Markov Decision Processes with a Constrained Stopping Time: Mathematical programming formulation

千葉大学大学院 自然科学研究科 数理物性科学専攻 堀口正之 (Masayuki HORIGUCHI)

Abstract

In this paper, the optimization problem for a stopped Markov decision process with finite states and actions is considered over stopping times τ constrained so that $\mathbb{E}\tau \leqq \alpha$ for some fixed $\alpha > 0$. The problem is solved through randomization of stopping times and mathematical programming formulation by occupation measures. Another representation, called F-representation, of randomized stopping times is given, by which the concept of Markov or stationary randomized stopping times is introduced. We treat two types of occupation measures, running and stopped, but stopped occupation measure is shown to be expressed by running one. We study the properties of the set of running occupation measures achieved by different classes of pairs of policies and randomized stopping times. Analyzing the equivalent mathematical programming problem formulated by running occupation measures corresponding with stationary policies and stationary randomized stopping times, we prove the existence of an optimal constrained pair of stationary policy and stopping time requiring randomization in at most one state .

Key words: Stopped Markov decision process, constrained stopping time, mathematical programming formulation.

1 Introduction

A constrained optimal stopping problem is originated by Nachman [17] and Kennedy [15], in which a Lagrangian approach was used to reduce the problem to an unconstrained stopping problem of a conventional type and the constrained optimal stopping time is characterized. Also, a constrained Markov decision process has been studied by many authors (cf. [1, 2, 3, 7, 10]). For the case of fixed entry time, Altman [2] has formed an equivalent infinite Linear Programming for the total cost criteria and by analyzing the corresponding LP formulation has shown that there exists an optimal constrained stationary policy. On the other hand, a combined model of the Markov decision process and stopping problem, called a stopped decision process, has been considered by Furukawa and Iwamoto [8] in that the existence of an optimal pair of policy and stopping time associated with some optimality criterions is discussed. Hordijk [9] has considered this model from a standpoint of potential theory. Also, the general utility-treatment for stopped decision processes has been studied by Kadota et al[13, 14]. Horiguchi et al.[11] has considered the optimization problem for the stopped decision process over stopping times τ constrained so that $\mathbb{E}\tau \leqq \alpha$ for some fixed $\alpha > 0$, which is analyzed by a Lagrange multiplier. In this paper, we develop mathematical programming methods in the framework similar to [11].

The problem is solved through randomization of stopping times and mathematical programming formulation by occupation measures. Another representation, called F-representation, of randomized stopping times is given, by which the concept of Markov or stationary randomized stopping times is introduced. We treat two types of occupation measures, running and stopped, but stopped occupation measure is shown to be expressed by running one. We study the properties of the set of running occupation measures achieved by different classes of pairs of policies and randomized stopping times. Analyzing the equivalent mathematical programming problem formulated by running occupation measures corresponding with stationary policies and stationary randomized stopping times, we prove the existence of an optimal constrained pair of stationary policy and stopping time requiring randomization in at most one state.

In the reminder of this section, we shall establish the notation that will be used throughout this paper and define the optimization problem. Also, an optimal constrained pair of policy and randomized stopping time is defined.

Let S and A be finite sets denoted by $S = \{1, 2, \ldots N\}$ and $A = \{1, 2, \ldots, K\}$. The stopped Markov decision model consists of five objects:

$$(S, A, \{p_{ij}(a) : i, j \in S, a \in A\}, c, r)$$

where S and A denote the state and action spaces respectively and $c = c(i, a)$ is a running cost function on $S \times A$ and $r = r(i)$ is a terminal reward function on S when selecting "stop" in state i, and $\{p_{ij}(a)\}$ is the law of motion i.e., for each $(i, a) \in S \times A, p_{ij}(a) \geqq 0$ and $\sum_{j \in S} p_{ij}(a) = 1$. When the system is in state $i \in S$, if we select "stop" the process terminates with the terminal reward $r(i)$. If we select "continue" and take $a \in A$, we move to a new state $j \in S$ selected according to the probability distribution $p_{i\cdot}(a)$ and the cost $c(i, a)$ is incurred. This process is repeated from the new state $j \in S$. Let x_t, a_t be the state and action at time t and $h_t = (x_1, a_1, \ldots, x_t) \in (S \times A)^{t-1} \times S$ the history up to time $t (t \geqq 1)$. A policy for a controlling the system is a sequence $\pi = (\pi_1, \pi_2, \ldots)$ such that, for each $t \geqq 1$, π_t is a conditional probability measure on A given history h_t with $\pi_t(A|(x_1, a_1, \ldots, x_t)) = 1$ for each $(x_1, a_1, \ldots, x_t) \in (S \times A)^{t-1} \times S$. Let Π denotes the set of all policies. A policy $\pi = (\pi_1, \pi_2, \ldots)$ is a Markov policy if π_t is a function of only x_t, i.e., $\pi_t(\cdot|x_1, a_1, \ldots, x_t) = \pi_t(\cdot|x_t)$ for all $(x_1, a_1, \ldots, x_t) \in (S \times A)^{t-1} \times S$. A Markov policy $\pi = (\pi_1, \pi_2, \ldots)$ is stationary if there exists a conditional probability on A, $w(\cdot|i)$, given $i \in S$ such that $\pi_t(\cdot|x_t) = w(\cdot|x_t)$ for all $x_t \in S$ and $t \geqq 1$, and denoted by $w^\infty = (w, w, \ldots)$, or simply by w. A stationary policy w is called deterministic if there exists a map $h : S \to A$ with $w(h(i)|i) = 1$ for all $i \in S$ and such a policy is identified by h. The sets of all Markov, stationary and deterministic policies will be denoted by Π_M, Π_S and Π_D respectively. Note that $\Pi_D \subset \Pi_S \subset \Pi_M \subset \Pi$. The sample spaces is the product space $\Omega = (S \times A)^\infty$. Let X_t, Δ_t be random quantities such that $X_t(\omega) = x_t$ and $\Delta_t(\omega) = a_t$ for all $\omega = (x_1, a_1, x_2, a_2, \ldots) \in \Omega$. For any given policy $\pi \in \Pi$ and initial distribution β on S we can specify the probability measure $\mathbb{P}_\beta^\pi$ on Ω in a usual way.

Let $H_t = (X_1, \Delta_1, \ldots, X_t)$. We denote by $\mathcal{B}(H_t)$ the σ-field induced by H_t. Let $\mathscr{F}_t = \mathcal{B}(H_t), (t \geqq 1)$ and $\mathscr{F}_\infty$ be the smallest σ-field containing each $\mathscr{F}_t, t \geqq 1$. Let $\overline{N} = \{1, 2, \ldots\} \cup \{\infty\}$. We call a map $\tau : \Omega \to \overline{N}$ a stopping time w.r.t. the filtration $\mathscr{F} = \{\mathscr{F}_t, t \in \overline{N}\}$ if $\{\tau = t\} \in \mathscr{F}_t$ for all $t \in \overline{N}$. In order to solve our problem described in the sequel, we need to introduce randomized stopping time (cf. [6, 15]). To this purpose, enlarging Ω to $\overline{\Omega} := \Omega \times [0, 1]$, we can embed $(\Omega, \mathscr{F}_\infty)$ to $(\overline{\Omega}, \mathscr{F}_\infty \times \mathbb{B}_1)$, where $\mathbb{B}_1$ is Borel subsets of $[0, 1]$. For a filtration $\mathscr{F}^* = \{\mathscr{F}_t^*, t \in \overline{N}\}$ with $\mathscr{F}_t^* = \mathscr{F}_t \times \mathbb{B}_1$ we can assume without loss of generality that for each $t \in \overline{N}$

$$\mathscr{F}_t \subset \mathscr{F}_t^*. \tag{1.1}$$

We call a map $\overline{\tau} : \overline{\Omega} \to \overline{N}$ a randomized stopping time (hereafter called RST) w.r.t. $\mathscr{F}^*$ if $\{\overline{\tau} = t\} \in \mathscr{F}_t^*$ for each $t \in \overline{N}$. For simplicity, the overline of RST $\overline{\tau}$ will be omitted and written by τ with some abuse of notation. The class of RSTs w.r.t. $\mathscr{F}^*$ will be denoted by $\mathcal{S}$. For each initial distribution β and each policy $\pi \in \Pi$, we denote the probability measure on $\overline{\Omega}$ by $\overline{\mathbb{P}}_\beta^\pi$, where $\overline{\mathbb{P}}_\beta^\pi = \mathbb{P}_\beta^\pi \times \lambda$ and λ is Lebesgue measure on $\mathbb{B}_1$. For any $\alpha > 0$ and initial distribution β on S, let

$$\Delta(\alpha, \beta) := \{(\pi, \tau) \in \Pi \times \mathcal{S} \mid \overline{\mathbb{E}}_\beta^\pi \tau \leqq \alpha\} \tag{1.2}$$

where $\overline{\mathbb{E}}_\beta^\pi$ is the expectation w.r.t. $\overline{\mathbb{P}}_\beta^\pi$. The pair belonging to $\Delta(\alpha, \beta)$ will be called a constrained one. In this paper, we will consider the constrained optimization problem(**COP**):

$$\textbf{COP}: \text{ Maximize } J(\beta, \pi, \tau) := \overline{\mathbb{E}}_\beta^\pi \left[\sum_{n=1}^{\tau-1} c(X_n, \Delta_n) + r(X_\tau) \right]$$

$$\text{subject to } (\pi, \tau) \in \Delta(\alpha, \beta)$$

The constrained pair $(\pi^*, \tau^*) \in \Delta(\alpha, \beta)$ is called optimal if

$$J(\beta, \pi, \tau) \leqq J(\beta, \pi^*, \tau^*) \quad \text{for all} \quad (\pi, \tau) \in \Delta(\alpha, \beta).$$

In Section 2, F-representation of RST is considered and after defining a Markov RST, it is shown that the class of pairs of Markov policies and Markov RSTs is sufficient for our problem. In Section 3, the running and stopped occupation measures are introduced, by which **COP** is reduced equivalently to mathematical programming. In Section 4, studying the properties of the set of running occupation measure we can prove the existence of an optimal constrained pair of stationary policy and stopping time requiring randomization in at most one state.

2 Preliminaries

In this section, F-representation of RSTs given by Irle[12](cf. [4]) will be extended to the case of the decision process considered in this paper by which Markov or stationary RSTs are defined.

For any RST $\tau \in \mathcal{S}$ and $t \in \overline{N}$, let $g_t(\omega) := \lambda\left(\{\tau = t\}_\omega\right), \omega \in \Omega$, where $\{\tau = t\}_\omega$ is the ω-section defined by $\{\tau = t\}_\omega = \{x \in [0,1] | (\omega, x) \in \{\tau = t\}\}$. Note that g_t is $\mathscr{F}_t$-measurable for $t \geq 1$. From this g_t $(t \in \overline{N})$, we define the set $f = (f_t)_{t \in \overline{N}}$ as follows:

$$f_t := \frac{g_t}{1 - \sum_{k=1}^{t-1} g_k}, \quad t \in \overline{N} \tag{2.1}$$

where if the denominator is 0 in (2.1) let $f_t = 1$. Let $F = \{a = (a_j)_{j \in \overline{N}} : 0 \leqq a_j \leqq 1, a_\infty = 1$, and if $a_j = 1 \Rightarrow a_i = 1$ for $i > j\}$. Then we have the following lemmas.

Lemma 2.1.
(i) *$f : \Omega \to F$ and for each $t \in \overline{N}$ f_t is $\mathscr{F}_t$-measurable.*
(ii) *For any initial distribution β and pair $(\pi, \tau) \in \Pi \times \mathcal{S}$ and $t \in \overline{N}$,*

$$f_t = \frac{\overline{\mathbb{P}}^\pi_\beta(\tau = t | H_t)}{\overline{\mathbb{P}}^\pi_\beta(\tau \geqq t | H_t)}, \quad \mathbb{P}^\pi_\beta\text{-a.s.} \tag{2.2}$$

(iii) *For any initial distribution β and pair $(\pi, \tau) \in \Pi \times \mathcal{S}$*

$$\begin{aligned} &\overline{\mathbb{E}}^\pi_\beta \left[\sum_{t=1}^{\tau-1} c(X_t, \Delta_t) + r(X_\tau) \right] \\ =&\sum_{t=1}^{\infty} \left(\mathbb{E}^\pi_\beta \left((1 - f_1) \cdots (1 - f_{t-1}) f_t \left(\sum_{k=1}^{t-1} c(X_k, \Delta_k) + r(X_t) \right) \right) \right) \end{aligned} \tag{2.3}$$

Proof. From Fubini's theorem (i) clearly follows.
For (ii), it suffices to prove

$$\lambda(\{\tau = t\}_\omega) = \overline{\mathbb{P}}^\pi_\beta(\{\tau = t\} | \mathscr{F}_t) \quad \mathbb{P}^\pi_\beta\text{-a.s.} \tag{2.4}$$

For any subset $D \in \mathscr{F}_t$, by Fubini's theorem

$$\int_{D \times [0,1]} 1_{\{\tau = t\}} d\overline{\mathbb{P}}^\pi_\beta = \int_D \lambda(\{\tau = t\}_\omega) d\mathbb{P}^\pi_\beta. \tag{2.5}$$

On the other hand, it holds that

$$\int_{D \times [0,1]} 1_{\{\tau = t\}} d\overline{\mathbb{P}}^\pi_\beta = \int_{D \times [0,1]} \overline{\mathbb{P}}^\pi_\beta(\{\tau = t\} | \mathscr{F}_t) d\overline{\mathbb{P}}^\pi_\beta = \int_D \overline{\mathbb{P}}^\pi_\beta(\{\tau = t\} | \mathscr{F}_t) d\mathbb{P}^\pi_\beta$$

which implies (2.4), together with (2.5).

By the definition (2.1), we get

$$1 - f_t = \frac{\lambda(\{\tau \geqq t+1\}_\omega)}{\lambda(\{\tau \geqq t\}_\omega)},$$

so that

$$\begin{aligned} g_t = \lambda(\{\tau = t\}_\omega\}) &= \frac{\lambda(\{\tau = t\}_\omega)}{\lambda(\{\tau \geqq t\}_\omega)} \cdot \frac{\lambda(\{\tau \geqq t\}_\omega)}{\lambda(\{\tau \geqq t-1\}_\omega)} \cdots \frac{\lambda(\{\tau \geqq 2\}_\omega)}{\lambda(\{\tau \geqq 1\}_\omega)} \\ &= f_t(1-f_{t-1})\cdots(1-f_2)(1-f_1), \quad (t \geqq 1). \end{aligned} \tag{2.6}$$

From (2.6),

$$\begin{aligned} \overline{\mathbb{E}}^\pi_\beta \left[\sum_{t=1}^{\tau-1} c(X_t, \Delta_t) r(X_\tau)\right] &= \overline{\mathbb{E}}^\pi_\beta \left[\sum_{t=1}^{\infty} 1_{\{\tau=t\}} \left(\sum_{k=1}^{t-1} c(X_k, \Delta_k) + r(X_t)\right)\right] \\ &= \mathbb{E}^\pi_\beta \left[\sum_{t=1}^{\infty} g_t \left(\sum_{k=1}^{t-1} c(X_k, \Delta_k) + r(X_t)\right)\right], \quad \text{by Fubini's theorem} \\ &= \sum_{t=1}^{\infty} \mathbb{E}^\pi_\beta \left[(1-f_1)\cdots(1-f_{t-1})f_t \left(\sum_{k=1}^{t-1} c(X_k, \Delta_k) + r(X_t)\right)\right], \end{aligned}$$

which completes the proof of (iii).▮

The set $f = (f_t)_{t \in \overline{N}}$ constructed from $\tau \in \mathcal{S}$ is called F-representation of τ, denoted by $f^\tau = (f^\tau_t)_{t \in \overline{N}}$.

Let $f = (f_t)_{t \in \overline{N}}$ be any function $f : \Omega \to F$ such that for each $t \in \overline{N}$ f_t is $\mathscr{F}_t$-measurable. From this f, we define $\tau^f : \Omega \times [0,1] \to \overline{N}$ by

$$\tau^f(\omega, x) := \begin{cases} t & \text{for } x \in \left[\sum_{k=1}^{t-1} \overline{g}_k(\omega), \sum_{k=1}^{t} \overline{g}_k(\omega)\right) \\ \infty & \text{for } x \in [\sum_{k=1}^{\infty} \overline{g}_k(\omega), 1]. \end{cases} \tag{2.7}$$

where

$$\overline{g}_t := (1-f_1)\cdots(1-f_{t-1})f_t \quad (t \geqq 1) \tag{2.8}$$

Then, we have:

Lemma 2.2.

(i) *τ^f is a RST w.r.t. $\mathscr{F}^* = \{\mathscr{F}^*_t, t \in \overline{N}\}$.*

(ii) *τ^f satisfies (ii) and (iii) of Lemma 2.1.*

Proof. For $\mathscr{F}_t$-measurable functions $\overline{g}_t$ $(t \geqq 1)$, we define the set of functions on $\Omega \times [0,1]$ as follows:

$$G_t(\omega, x) := \sum_{k=1}^{t-1} \overline{g}_k(\omega), \quad F(\omega, x) := x, \quad G'_t(\omega, x) := \sum_{k=1}^{t} \overline{g}_k(\omega) \ (t \geqq 1)$$

Since $\{\tau^f = t\} = \{G_t \leqq F\} \cap \{F < G'_t\}$ and $\{G_t \leqq F\} \in \mathscr{F}^*_{t-1}$ and $\{F < G'_t\} \in \mathscr{F}^*_t$, we have $\{\tau^f = t\} \in \mathscr{F}^*_t$, $(t \geqq 1)$, which proves (i).

From (2.7), $\lambda\left(\{\tau^f = n\}_\omega\right) = \overline{g}_n(\omega)$ and by (2.8) $f_t = \overline{g}_t/(1 - \sum_{n=1}^{t-1} \overline{g}_n)$, so that we have $f_t(\omega) = \lambda\left(\{\tau^f = t\}_\omega\right)/\lambda\left(\{\tau^f \geqq t\}_\omega\right)$. Hence (ii) follows similarly as the proof of (ii) and (iii) of Lemma 2.1. ▮

Note that Lemma 2.1 and 2.2 show there is one-to-one correspondence between $\mathcal{S}$ and the set of F-representations $f = (f_t)_{t \in \overline{N}}$. Using this fact, we define several types of RSTs. Let $\tau \in \mathcal{S}$. For the corresponding F-representation $f^\tau = (f_t^\tau)_{t \in \overline{N}}$, by Lemma 2.1, f_t^τ is $\mathscr{F}_t$-measurable $(t \geqq 1)$. So, f_t^τ is a function of $H_t = (X_1, \Delta_1, \ldots, X_t)$.

Definition 2.1. *If f_t^τ is depending only on X_t, that is, $f_t^\tau(H_t) = f_t^\tau(X_t)$ for all $t \geqq 1$, the RST τ is called Markov. A markov RST is called stationary if there exists a function $\delta : S \to [0,1]$ such that $f_t^\tau(X_t) = \delta(X_t)$ for all $t \geqq 1$, and denoted by δ^∞. When $\delta(i) \in \{0,1\}$ for all $i \in S$, the stationary RST δ^∞ is called deterministic.*

We denote the sets of all Markov RSTs, all stationary RSTs and all deterministic RSTs by $\mathcal{S}_M, \mathcal{S}_S$ and $\mathcal{S}_D$ respectively.

In the following, we say that the set of $\Pi_M \times \mathcal{S}_M$ is a sufficient class to our optimization problem.

Lemma 2.3. *For any pair $(\pi, \tau) \in \Pi \times \mathcal{S}$, there exist a pair $(v, \sigma) \in \Pi_M \times \mathcal{S}_M$ such that*

$$\overline{\mathbb{P}}_\beta^\pi (X_t = i, \Delta_t = a, \tau > t) = \overline{\mathbb{P}}_\beta^v (X_t = i, \Delta_t = a, \sigma > t), \quad \text{for } i \in S, a \in A. \tag{2.9}$$

Proof. We define a Markov policy $v = (v_1, v_2, \ldots)$ by

$$v_t(a|i) := \frac{\overline{\mathbb{P}}_\beta^\pi (X_t = i, \Delta_t = a, \tau > t)}{\overline{\mathbb{P}}_\beta^\pi (X_t = i, \tau > t)} \text{ for } t \geqq 1 \text{ and } i \in S, a \in A, \tag{2.10}$$

where if the denominator is zero, we let $v_t(\cdot|i)$ be an arbitrary probability measure over A. We also define the set $f = (f_t)_{t \in \overline{N}}$ by

$$1 - f_t(i) := \frac{\overline{\mathbb{P}}_\beta^\pi (X_t = i, \tau > t)}{\overline{\mathbb{P}}_\beta^\pi (X_t = i, \tau \geqq t)} \text{ for } t \geqq 1 \text{ and } i \in S, \tag{2.11}$$

where if the denominator is zero, we set $f_t(i) = 1$. Then, clearly $f = (f_t)_{t \in \overline{N}} : \Omega \to F$. Thus, we can define $\sigma \in \mathcal{S}_M$ from f through (2.11). Now we show that the pair (v, σ) satisfies (2.9) by induction. From (2.10) and (2.11),

$$\begin{aligned}\overline{\mathbb{P}}_\beta^\pi (X_1 = i, \Delta_1 = a, \tau > 1) &= v_1(a|i) \overline{\mathbb{P}}_\beta^\pi (X_1 = i, \tau > t) \\ &= v_1(a|i)(1 - f_1(i)) \overline{\mathbb{P}}_\beta^\pi (X_1 = i, \tau \geqq 1)\end{aligned}$$

Since $\overline{\mathbb{P}}_\beta^\pi(X_1, \tau \geqq 1) = \beta(i)$, we get

$$\overline{\mathbb{P}}_\beta^\pi(X_1 = i, \Delta_1 = a, \tau > 1) = \beta(i) v_1(a|i)(1 - f_1(i)) = \overline{\mathbb{P}}_\beta^v (X_1 = i, \Delta_1 = a, \sigma > 1),$$

which shows that (2.9) holds for $t = 1$.

Assume (2.9) holds for t $(t \geqq 1)$. Combining (2.10) and (2.11) with $t + 1$, we get

$$\begin{aligned}&\overline{\mathbb{P}}_\beta^\pi (X_{t+1} = i, \Delta_{t+1} = a, \tau > t + 1) \\ =&\overline{\mathbb{P}}_\beta^\pi (X_{t+1} = i, \tau \geqq t + 1)(1 - f_{t+1}(i)) v_{t+1}(a|i) \\ =&\overline{\mathbb{E}}_\beta^\pi \left[1_{\{X_{t+1} = i\}} 1_{\{\tau > t\}}\right] (1 - f_{t+1}(i)) v_{t+1}(a|i) \\ =& \sum_{j \in S, a' \in A} \overline{\mathbb{P}}_\beta^v (X_t = j, \Delta_t = a', \sigma > t) p_{ji}(a')(1 - f_{t+1}(i)) v_{t+1}(a|i), \\ &\qquad\qquad\qquad \text{from the hypothesis of induction,} \\ =&\overline{\mathbb{P}}_\beta^v (X_{t+1} = j, \Delta_{t+1} = a, \sigma > t + 1).\end{aligned}$$

This completes the proof of Lemma.∎

3 Running and stopped occupation measures

We introduce, in this section, two types of occupation measures and consider the properties of them. Also, we formulate the Mathematical Programming problem which is proved to be equivalent to **COP**.

Definition 3.1. *For any initial distribution β and a pair (π, τ) with $\mathbb{E}_\beta^\pi[\tau] < \infty$, we define the measure $x(\beta, \pi, \tau)$ on $S \times A$, called running occupation measure, by*

$$x\left(\beta, \pi, \tau; i, a\right) := \sum_{t=1}^{\infty} \mathbb{P}_\beta^\pi \left(X_t = i, \Delta_t = a, \tau > t\right) \quad \textit{for} \quad i \in S, a \in A. \tag{3.1}$$

Definition 3.2. *For any initial distribution β and a pair (π, τ) with $\mathbb{E}_\beta^\pi[\tau] < \infty$, we define the measure $y(\beta, \pi, \tau)$ on $S \times A$, called the stopped occupation measure, by*

$$y\left(\beta, \pi, \tau; i, a\right) := \sum_{t=1}^{\infty} \mathbb{P}_\beta^\pi \left(X_t = i, \Delta_t = a, \tau = t\right) \quad \textit{for} \quad i \in S, a \in A. \tag{3.2}$$

The state running and stopped occupation measures will be defined by

$$x(\beta, \pi, \tau; i) := \sum_{a \in A} x(\beta, \pi, \tau; i, a) \text{ and } y(\beta, \pi, \tau; i) := \sum_{a \in A} y(\beta, \pi, \tau; i, a) \text{ for all } i \in S.$$

Then, in the following lemma, the state stopped occupation measure is proved to be represented by the running one.

Lemma 3.1. *For any β and pair $(\pi, \tau) \in \Pi \times \mathcal{S}$ with $\mathbb{E}_\beta^\pi[\tau] < \infty$ we have the following:*

(i) $x\left(\beta, \pi, \tau; i\right) < \infty$ *and* $y\left(\beta, \pi, \tau; i\right) < \infty$ *for all* $i \in S$.

(ii) $\mathbb{E}_\beta^\pi[\tau] = \sum_{i \in S} x(\beta, \pi, \tau; i) + 1$.

(iii) $y(\beta, \pi, \tau; i) = \beta(i) + \sum_{j \in S, a \in A} x(\beta, \pi, \tau; j, a) p_{ji}(a) - x(\beta, \pi, \tau; i)$.

Proof. Observing that $\mathbb{E}_\beta^\pi[\tau] = \sum_{t=1}^{\infty} \mathbb{P}_\beta^\pi(\tau \geqq t)$, it holds that $\sum_{i \in S} x\left(\beta, \pi, \tau; i\right) = \sum_{t=1}^{\infty} \mathbb{P}_\beta^\pi(\tau > t) = \mathbb{E}_\beta^\pi[\tau] - 1$ for all $i \in S$, which from the $\mathbb{E}_\beta^\pi[\tau] < \infty$ proves the first part of (i) and (ii). Also, $y\left(\beta, \pi, \tau; i\right) < \infty$ follows obviously.
For (iii), we have

$$\begin{aligned}
y(\beta, \pi, \tau; i) &= \sum_{t=1}^{\infty} \mathbb{P}_\beta^\pi(X_t = i, \tau = t) = \sum_{t=1}^{\infty} \mathbb{P}_\beta^\pi(X_t = i, \tau \geqq t) - \sum_{t=1}^{\infty} \mathbb{P}_\beta^\pi(X_t = i, \tau > t) \\
&= \beta(i) + \sum_{j \in S, a \in A} \left(\sum_{t=1}^{\infty} \mathbb{P}_\beta^\pi(X_t = j, \Delta_t = a, \tau > t) \right) p_{ji}(a) - x(\beta, \pi, \tau; i) \\
&= \beta(i) + \sum_{j \in S, a \in A} x(\beta, \pi, \tau; j, a) p_{ji}(a) - x(\beta, \pi, \tau; i)
\end{aligned}$$

as required. ▮

For any $\delta : S \to [0, 1]$ and conditional distribution $w(\cdot | i)$ on A given $i \in S$, we define by $P^\delta(w)$ the $N \times N$ matrix where (i, j)th element is $\sum_{a \in A} p_{ij}(a) w(a|i)(1 - \delta(j)) := p_{ij}(w)(1 - \delta(j))$ or simply $(P^\delta(w))_{ij}$. Let $\mathbb{R}^N$ be the set of real N-dimensional row vectors. For any initial distribution β and $(\pi, \tau) \in \Pi \times \mathcal{S}$, the row vector $x(\beta, \pi, \tau) \in \mathbb{R}^N$ is defined by

$$x(\beta, \pi, \tau) := (x(\beta, \pi, \tau; 1), \ldots, x(\beta, \pi, \tau; N)).$$

If the distribution β on S is degenerate as $i \in S$, it is simply denoted by i.

Lemma 3.2. *Let $(w, \tau) \in \Pi_S \times \mathcal{S}_S$ with $\overline{\mathbb{E}}_i^w(\tau) < \infty$ for all $i \in S$. Then the state running occupation measure $x(\beta, w, \tau)$ is the unique solution to*

$$x = \beta(1-\delta) + xP^{\delta}(w), \ \ x \in \mathbb{R}^N \tag{3.3}$$

where $\beta(1-\delta)$ is in $\mathbb{R}^n$ whose i-th component is $\beta(i)(1-\delta(i))$ and $\delta := f^{\tau} : S \to [0,1]$ is F-representation of τ.

Proof. By virtue of stationary of τ and w, we have

$$\begin{aligned}
&\overline{\mathbb{P}}_{\beta}^{\pi}(X_t = i, \tau > t) \\
=&\overline{\mathbb{E}}_{\beta}^{w}\left[\overline{\mathbb{E}}_{\beta}^{w}\left[1_{\{X_t=i\}}1_{\{\tau>t\}}|H_t\right]\right] \\
=&\overline{\mathbb{E}}_{\beta}^{w}\left[(1-\delta(X_1))(1-\delta(X_2))\cdots(1-\delta(X_{t-1}))(P^{\delta}(w))_{X_{t-1}i}\right] \\
=&\overline{\mathbb{E}}_{\beta}^{w}\left[(1-\delta(X_1))\cdots(1-\delta(X_{t-2}))\overline{\mathbb{E}}_{\beta}^{w}\left[(1-\delta(X_{t-1}))(P^{\delta}(w))_{X_{t-1}i}|H_{t-2}\right]\right] \\
=&\overline{\mathbb{E}}_{\beta}^{w}\left[(1-\delta(X_1))\cdots(1-\delta(X_{t-2}))\left(P^{\delta}(w)\right)^2_{X_{t-2}i}\right] \\
=&\cdots \\
=&\overline{\mathbb{E}}_{\beta}^{w}\left[(1-\delta(X_1)\left(P^{\delta}(w)\right)^{t-1}_{X_1 i}\right] \\
=&\sum_{j\in S}\beta(j)(1-\delta(j))\left(P^{\delta}(w)\right)^{t-1}_{ji}.
\end{aligned}$$

Thus,

$$\begin{aligned}
x(\beta,\pi,\tau;i) = \sum_{t=1}^{\infty}\overline{\mathbb{P}}_{\beta}^{w}(X_t = i, \tau > t) &= \sum_{t=1}^{\infty}\left(\sum_{k\in S}\beta(k)(1-\delta(k))\left(P^{\delta}(w)\right)^{t-1}_{ki}\right) \\
&= \beta(i)(1-\delta(i)) + \sum_{j\in S}x(\beta, w, \tau; j)\left(P^{\delta}(w)\right)_{ji},
\end{aligned}$$

which shows $x(\beta, \pi, \tau)$ is a solution of (3.3).

To prove the uniqueness, let x, z be the solutions of (3.3), that is, x, z satisfy that

$$x = \beta(1-\delta) + xP^{\delta}(w), \quad z = \beta(1-\delta) + zP^{\delta}(w).$$

Then we have $x - z = (x - z)P^{\delta}(w)$. By iterating this equation, we get

$$x - z = (x - z)\left(P^{\delta}(w)\right)^t \quad (t \geqq 1). \tag{3.4}$$

By Lemma 3.2 (ii) and $\overline{\mathbb{E}}[\tau] < \infty$, we have

$$\overline{\mathbb{P}}_i^{w}(X_t = j, \tau > t) = (1-\delta(i))\left(P^{\delta}(w)\right)^t_{ij} \to 0 \ \text{ as } \ t \to \infty.$$

Noting $x(i) = z(i) = 0$ for $\delta(i) = 1$, we get $P^{\delta}(w) \to \infty$ from (3.4), which implies $x = z$. ▮

Next, we present that the objective function $J(\beta, \pi, \tau)$ of **COP** is written by running and stopped occupation measures.

Lemma 3.3. *For* $(\pi, \tau) \in \Pi \times \mathcal{S}$ *with* $\overline{\mathbb{E}}_\beta^\pi[\tau] < \infty$, *we have*

$$J(\beta, \pi, \tau) = \sum_{i \in S, a \in A} c(i, a) x(\beta, \pi, \tau; i, a) + \sum_{i \in S} r(i) y(\beta, \pi, \tau; i). \tag{3.5}$$

Proof. By the generalized convergence theorem (Royden [18] p.232) we get

$$\begin{aligned}
\overline{\mathbb{E}}_\beta^\pi \left[\sum_{t=1}^{\tau-1} c(X_t, \Delta_t) \right] &= \overline{\mathbb{E}}_\beta^\pi \left[\sum_{t=1}^{\infty} c(X_t, \Delta_t) 1_{\{\tau > t\}} \right] \\
&= \sum_{t=1}^{\infty} \left(\sum_{i \in S, a \in A} c(i, a) \overline{\mathbb{P}}_\beta^\pi (X_t = i, \Delta_t = a, \tau > t) \right) \\
&= \sum_{i \in S, a \in A} c(i, a) \left(\sum_{t=1}^{\infty} \overline{\mathbb{P}}_\beta^\pi (X_t = i, \Delta_t = a, \tau > t) \right) \\
&= \sum_{i \in S, a \in A} c(i, a) x(\beta, \pi, \tau; i, a).
\end{aligned}$$

Also, using a similar argument for stopped occupation measure we get

$$\begin{aligned}
\overline{\mathbb{E}}_\beta^\pi [r(X_\tau)] = \overline{\mathbb{E}}_\beta^\pi \left[\sum_{t=1}^{\infty} r(X_\tau) 1_{\{\tau = t\}} \right] &= \sum_{t=1}^{\infty} \left(\sum_{i \in S} r(i) \overline{\mathbb{P}}_\beta^\pi (X_t = i, \tau = t) \right) \\
&= \sum_{i \in S, a \in A} r(i) \left(\sum_{t=1}^{\infty} \overline{\mathbb{P}}_\beta^\pi (X_t = i, \tau = t) \right) \\
&= \sum_{i \in S, a \in A} r(i) y(\beta, \pi, \tau; i).
\end{aligned}$$

Hence, the lemma follows. ▮

Let $\mathbb{R}^{N \times K}$ be the set of real $N \times K$ matrices. For any subset $U \subset \Pi \times \mathcal{S}$, let

$$\mathbb{X}_{\{\leqq\}\alpha}(U) = \{ x(\beta, \pi, \tau; i, a)_{i \in S, a \in A} : (\pi, \tau) \in U,\ \overline{\mathbb{E}}_\beta^\pi[\tau] \leqq \alpha \}. \tag{3.6}$$

Note that $\mathbb{X}_{\{\leqq\}\alpha}(U) \subset \mathbb{R}^{N \times K}$. We introduce the Mathematical Programming(**MP(I)**) as follows.

$$\begin{aligned}
\textbf{MP(I)}\text{: Maximize } & \sum_{i \in S, a \in A} c(i, a) x(i, a) + \sum_{i \in S} r(i) y(i) \\
\text{subject to } & x \in \mathbb{X}_{\{\leqq\}\alpha}(\Pi \times \mathcal{S}),\ y \in \mathbb{R}^N \text{and} \\
& y(i) = \beta(i) + \sum_{j \in S, a \in A} x(j, a) p_{ji}(a) - x(i),\quad i \in S
\end{aligned}$$

Then, we have the following theorem whose proof follows easily from Lemma 3.3.

Theorem 3.1. COP *is equivalent to* **MP(I)**, *i.e., a pair* (π^*, τ^*) *is optimal for* **COP** *if and only if the corresponding* $\{x(\beta, \pi^*, \tau^*; i, a)\} \in \mathbb{X}_{\{\leqq\}\alpha}(\Pi \times \mathcal{S})$ *is optimal for* **MP(I)**.

4 Mathematical programming and optimal pair

In this section, we present another Mathematical programming formulation by which **COP** is explicitly solved.

For any $U \subset \Pi \times \mathcal{S}$, let $\mathbb{X}^{\beta}_{\{=\}\alpha}(U)$ be the set of $\mathbb{X}^{\beta}_{\{\leqq\}\alpha}(U)$ which is defined by replacing $\overline{\mathbb{E}}^{\pi}_{\beta}[\tau] \leqq \alpha$ with $\overline{\mathbb{E}}^{\pi}_{\beta}[\tau] = \alpha$ in (3.6).

Theorem 4.1.

$$\mathbb{X}^{\beta}_{\{\leqq\}\alpha}(\Pi \times \mathcal{S}) = \mathbb{X}^{\beta}_{\{\leqq\}\alpha}(\Pi_M \times \mathcal{S}_M) = \mathbb{X}^{\beta}_{\{\leqq\}\alpha}(\Pi_S \times \mathcal{S}_S), \quad \textit{and} \tag{4.1}$$

$$\mathbb{X}^{\beta}_{\{=\}\alpha}(\Pi \times \mathcal{S}) = \mathbb{X}^{\beta}_{\{=\}\alpha}(\Pi_M \times \mathcal{S}_M) = \mathbb{X}^{\beta}_{\{=\}\alpha}(\Pi_S \times \mathcal{S}_S). \tag{4.2}$$

Proof. It is sufficient to prove (4.2). From lemma 2.4 the first equality of (4.2) is shown. To prove the second part, for any running occupation measure $\{x(\beta, \pi, \tau; i, a)\} \in \mathbb{X}^{\beta}_{\{=\}\alpha}(\Pi \times \mathcal{S})$, we define $w \in \Pi_S$ and $\sigma^{\delta} \in \mathcal{S}_S$ with $\delta = f^{\sigma}$ by the following.

$$w(a|i) := \frac{x\,(\beta, \pi, \tau; i, a)}{x(\beta, \pi, \tau; i)} \quad \text{for } i \in S \text{ and } a \in A, \tag{4.3}$$

$$1 - \delta(i) := \frac{x(\beta, \pi, \tau; i)}{\sum_{t=1}^{\infty} \overline{\mathbb{P}}^{\pi}_{\beta}(X_t = i, \tau \geqq t)} \quad \text{for } i \in S \qquad (4.4).$$

We note that

$$\overline{\mathbb{P}}^{\pi}_{\beta}(X_t = i, \tau \geqq t) = \overline{\mathbb{P}}^{\pi}_{\beta}(X_t = i, \tau > t-1) = \sum_{j \in S, a \in A} \overline{\mathbb{P}}^{\pi}_{\beta}(X_{t-1} = j, \Delta_{t-1} = a, \tau > t-1) p_{ji}(a).$$

So, we get

$$\begin{aligned}
x(\beta, \pi, \tau; i) &= (1 - \delta(i)) \sum_{t=1}^{\infty} \overline{\mathbb{P}}^{\pi}_{\beta}(X_t = i, \tau \geqq t) \quad \text{from (4.4)} \\
&= (1 - \delta(i))\beta(i) + (1 - \delta(i)) \sum_{j \in S, a \in A} x(\beta, \pi, \tau; j, a) p_{ji}(a) \\
&= (1 - \delta(i))\beta(i) + \sum_{j \in S} x(\beta, \pi, \tau; j) \left(\sum_{a \in A} p_{ji}(a) w_j(a) \right) (1 - \delta(i)) \quad \text{from (4.3)} \\
&= (1 - \delta(i))\beta(i) + \sum_{j \in S} x(\beta, \pi, \tau; j) P^{\delta}_{ji}(w).
\end{aligned}$$

Applying Lemma 3.2, we have

$$x(\beta, \pi, \tau; i) = x(\beta, w, \sigma^{\delta}; i), \quad i \in S,$$

as required.▮

In order to drive another mathematical programming formulation, we need the definition of several basic sets. For simplicity, we put $(x_{ia}) = \{x_{ia}\}_{i \in S, a \in A} \in \mathbb{R}^{N \times K}$ and $\delta = \{\delta(i)\}_{i \in S} \in \mathbb{R}^{N}$.

For any distribution β on S and $\alpha(>1)$, let

$$\hat{\mathbb{Q}}_{\{\leqq\}\alpha} := \left\{ \begin{array}{l} ((x_{ia}),\delta) \in \mathbb{R}^{N\times K} \times \mathbb{R}^{N} : \\ \text{(i)} \displaystyle\sum_{a\in A} x_{ia} = \beta(i)(1-\delta(i)) + \sum_{j\in S, a\in A} x_{ja}p_{ji}(a)(1-\delta(i)) \ (i \in S) \\ \text{(ii)} \ 0 \leqq \delta(i) \leqq 1 (i \in S) \\ \text{(iii)} \displaystyle\sum_{i\in S, a\in A} x_{ia} \leqq \alpha - 1 \\ \text{(iv)} \ x_{ia} \geqq 0 \ \ (i \in S, a \in A) \end{array} \right\}. \tag{4.5}$$

Let

$$\mathbb{Q}_{\{\leqq\}\alpha} := \left\{ (x_{ia}) \in \mathbb{R}^{N\times K} : ((x_{ia}),\delta) \in \hat{\mathbb{Q}}_{\{\leqq\}\alpha} \text{ for some } \delta \right\}. \tag{4.6}$$

We denote by $\hat{\mathbb{Q}}_{\{=\}\alpha}$ the subset of $\hat{\mathbb{Q}}_{\{\leqq\}\alpha}$ obtained replacing (iii) in (4.5) by $\sum_{i\in S, a\in A} x_{ia} = \alpha - 1$ and by $\mathbb{Q}_{\{=\}\alpha}$ the set defined in (4.6) replacing $\hat{\mathbb{Q}}_{\{\leqq\}\alpha}$ by $\hat{\mathbb{Q}}_{\{=\}\alpha}$.

Lemma 4.1. *Both* $\mathbb{Q}_{\{\leqq\}\alpha}$ *and* $\mathbb{Q}_{\{=\}\alpha}$ *are compact and convex.*

Proof. Compactness is obvious. To prove the convexity, we show that, for $x^1 = (x^1_{ia}), x^2 = (x^2_{ia}) \in \mathbb{Q}_{\{\leqq\}\alpha}$ and $\gamma \in (0,1)$, $x = (x_{ia}) \in \mathbb{Q}_{\{\leqq\}\alpha}$ with $x_{ia} = \gamma x^1_{ia} + (1-\gamma)x^2_{ia}$, $i \in S, a \in A$. Since $x^1, x^2 \in \mathbb{Q}_{\{\leqq\}\alpha}$, there exist $\delta^1 = (\delta^1(i)), \delta^2 = (\delta^2(i))$ such that

$$x^k_i = \beta(i)(1-\delta^k(i)) + \sum_{j\in S, a\in A} x^k_{ja}p_{ji}(a)(1-\delta^k(i)) \ \ (i \in S, k = 1,2), \tag{4.7}$$

where $x^k_i = \sum_{a\in A} x^k_{ia}$. Now, define $\delta = (\delta(i))$ as follows.

$$1 - \delta(i) = \frac{\gamma x^1_i + (1-\gamma)x^2_i}{\gamma\left(\beta(i) + \sum_{j,a} x^1_{ja}p_{ji}(a)\right) + (1-\gamma)\left(\beta(i) + \sum_{j,a} x^2_{ja}p_{ji}(a)\right)}, \ \ (i \in S). \tag{4.8}$$

where if the denominator is zero, $0 \leqq \delta(i) \leqq 1$ is chosen arbitrary. From (4.7) and (4.8), it follows that $0 \leqq \delta(i) \leqq 1$ and from (4.8) it follows that

$$x_i = \beta(i)(1-\delta(i)) + \sum_{j\in S, a\in A} x_{ja}p_{ji}(a)(1-\delta(i)), \ \ (i \in S)$$

which implies $x \in \mathbb{Q}_{\{\leqq\}\alpha}$. Also, if $x^k \in \mathbb{Q}_{\{=\}\alpha}$ $(k = 1,2), x \in \mathbb{Q}_{\{=\}\alpha}$. Thus, $\mathbb{Q}_{\{=\}\alpha}$ is convex.$\blacksquare$

Theorem 4.2. $\mathbb{Q}_{\{\leqq\}\alpha} = \mathbb{X}^{\beta}_{\{\leqq\}\alpha}(\Pi_S \times \mathcal{S}_S)$.

Proof. From Lemma 3.1 (ii) and Lemma 3.2, the right hand side is clearly contained in the left. To prove the converse, let $x \in \mathbb{Q}_{\{\leqq\}\alpha}$. Then, there exists $\delta = (\delta(i))$ such that $(x,\delta) \in \hat{\mathbb{Q}}_{\{\leqq\}\alpha}$. Define a stationary policy w, for any $a \in A$ and $i \in S$, by

$$w(a|i) = \begin{cases} \dfrac{x_{ia}}{x_i}, & \text{if } x_i > 0 \\ \text{any probability distribution on } A, & \text{if } \ x_i = 0, \end{cases}$$

and consider the pair $(w,\tau) \in \Pi_S \times \mathcal{S}_S$. From the definition of $\hat{\mathbb{Q}}_{\{\leqq\}\alpha}$, we have $x_i = \beta(i)(1-\delta(i)) + \sum_{j\in S} x_j P^{\delta}_{ji}(w)$, where $x_i = \sum_{a\in A} x_{ia}$. Hence, from lemma 3.2, $x_i = x(\beta, w, \tau; i)$. Also, by the definition of w, we get

$$x_{ia} = x_i \frac{x_{ia}}{x_i} = x(\beta, w, \tau; i)\frac{x_{ia}}{x_i} = x(\beta, w, \tau; i, a),$$

which implies $\{x\} = \{x(\beta, w, \tau; i, a)\} \in \mathbb{X}^{\beta}_{\{\leqq\}\alpha}(\Pi_S \times \mathcal{S}_S)$.∎

From this theorem, we have the following corollary.

Corollary 4.1. $\mathbb{X}^{\beta}_{\{\leqq\}\alpha}(\Pi_S \times \mathcal{S}_S)$ *is compact and convex.*

Now, define another Mathematical Programming formulation(MP(II)) for **COP**:

$$\begin{aligned} \textbf{MP(II)}: \ \text{Maximize} \quad & \sum_{i\in S, a\in A} c(i,a)x_{ia} + \sum_{i\in S} r(i)y_i \\ \text{subject to} \quad & (x,\delta) \in \hat{\mathbb{Q}}_{\{\leqq\}\alpha}, y_i = \beta(i) + \sum_{j,a} x_{ja}p_{ji}(a) - \sum_{a} x_{ia}, \ i \in S \end{aligned}$$

From Theorem 4.1 and 4.2, the following corollary easily follows.

Corollary 4.2. **COP** *and* **MP(II)** *are equivalent.*

Let

$\Pi'_S := \{w \in \Pi_S : w$ requires randomization between two actions in at most one state$\}$,
and

$\mathcal{S}'_S := \{\tau \in \mathcal{S}_S | f^{\tau}(i) \in \{0,1\}$ except at most one state $i \in S\}$.

For any compact convex set D we denote by $\mathrm{ext}(D)$ the set of extreme points of D.

Lemma 4.2.

$$\mathrm{ext}\left(\mathbb{X}^{\beta}_{\{=\}\alpha}(\Pi_S \times \mathcal{S}_S)\right) \subset \{x(\beta, w, \tau) : (w,\tau) \in \Pi'_S \times \mathcal{S}'_S\}. \tag{4.9}$$

Proof. By the entire analogy to the proof of Theorem 3.8[3], we can show that

$$\mathrm{ext}\left(\mathbb{X}^{\beta}_{\{=\}\alpha}(\Pi_S \times \mathcal{S}_S)\right) \subset \{x(\beta, w, \tau) : (w,\tau) \in \Pi'_S \times \mathcal{S}_S\}. \tag{4.10}$$

Let $(w,\tau) \in \Pi'_S \times \mathcal{S}_S$. For simplicity, let $\delta = f^{\tau}$. Suppose that there exists $i_1, i_2 \in S (i_1 \neq i_2)$ with $0 < \delta(i_1) < 1, 0 < \delta(i_2) < 1, \mathbb{P}^w_{\beta}(X_t = i_1$ for some $t \geq 1) > 0$ and $\mathbb{P}^w_{\beta}(X_t = i_2$ for some $t \geq 1) > 0$. We consider $\delta^1 = (\delta^1(i)), \delta^2 = (\delta^2(i))$ satisfying the following (4.11) and (4.12):

$$\begin{cases} \delta^k(i) = \delta(i) \ \text{ if } i \neq i_1, i_2 \text{ for } k = 1,2, \\ 0 < \delta^1(i_1) < \delta(i_1) < \delta^2(i_1) < 1, \\ 0 < \delta^2(i_2) < \delta(i_2) < \delta^1(i_2) < 1, \end{cases} \tag{4.11}$$

and

$$\begin{cases} \sum_{i\in S} x(\beta, w, \tau^{\delta^1} : i) = \sum_{i\in S} x(\beta, w, \tau^{\delta^2} : i) = \alpha - 1 \\ x(\beta, w, \tau^{\delta^1}) \neq x(\beta, w, \tau^{\delta^2}). \end{cases} \tag{4.12}$$

Note that the existence of such δ^k $(k = 1,2)$ is easily shown. For simplicity, let $x^{\delta^1}(i) := x(\beta, w, \tau^{\delta^1}; i)$ and $x^{\delta^2}(i) := x(\beta, w, \tau^{\delta^2}; i)$, $i \in S$. Let $b \in (0,1)$ be such that

$$1 - \delta(i) = \frac{bx^{\delta^1}(i_1) + (1-b)x^{\delta^2}(i)}{\beta(i) + (\sum_{k\in S}(bx^{\delta^1}(k) + (1-b)x^{\delta^2}(k))(P(w))_{ki_1})} \quad \text{for all} \quad i \in S (i \neq i_2) \tag{4.13}$$

By the definition of δ^1 and δ^2 we observe that such a b exists. Using this $b \in (0,1)$, we define $\tilde{\delta} = (\tilde{\delta}(i))$ as follows:

$$1 - \tilde{\delta}(i_2) = \frac{bx^{\delta^1}(i_2) + (1-b)x^{\delta^2(i_2)}}{\beta(i_2) + \left(\sum_{k \in S}(bx_k^{\delta^1} + (1-b)x_k^{\delta^2})(P(w))_{ki_2}\right)}, \tag{4.14}$$

$$\text{and} \qquad \tilde{\delta}(i) = \delta(i) \text{ if } i \neq i_2.$$

Then, applying Lemma 3.2, by (4.13) and (4.14), we get

$$x(\beta, w, \tau^{\tilde{\delta}}) = bx^{\delta^1} + (1-b)x^{\delta^2}. \tag{4.15}$$

By (4.15), $\sum_{i \in S} x(\beta, w, \tau^{\tilde{\delta}}; i) = \alpha - 1$, so that from (4.14), we can assume that $\tilde{\delta} = \delta$. Thus, by (4.15), $x(\beta, w, \tau^{\delta})$ is not an extreme point. The above discussion shows that

$$\operatorname{ext}(\{x(\beta, w, \tau) : (w, \tau) \in (\Pi'_S \times \mathcal{S}_S)\}) \subset \{x(\beta, w, \tau) : (w, \tau) \in \Pi'_S \times \mathcal{S}'_S\}.$$

which implies, together with (4.10), that (4.9) holds. ▮

Theorem 4.3. *For* **COP**, *there exists an optimal pair in* $\Pi'_S \times \mathcal{S}'_S$.

Proof. There exists an optimal pair $(w^*, \tau^*) \in \Pi_S \times \mathcal{S}_S$ from Theorem 4.1, Corollary 4.1 and Theorem 3.1. For $\alpha' := \mathbb{E}_\beta^{w^*}[\tau^*] \leqq \alpha, (w^*, \tau^*) \in \mathbb{X}^\beta_{\{=\}\alpha'}(\Pi_S \times \mathcal{S}_S)$. Hence, since the objective function of **MP(II)** is linear, from Lemma 4.2 the theorem follows. ▮

Here, we give the following numerical example:
$S = \{1,2,3,4\}, A = \{1\}, \alpha = 3, \beta = (0.25, 0.25, 0.25, 0.25)$,

$$(p_{ij}(1)) = \begin{pmatrix} 0.3 & 0.4 & 0.1 & 0.2 \\ 0.4 & 0.1 & 0.2 & 0.3 \\ 0.2 & 0.3 & 0.4 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.3 \end{pmatrix},$$

$c(1,1) = 0.6, c(2,1) = 0.1, c(3.1) = 0.5, c(4,1) = 0.4, r(1) = 4, r(2) = 3, r(3) = 2, r(4) = 2$.
Letting $x_i = x_{i1}$ $(i \in S)$, the Mathematical Programming formulation(**MP(II)**) for the corresponding **COP** is given as follows:

$$\begin{array}{ll} \text{Maximize} & -1.6x_1 - 0.2x_2 + 0.2x_3 + 0.5x_4 + 2.75 \\ \text{subject to} & x_1 = (0.25 + 0.3x_1 + 0.4x_2 + 0.2x_3 + 0.3x_4)(1 - \delta(1)), \\ & x_2 = (0.25 + 0.4x_1 + 0.1x_2 + 0.3x_3 + 0.3x_4)(1 - \delta(2)), \\ & x_3 = (0.25 + 0.1x_1 + 0.2x_2 + 0.4x_3 + 0.1x_4)(1 - \delta(3)), \\ & x_4 = (0.25 + 0.2x_1 + 0.3x_2 + 0.1x_3 + 0.3x_4)(1 - \delta(4)), \\ & x_1 + x_2 + x_3 + x_4 \leqq 2, \\ & x_1, x_2, x_3, x_4 \geqq 0, 1 \geqq \delta(1), \delta(2), \delta(3), \delta(4) \geqq 0. \end{array}$$

After a simple calculation, we find that the optimal solution of the above is $x_1^* = 0, x_2^* = 89/156, x_3^* = 113/156, x_4^* = 55/78, \delta^*(1) = 1, \delta^*(2) = 129/574, \delta^*(3) = \delta^*(4) = 0$ and the optimal value is $611/195 (\doteqdot 3.13)$. Note that the value is $75/82 (\doteqdot 3.06)$ for $\delta(1) = \delta(2) = 1$ and $\delta(3) = \delta(4) = 0$.
Thus, by Corollary 4.2 and Theorem 4.3, the pair $(w^*, \tau^*) \in \Pi'_S \times \mathcal{S}'_S$ with $w^*(i) = 1$ for all $i \in S$ and $f^{\tau^*}(1) = \delta^*(1) = 1, f^{\tau^*}(2) = \delta^*(2) = 129/574, f^{\tau^*}(3) = \delta^*(3) = 0, f^{\tau^*}(4) = \delta^*(4) = 0$ is optimal for the corresponding **COP** and the optimal reward $J(\beta, w^*, \tau^*) = 611/195$.

References

[1] Altman E (1994) Denumerable constrained Markov decision process and finite approximations. Math. Oper. Res. 19:169–191

[2] Altman E (1996) Constrained Markov decision processes with total cost criteria: occupation measures and primal LP. Math. Methods Oper. Res. 43:45–72

[3] Altman E (1999) Constrained Markov Decision Processes. Chapmann & Hall/CRC

[4] Assaf D, Samuel-Cahn E (1998) Optimal multivariate stopping rules. J. Appl. Probab. 35:693–706

[5] Borkar VS (1991) Topics in controlled Markov chains. Longman Scientific & Technical. John Wiley & Sons, Inc., New York

[6] Chow YS, Robbins H, Siegmund D (1976) Great expectations: the theory of optimal stopping. Houghton Mifflin Co. Boston, Mass.

[7] Frid EB On optimal strategies in control problems with constraints. Theory of Probability and its Applications. 17:188–192

[8] Furukawa N, Iwamoto S (1970) Stopped decision processes on complete separable metric spaces. J. Math. Anal. Appl. 31:615–658

[9] Hordijk A (1974) Dynamic programming and Markov potential theory. Mathematical Centre Tracts, No. 51. Mathematisch Centrum Amsterdam.

[10] Hordijk A, Kallenberg LCM (1984) Constrained undiscounted stochastic dynamic programming. Math. Oper. Res. 9:276–289

[11] Horiguchi M, Kurano M, Yasuda M (preprint) Markov decision process with constrained stopping times.

[12] Irle A (1998) Minimax results and randomization for certain stochastic games. In: Ricceri B, Stephen S (eds.) Minimax Theory and Applications. Kluwer Acad. Publ., Dordrecht, pp. 91–103

[13] Kadota Y, Kurano M, Yasuda M (1996) Utility-optimal stopping in a denumerable Markov chain. Bull. Inform. Cybernet. 28:15–21

[14] Kadota Y, Kurano M, Yasuda M (1998) On the general utility of discounted Markov decision processes. Int. Trans. Opl. Res. 5:27–34

[15] Kennedy DP (1982) On a constrained optimal stopping problem. J. Appl. Probab. 19:631–641

[16] Luenberger DG (1969) Optimization by vector space methods. John Wiley & Sons, Inc. New York-London-Sydney.

[17] Nachman DC (1980) Optimal stopping with a horizon constraint. Math. Oper. Res. 5:126–134

[18] Royden HL (1968) Real Analysis. 2nd Edition. Macmillan, New York.