu Q$ and so, $\mu(\Delta, a) = +\infty$; $\mu(x, a) = \mu(0, a)$ for $x \leq 0$; $\mu(1, a) + \mu(1, b) = 1/2 + \mu(0, a)$ and finally, $\mu(2, a) = (1/2)\mu(1, a)$ and $\mu(x, a) = (1/2)\mu(x-1, a) + (1/2)\mu(x-2, a)$ for $x \geq 3$ showing that for $x \geq 2$, $\mu(x, a) = (1/6)[4 - (-1/2)^{x-2}]\mu(1, a)$. Therefore,

$$\sup\left\{\mu(r) : \mu \in \boldsymbol{\mathcal{M}}(\mathbf{X}),\ \mu_{\mathbf{X}} = \nu + \mu Q\right\} \geq \sup\left\{\mu(0, a) : \mu(0, a) \in \overline{\mathbb{R}}_+\right\} = +\infty$$

since $\mu(r) = \sum_{x \neq 1} (1/2)^{|x|}\mu(x, a) + (1/2)[\mu(1, a) + \mu(1, b)]$. This implies that Assumption (A2) in [8] is not satisfied and therefore, the approach developed there cannot be applied.

The stochastic kernel P on $\mathbf{X}$ given $\mathbf{X}$ defined by $P(x|y) = Q(x|y, a)$ for $y \in \{\Delta\} \cup \mathbb{Z} \setminus \{1\}$ and $P(2|1) = P(3|1) = P(\Delta|1) = 1/3$ satisfies Assumption A.

The probability p associated to P and given by (2) satisfies $p(x) = 0$ for $x \leq 0$. As a consequence, $\eta^\Phi(x, a) = 0$ for any $x \leq 0$ and $\Phi \in \boldsymbol{\mathcal{K}}_p$. Moreover, since η^Φ satisfies the characteristic equation, it follows that $\eta^\Phi(1, a) + \eta^\Phi(1, b) = 1/2$ and $\eta^\Phi(x, a) = (1/6)[4 - (-1/2)^{x-2}]\eta^\Phi(1, a)$ for $x \geq 2$ and $\eta^\Phi(\Delta, a) = +\infty$. Thus,

$$\begin{aligned}
\eta^\Phi(r) &= r(1, a)\eta^\Phi(1, a) + r(1, b)\eta^\Phi(1, b) + \sum_{x \geq 2} r(x, a)\eta^\Phi(x, a) \\
&= (1/2)[\eta^\Phi(1, a) + \eta^\Phi(1, b)] + \sum_{x \geq 2} (1/2)^x (1/6)[4 - (-1/2)^{x-2}]\eta^\Phi(1, a) \\
&= 1/4 + (3/10)\,\eta^\Phi(1, a)
\end{aligned}$$

and similarly,

$$\begin{aligned}
\eta^\Phi(c_1) &= (-1/18)\eta^\Phi(1, a) + \eta^\Phi(1, b) + \sum_{x \geq 2} (-1/2)^x (1/6)[4 - (-1/2)^{x-2}]\eta^\Phi(1, a) \\
&= 1/2 - \eta^\Phi(1, a)
\end{aligned}$$

where $\eta^\Phi(1, a) \in [0, 1/2]$. Clearly, we have $\eta^\Phi(r^+) < +\infty$ and $\eta^\Phi(c_1^+) < +\infty$ for any $\Phi \in \boldsymbol{\mathcal{K}}_p$ showing that Assumption (B.1) is satisfied.

Now, let π_a (respectively, π_b) be the deterministic stationary policy given by $\pi_a(\{a\}|x) = 1$ for $x \in \mathbb{Z} \cup \{\Delta\}$ (respectively, $\pi_b(\{a\}|x) = 1$ if $x \in \mathbb{Z} \cup \{\Delta\} \setminus \{1\}$ and $\pi_b(\{b\}|1) = 1$). It is easy to see that the occupation measure μ^{π_a} is given by $\mu^{\pi_a}(1, a) = 1/2$; $\mu^{\pi_a}(1, b) = 0$; $\mu^{\pi_a}(\Delta, a) = +\infty$; $\mu^{\pi_a}(x, a) = 0$ for any $x \leq 0$ and $\mu^{\pi_a}(x, a) = (1/12)[4 - (-1/2)^{x-2}]$ for $x \geq 2$ and the occupation measure μ^{π_b} satisfies $\mu^{\pi_b}(x, a) = 0$ for any $x \in \mathbb{Z}$; $\mu^{\pi_b}(1, b) = 1/2$ and $\mu^{\pi_b}(\Delta, a) = +\infty$. It follows easily $\mu^{\pi_a}(r) = \sum_{x \geq 2} (1/2)^x (1/12)[4 - (-1/2)^{x-2}] + 1/4 = 2/5$ and $\mu^{\pi_b}(r) = r(1, b)\mu^{\pi_b}(1, b) = 1/4$.

Observe also that $\mu^{\pi_a}(c_1) = -1/18 + \sum_{x\geq 2}(-1/2)^x(1/6)[4-(-1/2)^{x-2}] = 0$ and $\mu^{\pi_b}(c_1) = 1/2$. Clearly, the reward $\mathcal{J}_\nu(r,\pi)$ takes values in the interval $[\mathcal{J}_\nu(r,\pi_b), \mathcal{J}_\nu(r,\pi_a)]$ when the policy π ranges over Π and the constraint $\mathcal{J}_\nu(c_1,\pi)$ takes values in $[\mathcal{J}_\nu(c_1,\pi_a), \mathcal{J}_\nu(c_1,\pi_b)]$. Therefore, Assumption (B.2) is satisfied.

Finally, Condition (W) is obviously satisfied for this model.

Remark that for any $\alpha \in [0,1]$, the stationary randomized policy given by $\pi(\{a\}|1) = \alpha$, $\pi(\{b\}|1) = 1-\alpha$ and $\pi(\{a\}|x) = 1$ for $x \in \mathbb{Z}\setminus\{1\}$ yields $\mathcal{J}_\nu(r,\pi) = (1-\alpha)\mathcal{J}_\nu(r,\pi_b) + \alpha\mathcal{J}_\nu(r,\pi_a)$ and $\mathcal{J}_\nu(c_1,\pi) = (1-\alpha)\mathcal{J}_\nu(c_1,\pi_b) + \alpha\mathcal{J}_\nu(c_1,\pi_a)$.

The case where $\theta_1^* = 1/4$. From the previous discussion, we have

$$\sup\left\{\mathcal{J}_\nu(r,\pi) : \pi\in\Pi \text{ and } \mathcal{J}_\nu(c_1,\pi)\geq\theta_1^*\right\} = \mathcal{J}_\nu(r,\pi^*) = 13/40$$

where π^* is the stationary randomized policy given by $\pi^*(\{a\}|1) = \pi^*(\{b\}|1) = 1/2$, $\pi^*(\{a\}|x) = 1$ for $x \in \mathbb{Z}\setminus\{1\}$. Moreover,

$$\begin{aligned}
&\sup\left\{\eta^\Phi(r) : \Phi\in\boldsymbol{\mathcal{K}}_p \text{ and } \eta^\Phi(c_1)\geq\theta_1^*\right\}\\
&\qquad = \sup\{(3/10)\,\eta^\Phi(1,a) + 1/4 : \eta^\Phi(1,a)\in[0,1/2] \text{ and } (1/2-\eta^\Phi(1,a))\geq 1/4\}\\
&\qquad = 13/40.
\end{aligned}$$

Therefore, the values of the original control problem and the convex program $\boldsymbol{\mathcal{KP}}_p$ agree as claimed by Theorem 5.1 since the Slater condition holds.

Observe that the optimal value of the convex program $\boldsymbol{\mathcal{KP}}_p$ is achieved for $\eta^{\hat\Phi}(1,a) = 1/4$ where $\hat\Phi\in\boldsymbol{\mathcal{K}}_p$ is an optimal solution to the convex program $\boldsymbol{\mathcal{KP}}_p$. Since $p(1) = 1/4$, the stationary policy $\varphi_{\hat\Phi}$ induced by $\hat\Phi$ is given by $\varphi_{\hat\Phi}(\{a\}|1) = \varphi_{\hat\Phi}(\{b\}|1) = 1/2$ and $\varphi_{\hat\Phi}(\{a\}|\Delta) = \varphi_{\hat\Phi}(\{a\}|x) = 1$ for $x\in\mathbb{Z}\setminus\{1\}$. This optimal policy corresponds to π^* as determined above.

The case where $\theta_1^* = 1/2$. We have for this value of the constraint limit,

$$\sup\left\{\mathcal{J}_\nu(r,\pi) : \pi\in\Pi \text{ and } \mathcal{J}_\nu(c_1,\pi)\geq\theta_1^*\right\} = \mathcal{J}_\nu(r,\pi^*) = 1/4$$

where π^* is the stationary randomized policy given by $\pi^*(\{a\}|1) = 0$, $\pi^*(\{b\}|1) = 1$, $\pi^*(\{a\}|x) = 1$ for $x\in\mathbb{Z}\setminus\{1\}$.

However, we cannot apply the results of the present paper because in this case the Slater condition is not satisfied. Indeed, for any $\pi\in\Pi$, $\mathcal{J}_\nu(c_1,\pi)\leq 1/2$. But, the values of the original control problem and the convex program $\boldsymbol{\mathcal{KP}}_p$ still agree since

$$\begin{aligned}
&\sup\left\{\eta^\Phi(r) : \Phi\in\boldsymbol{\mathcal{K}}_p \text{ and } \eta^\Phi(c_1)\geq\theta_1^*\right\}\\
&\qquad = \sup\{(3/10)\,\eta^\Phi(1,a) + 1/4 : \eta^\Phi(1,a)\in[0,1/2] \text{ and } (1/2-\eta^\Phi(1,a))\geq 1/2\}\\
&\qquad = 1/4.
\end{aligned}$$

A Appendix

In this appendix, let m be an integer in $\mathbb{N}^*$. Consider the functions $h\in\mathfrak{M}(\mathbf{K})$ and $g_i\in\mathfrak{M}(\mathbf{K})$ for $i\in\mathbb{N}_m$. We will first present a slightly different version of a result derived by M. Schäl in [17, Theorem 1]. The only difference is that, we consider here the expected total reward criterion while in [17], Schäl deals with the conditional version of that performance criterion. We will use

it repeatedly in this paper. In this section we will also establish a technical result that is used in section 4.2 to show that in the framework of control problems with constraints, the supremum of the expected total rewards over the set of randomized policies is equal to the supremum of the expected total rewards over the set of *stationary* randomized policies.

To use Theorem 1 in [17], we need to introduce the following two sets of conditions:

($\boldsymbol{\mathcal{S}1}$) For any $x \in \mathbf{X}$, $\mathbf{A}(x)$ is compact.

($\boldsymbol{\mathcal{S}2}$) For any $x \in \mathbf{X}$ and $\Lambda \in \mathfrak{B}(\mathbf{X})$, $Q(\Lambda|x,\cdot)$ is continuous on $\mathbf{A}(x)$.

($\boldsymbol{\mathcal{S}3}$) For any $x \in \mathbf{X}$, $h(x,\cdot)$ is upper-semicontinuous on $\mathbf{A}(x)$.

($\boldsymbol{\mathcal{S}4}$) For any $x \in \mathbf{X}$, $g_i(x,\cdot)$ for $i \in \mathbb{N}_m$ are upper-semicontinuous on $\mathbf{A}(x)$.

or

($\boldsymbol{\mathcal{W}1}$) For any $x \in \mathbf{X}$, the action set $\mathbf{A}(x)$ is compact and the multifunction from $\mathbf{X}$ to $\mathbf{A}$ defined by $x \to \mathbf{A}(x)$ is upper-semicontinuous.

($\boldsymbol{\mathcal{W}2}$) For any $f \in \boldsymbol{\mathcal{C}}(\mathbf{X})$, Qf is continuous on $\mathbf{K}$.

($\boldsymbol{\mathcal{W}3}$) The function h is upper-semicontinuous on $\mathbf{K}$.

($\boldsymbol{\mathcal{W}4}$) The functions g_i for $i \in \mathbb{N}_m$ are upper-semicontinuous on $\mathbf{K}$.

Theorem A.1 *Suppose $\mu(h^+) < +\infty$ or $\mu(h^-) < +\infty$ for any $\mu \in \boldsymbol{\mathcal{O}}$ and either conditions ($\mathcal{S}1$)-($\mathcal{S}3$) or ($\mathcal{W}1$)-($\mathcal{W}3$) are satisfied. Then*

$$\sup\big\{\mu(h) : \mu \in \boldsymbol{\mathcal{O}}\big\} = \sup\big\{\mu(h) : \mu \in \boldsymbol{\mathcal{O}}_s\big\}. \tag{29}$$

Proof: The proof of this result is essentially the same as Theorem 1 in [17]. The only difference is that, we consider here the expected total reward criterion while in [17], Schäl deals with the conditional version of that performance criterion. By adapting the arguments developed in [17], we obtain easily the result. □

Proposition A.2 *Consider $\tilde{\theta} \in \mathbb{R}^m$. Assume $\sup\big\{\mu(h^+ + g_i^+) : \mu \in \boldsymbol{\mathcal{O}} \cup \{\eta^\Phi : \Phi \in \boldsymbol{\mathcal{K}}_p\}\big\} < +\infty$; $\mu(h^-) < +\infty$ and $\mu(g_i^-) < +\infty$ for $\mu \in \boldsymbol{\mathcal{O}} \cup \{\eta^\Phi : \Phi \in \boldsymbol{\mathcal{K}}_p\}$. Suppose also that Assumption A and either conditions ($\mathcal{S}1$)-($\mathcal{S}4$) or ($\mathcal{W}1$)-($\mathcal{W}4$) are satisfied. If there exists $\widetilde{\mu} \in \boldsymbol{\mathcal{O}}_s$ satisfying $\tilde{\theta}_i < \widetilde{\mu}(g_i)$ for any $i \in \mathbb{N}_m$ then*

$$\begin{aligned}\sup\big\{\mu(h) : \mu \in \boldsymbol{\mathcal{O}} \text{ and } \mu(g_i) \geq \tilde{\theta}_i \text{ for } i \in \mathbb{N}_m\big\}& \\ = \sup\big\{\mu(h) : \mu \in \boldsymbol{\mathcal{O}}_s \text{ and } \mu(g_i) \geq \tilde{\theta}_i \text{ for } i \in \mathbb{N}_m\big\}.&\end{aligned} \tag{30}$$

Proof: Let $\mathfrak{R}$ be either $\boldsymbol{\mathcal{O}}$ or $\{\eta^\Phi : \Phi \in \boldsymbol{\mathcal{K}}_p\}$. Clearly $\beta\mu_1 + (1-\beta)\mu_2 \in \mathfrak{R}$ for any μ_1, μ_2 in $\mathfrak{R}$ and $\beta \in [0,1]$. Let us define $\mathcal{C} = \bigcup_{\mu \in \mathfrak{R}} \{\theta \in \mathbb{R}^p : \mu(g_i) \geq \theta_i \text{ for } i \in \mathbb{N}_m\}$. $\mathcal{C}$ is clearly a non-empty convex subset of $\mathbb{R}^p$. Define the function $\mathcal{V}$ on $\mathcal{C}$ by

$$\mathcal{V}(\theta) := \sup\{\mu(h) : \mu \in \mathfrak{R} \text{ and } \mu(g_i) \geq \theta_i \text{ for } i \in \mathbb{N}_m\}.$$

By hypothesis, $\mathcal{V}$ takes values in $\mathbb{R}$ for any $\theta \in \mathcal{C}$. Observe that $\mathcal{V}$ is a proper concave on $\mathcal{C}$. Indeed, consider $\theta_1 = (\theta_{1,1}, \ldots, \theta_{1,m})$ and $\theta_2 = (\theta_{2,1}, \ldots, \theta_{2,m})$ in $\mathcal{C}$ and $\alpha \in [0,1]$. For any $\epsilon > 0$, there

exist $\mu_{j,\epsilon} \in \mathfrak{R}$ for $j = 1, 2$ satisfying $\mu_{j,\epsilon}(g_i) \geq \theta_{j,i}$ and $\mu_{j,\epsilon}(h) \geq \mathcal{V}(\theta_j) - \epsilon/2$ for $i \in \mathbb{N}_m$. Clearly, we have $\big(\beta\mu_{1,\epsilon} + (1-\beta)\mu_{2,\epsilon}\big)(g_i) \geq \beta\theta_{1,i} + (1-\beta)\theta_{2,i}$ for any $i \in \mathbb{N}_m$. Therefore,

$$\mathcal{V}(\beta\theta_1 + (1-\beta)\theta_2) \geq \big(\beta\mu_{1,\epsilon} + (1-\beta)\mu_{2,\epsilon}\big)(h) \geq \beta\mathcal{V}(\theta_1) + (1-\beta)\mathcal{V}(\theta_2) - \epsilon,$$

showing that $\mathcal{V}$ is a proper concave function on $\mathcal{C}$. Now, $\tilde{\theta}$ is in the interior of $\mathcal{C}$, and so $\mathcal{V}$ is continuous at $\tilde{\theta}$ by Proposition 2.17 in [3] and therefore, we can apply Proposition 2.36 in [3] to claim the existence of $\tilde{\lambda} \in \mathbb{R}^m$ such that, for all $\theta \in \mathcal{C}$,

$$\mathcal{V}(\theta) \quad \leq \quad \mathcal{V}(\tilde{\theta}) + \langle \tilde{\lambda}, \theta - \tilde{\theta} \rangle.$$

Remark that $\tilde{\lambda} \leq \mathbf{0}_m$ since $\mathcal{V}(\theta) \geq \mathcal{V}(\tilde{\theta})$ for all $\theta \leq \tilde{\theta}$. Now, fix an arbitrary $\mu \in \mathfrak{R}$. Then $(\mu(g_1), \cdots, \mu(g_p)) \in \mathcal{C}$ and so,

$$\mathcal{V}(\tilde{\theta}) \geq \mu\big(h - \langle \tilde{\lambda}, g \rangle\big) + \langle \tilde{\lambda}, \tilde{\theta} \rangle.$$

Therefore,

$$\mathcal{V}(\tilde{\theta}) \geq \sup\{\mu\big(h - \langle \tilde{\lambda}, g \rangle\big) : \mu \in \mathfrak{R}\} + \langle \tilde{\lambda}, \tilde{\theta} \rangle. \tag{31}$$

For any $\epsilon > 0$, there exists $\mu_\epsilon \in \mathfrak{R}$ with $\mu_\epsilon(g_i) \geq \tilde{\theta}_i$ for any $i \in \mathbb{N}_m$ such that $\mu_\epsilon(h) \geq \mathcal{V}(\tilde{\theta}) - \epsilon$ implying

$$\sup\{\mu\big(h - \langle \tilde{\lambda}, g \rangle\big) : \mu \in \mathfrak{R}\} + \langle \tilde{\lambda}, \tilde{\theta} \rangle \geq \mu_\epsilon(h) - \mu_\epsilon\big(\langle \tilde{\lambda}, g \rangle\big) + \langle \tilde{\lambda}, \tilde{\theta} \rangle \geq \mu_\epsilon(h) \geq \mathcal{V}(\tilde{\theta}) - \epsilon$$

since $\tilde{\lambda} \leq \mathbf{0}_m$. Together with (31), this shows

$$\sup\{\mu(h) : \mu \in \mathfrak{R} \text{ and } \mu(g_i) \geq \tilde{\theta}_i \text{ for } i \in \mathbb{N}_m\} = \sup\{\mu\big(h - \langle \tilde{\lambda}, g \rangle\big) : \mu \in \mathfrak{R}\} + \langle \tilde{\lambda}, \tilde{\theta} \rangle. \tag{32}$$

Now, we have for $\lambda \leq \mathbf{0}_m$,

$$\begin{aligned} \sup\big\{\mu\big(h - \langle \lambda, g \rangle\big) : \mu \in \mathfrak{R}\big\} + \langle \lambda, \tilde{\theta} \rangle & \\ \geq \sup\big\{\mu\big(h - \langle \lambda, g \rangle\big) : \mu \in \mathfrak{R} \text{ and } \mu(g_i) \geq \tilde{\theta}_i \text{ for } i \in \mathbb{N}_m\big\} + \langle \lambda, \tilde{\theta} \rangle & \\ \geq \sup\big\{\mu(h) : \mu \in \mathfrak{R} \text{ and } \mu(g_i) \geq \tilde{\theta}_i \text{ for } i \in \mathbb{N}_m\big\}, & \end{aligned}$$

implying

$$\begin{aligned} \inf\Big\{ \sup\big\{\mu\big(h - \langle \lambda, g \rangle\big) : \mu \in \mathfrak{R}\big\} + \langle \lambda, \tilde{\theta} \rangle : \lambda \leq \mathbf{0}_m \Big\} & \\ \geq \sup\big\{\mu(h) : \mu \in \mathfrak{R} \text{ and } \mu(g_i) \geq \tilde{\theta}_i \text{ for } i \in \mathbb{N}_m\big\}, & \end{aligned}$$

and so with (32) we obtain

$$\begin{aligned} \inf\Big\{ \sup\big\{\mu\big(h - \langle \lambda, g \rangle\big) : \mu \in \mathfrak{R}\big\} + \langle \lambda, \tilde{\theta} \rangle : \lambda \leq \mathbf{0}_m \Big\} & \\ = \sup\big\{\mu(h) : \mu \in \mathfrak{R} \text{ and } \mu(g_i) \geq \tilde{\theta}_i \text{ for } i \in \mathbb{N}_m\big\}. & \end{aligned}$$

Therefore, with $\mathfrak{R} = \boldsymbol{\mathcal{O}}$

$$\begin{aligned} \inf\Big\{ \sup\big\{\mu\big(h - \langle \lambda, g \rangle\big) : \mu \in \boldsymbol{\mathcal{O}}\big\} + \langle \lambda, \tilde{\theta} \rangle : \lambda \leq \mathbf{0}_m \Big\} & \\ = \sup\big\{\mu(h) : \mu \in \boldsymbol{\mathcal{O}} \text{ and } \mu(g_i) \geq \tilde{\theta}_i \text{ for } i \in \mathbb{N}_m\big\}, & \end{aligned} \tag{33}$$

and with $\mathfrak{R} = \{\eta^\Phi : \Phi \in \boldsymbol{\mathcal{K}}_p\}$

$$\inf\Big\{\sup\big\{\eta^\Phi\big(h - \langle\lambda, g\rangle\big) : \Phi \in \boldsymbol{\mathcal{K}}_p\big\} + \langle\lambda, \tilde{\theta}\rangle : \lambda \leq \mathbf{0}_m\Big\}$$
$$= \sup\big\{\eta^\Phi(h) : \Phi \in \boldsymbol{\mathcal{K}}_p \text{ and } \eta^\Phi(g_i) \geq \tilde{\theta}_i \text{ for } i \in \mathbb{N}_m\big\}. \quad (34)$$

Now, for $\lambda \leq \mathbf{0}_m$ we have $\sup\Big\{\eta^\Phi\Big(\big(h - \langle\lambda, g\rangle\big)^+\Big) : \Phi \in \boldsymbol{\mathcal{K}}_p\Big\} < +\infty$ by hypothesis and we obtain from Lemma 4.2 and Theorem 4.3 that

$$\sup\big\{\eta^\Phi\big(h - \langle\lambda, g\rangle\big) : \Phi \in \boldsymbol{\mathcal{K}}_p\big\} = \sup\{\mu\big(h - \langle\lambda, g\rangle\big) : \mu \in \boldsymbol{\mathcal{O}}_s\} \quad (35)$$

and also,

$$\sup\big\{\eta^\Phi(h) : \Phi \in \boldsymbol{\mathcal{K}}_p \text{ and } \eta^\Phi(g_i) \geq \tilde{\theta}_i \text{ for } i \in \mathbb{N}_m\big\}$$
$$= \sup\big\{\mu(h) : \mu \in \boldsymbol{\mathcal{O}}_s \text{ and } \mu(g_i) \geq \tilde{\theta}_i \text{ for } i \in \mathbb{N}_m\big\}. \quad (36)$$

Therefore, combining equations (34)-(36) we obtain that

$$\inf\Big\{\sup\big\{\mu\big(h - \langle\lambda, g\rangle\big) : \mu \in \boldsymbol{\mathcal{O}}_s\big\} + \langle\lambda, \tilde{\theta}\rangle : \lambda \leq \mathbf{0}_m\Big\}$$
$$= \sup\big\{\mu(h) : \mu \in \boldsymbol{\mathcal{O}}_s \text{ and } \mu(g_i) \geq \tilde{\theta}_i \text{ for } i \in \mathbb{N}_m\big\}, \quad (37)$$

Moreover, Theorem A.1 can be applied to show that

$$\sup\{\mu\big(h - \langle\lambda, g\rangle\big) : \mu \in \boldsymbol{\mathcal{O}}\} = \sup\{\mu\big(h - \langle\lambda, g\rangle\big) : \mu \in \boldsymbol{\mathcal{O}}_s\}. \quad (38)$$

Combining equations (33), (37) and (38), we obtain the result. □

References

[1] C. Aliprantis and K. Border. *Infinite dimensional analysis*. Springer, Berlin, third edition, 2006. A hitchhiker's guide.

[2] E. Altman. *Constrained Markov decision processes*. Stochastic Modeling. Chapman & Hall/CRC, Boca Raton, FL, 1999.

[3] V. Barbu and T. Precupanu. *Convexity and optimization in Banach spaces*. Springer Monographs in Mathematics. Springer, Dordrecht, fourth edition, 2012.

[4] V. Borkar. A convex analytic approach to Markov decision processes. *Probab. Theory Related Fields*, 78(4):583–602, 1988.

[5] V. Borkar. Convex analytic methods in Markov decision processes. In *Handbook of Markov decision processes*, volume 40 of *Internat. Ser. Oper. Res. Management Sci.*, pages 347–375. Kluwer Acad. Publ., Boston, MA, 2002.

[6] F. Dufour, M. Horiguchi, and A. Piunovskiy. The expected total cost criterion for Markov decision processes under constraints: a convex analytic approach. *Advances in Applied Probability*, 44(3):774–793, 2012.

[7] F. Dufour and A. Piunovskiy. Multiobjective stopping problem for discrete-time Markov processes: convex analytic approach. *J. Appl. Probab.*, 47(4):947–966, 2010.

[8] F. Dufour and A. Piunovskiy. The expected total cost criterion for Markov decision processes under constraints. *Advances in Applied Probability*, 45(3):837–859, 2013.

[9] E. Feinberg. Total reward criteria. In *Handbook of Markov decision processes*, volume 40 of *Internat. Ser. Oper. Res. Management Sci.*, pages 173–207. Kluwer Acad. Publ., Boston, MA, 2002.

[10] E. Feinberg and U. Rothblum. Splitting randomized stationary policies in total-reward Markov decision processes. *Math. Oper. Res.*, 37(1):129–153, 2012.

[11] O. Hernández-Lerma and J.B. Lasserre. *Discrete-time Markov control processes: Basic optimality criteria*, volume 30 of *Applications of Mathematics*. Springer-Verlag, New York, 1996.

[12] O. Hernández-Lerma and J.B. Lasserre. *Further topics on discrete-time Markov control processes*, volume 42 of *Applications of Mathematics*. Springer-Verlag, New York, 1999.

[13] A. Nowak. On the weak topology on a space of probability measures induced by policies. *Bull. Polish Acad. Sci. Math.*, 36(3-4):181–186 (1989), 1988.

[14] A. Piunovskiy. *Optimal control of random sequences in problems with constraints*, volume 410 of *Mathematics and its Applications*. Kluwer Academic Publishers, Dordrecht, 1997.

[15] A. Piunovskiy. *Examples in Markov decision processes*, volume 2 of *Imperial College Press Optimization Series*. Imperial College Press, London, 2013.

[16] M. Schäl. On dynamic programming: compactness of the space of policies. *Stochastic Processes Appl.*, 3(4):345–364, 1975.

[17] M. Schäl. Stationary policies in dynamic programming models under compactness assumptions. *Math. Oper. Res.*, 8(3):366–372, 1983.