A linear programming approach to constrained nonstationary infinite-horizon Markov decision processes

Ilbin Lee
Marina A. Epelman
H. Edwin Romeijn
Robert L. Smith

Technical Report 13-01

March 6, 2013

University of Michigan
Industrial and Operations Engineering
1205 Beal Avenue
Ann Arbor, MI 48109

A linear programming approach to constrained nonstationary infinite-horizon Markov decision processes

Ilbin Lee, Marina A. Epelman, H. Edwin Romeijn, Robert L. Smith*

March 6, 2013

Abstract

Constrained Markov decision processes (MDPs) are MDPs optimizing an objective function while satisfying additional constraints. We study a class of infinite-horizon constrained MDPs with nonstationary problem data, finite state space, and discounted cost criterion. This problem can equivalently be formulated as a *countably infinite linear program* (CILP), *i.e.*, a linear program (LP) with a countably infinite number of variables and constraints. Unlike finite LPs, CILPs can fail to satisfy useful theoretical properties such as duality, and to date there does not exist a general solution method for such problems. Specifically, the characterization of extreme points as basic feasible solutions in finite LPs does not extend to general CILPs. In this paper, we provide duality results and a complete characterization of extreme points of the CILP formulation of constrained nonstationary MDPs with finite state space, and illustrate the characterization for special cases. As a corollary, we obtain the existence of a K-randomized optimal policy, where K is the number of constraints.

1 Introduction

For the last couple of decades, growing attention has been given to solving constrained Markov decision processes (MDPs). Constrained MDPs are MDPs optimizing an objective function while satisfying constraints, typically on budget, quality, and so on. In addition, decision making problems with multiple criteria are often approached by optimizing one criterion while satisfying constraints on the other criteria, which also turns into a constrained MDP. Such problems often arise in data communications. In queueing systems with service rate control, the average throughput is maximized with constraints on the average delay [13, 15]. Priority queueing systems with a fixed service rate are another example [4, 16, 19]. Here, one optimizes the queueing time of noninteractive traffic while satisfying a constraint on the average end-to-end delay of interactive traffic. For these problems, [20] considered a case where service rate costs and penalty costs of delay are actually incurred in discrete time periods and it is desired to minimize the discounted service rate cost with constraints on the discounted delay cost. Facility maintenance is another type of problems modeled by constrained MDPs. [11] considered finding an optimal maintenance policy for each mile of a network of highways and [22] studied a building management problem. In the models for these problems, the total cost is minimized subject to constraints on quality of facilities.

In this paper we study an infinite-horizon constrained MDP minimizing the discounted cost criterion

*Industrial and Operations Engineering Department, University of Michigan, Email addresses: ilbinlee@umich.edu, mepelman@umich.edu, romeijn@umich.edu, rlsmith@umich.edu

with nonstationary problem data and finite state space, which we call *constrained nonstationary MDP with finite state space*. Constrained nonstationary MDP with finite state space is obtained from a constrained stationary MDP with finite state space by dropping the stationary data assumption which often does not hold in practice. It is also well-known that constrained nonstationary MDPs with finite state space form a subclass of constrained MDPs with stationary data and countably infinite number of states. Constrained nonstationary MDP with finite state space has an equivalent LP formulation and the LP has countably infinite number of variables and countably infinite number of constraints [3], which we call *countably infinite linear program* (CILP). Unlike finite LPs, CILPs lack a general solution method and may fail useful theoretical properties such as duality, which make them hard to analyze [5]. By Bauer's Maximum Principle [1], for finite LPs the minimum is achieved at an extreme point, and often for CILPs as well. For finite LPs, a feasible solution is an extreme point if and only if it is a basic solution. This equivalency translates the geometric concept of an extreme point to the algebraic object of a basic solution. However, such an algebraic characterization of extreme points does not extend to CILPs [9]. In this paper we provide algebraic necessary conditions for a feasible solution of the CILP formulation of constrained nonstationary MDP with finite state space to be an extreme point of its feasible region. Using those necessary conditions, we also establish a necessary and sufficient condition for a feasible solution to be an extreme point that can be checked in a finite dimensional polyhedron with which we are familiar. This is a complete characterization of extreme points for a class of CILPs representing constrained nonstationary MDPs with finite state space. [10] recently studied the CILP representation of unconstrained nonstationary MDPs with finite state space and established duality results and algebraic characterization of extreme points. They also developed a simplex algorithm for the CILP and showed that it achieved optimality at the limit. For the CILP representation of constrained nonstationary MDPs with finite state space, we introduce the duality results proven in [3], define complementary slackness, and establish its relation to optimality. Thus, this report sets important foundations for developing a simplex-type algorithm for constrained nonstationary MDPs with finite state space.

Under typical settings for constrained MDPs, there exists a stationary optimal policy but a deterministic stationary optimal policy may not exist [8]. Thus, a pursued goal in literature is proving existence of an optimal policy that is as close to deterministic as possible, that is, the existence of K-randomized optimal policy, where K is the number of constraints and a policy is K-randomized if it uses K "more" actions than a deterministic stationary policy (for a more precise definition, see Section 4). It is well-known that extreme points of the LP formulation of unconstrained MDPs with finite number of states correspond to deterministic policies. Now consider a constrained MDP obtained by adding linear constraints to an unconstrained MDP. Then, an extreme points of the LP formulation of the constrained MDP is a convex combination of extreme points of the unconstrained MDP, *i.e.*, deterministic policies, and this explains how the randomization is introduced. For constrained stationary MDPs with finite state space, there exists a K-randomized optimal policy and it can be found by obtaining an optimal basic feasible solution of their equivalent finite LP formulation [12, 14, 18]. For constrained stationary MDPs with countably infinite number of states, a K-randomized optimal policy is proven to exist for single constraint case in [20] using Lagrangian multiplier approach and general case in [7] by studying the Pareto frontier of the performance set. We obtain the existence of K-randomized optimal policy for constrained nonstationary MDPs as a byproduct of characterizing extreme points of their CILP formulation.

2 Problem Formulation

Consider a dynamic system operating in discrete time periods on a finite state space. In periods $n \in \mathbb{N}$, the system is observed in a state $s \in \mathcal{S}$ and an action $a \in \mathcal{A}$ is chosen, where $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$ are both finite. After multiple kinds of costs, $c_n(s,a)$ and $d_n^k(s,a)$ for $k = 1, 2, \ldots, K$ are incurred where K is a finite integer, the system makes a transition to be observed in a state s' at the beginning of period $n+1$, with probability $p_n(s'|s,a)$. This process continues indefinitely. The costs are assumed to be nonnegative and uniformly bounded, *i.e.*, there exist c and d^k for $k = 1, 2, \ldots, K$ such that $0 \leq c_n(s,a) \leq c$, $0 \leq d_n^k(s,a) \leq d^k$ for $n \in \mathbb{N}, s \in \mathcal{S}, a \in \mathcal{A}$ and $k = 1, 2, \ldots, K$. The goal is to minimize the expected discounted "c-cost" satisfying K constraints on the expected discounted "d^k-costs" for $k = 1, 2, \ldots, K$, with a common discount factor $0 < \alpha < 1$. A policy π is a sequence $\pi = \{\pi_1, \pi_2, \ldots\}$, where π_n is a probability measure over $\mathcal{A}$ conditioned on the whole history of states and actions before period n plus the current state at the beginning of period n. Given an initial state distribution β, each policy π induces a probability measure P_β^π on which the state process $\{S_n\}_{n=1}^\infty$ and the action process $\{A_n\}_{n=1}^\infty$ are defined. The corresponding expectation operator is denoted as E_β^π. Let

$$C(\beta,\pi) \triangleq E_\beta^\pi\big[\sum_{n=1}^\infty \alpha^{n-1} c_n(S_n, A_n)\big]$$

$$D^k(\beta,\pi) \triangleq E_\beta^\pi\big[\sum_{n=1}^\infty \alpha^{n-1} d_n^k(S_n, A_n)\big] \text{ for } k = 1, 2, \ldots, K,$$

and let $\Pi \triangleq \{\pi | D^k(\beta,\pi) \leq V_k \text{ for } k = 1, 2, \ldots, K\}$. The optimization problem at hand can be written as

$$(Q) \ \min_{\pi \in \Pi} C(\beta,\pi).$$

It was shown that optimal policy for constrained MDP (Q) may depend on the initial state [8], thus (Q) is formulated with a fixed initial state distribution β. This problem can be viewed as a constrained stationary MDP with countable number of states by appending the states $s \in \mathcal{S}$ with time-indices $n \in \mathbb{N}$. For constrained stationary MDPs, it was shown in [3] that, without loss of optimality, we can restrict attention to Markov policies. In the stationary MDP counterpart of constrained nonstationary MDPs with finite state space, a Markov policy is also stationary because each period-state pair is visited only once. Moreover, any stationary policy in the stationary MDP counterpart corresponds to a Markov policy in the original constrained nonstationary MDP with finite state space, and thus, we can restrict our attention to Markov policies for constrained nonstationary MDPs with finite state space.

It was proven that (Q) has an equivalent CILP formulation [2, 3], which can be written as:

$$(P)\ \min\ f(x) = \sum_{n\in\mathbb{N}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1} c_n(s,a) x_n(s,a) \tag{1}$$

$$\text{s.t.}\ \sum_{a\in\mathcal{A}} x_1(s,a) = \beta(s) \text{ for } s\in\mathcal{S} \tag{2}$$

$$\sum_{a\in\mathcal{A}} x_n(s,a) - \sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_{n-1}(s|s',a) x_{n-1}(s',a) = 0 \text{ for } n\in\mathbb{N}\setminus\{1\}, s\in\mathcal{S} \tag{3}$$

$$\sum_{n\in\mathbb{N}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a) x_n(s,a) \le V_k \text{ for } k = 1,2,\ldots,K \tag{4}$$

$$x \ge 0. \tag{5}$$

We call the constraints (2) and (3) the *flow balance constraints*. Let $\mathcal{P}$ be the feasible region of (P). (2) and (3) imply that for $x\in\mathcal{P}$,

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} x_n(s,a) = 1 \text{ for } n\in\mathbb{N}. \tag{6}$$

Since x is nonnegative, we have $0\le x_n(s,a)\le 1$ for $n\in\mathbb{N}, s\in\mathcal{S}, a\in\mathcal{A}$. Because all objective and constraint cost functions are uniformly bounded, the infinite sums in (1) and (4) exist.

Let a hypernetwork be defined as a network in which edges can connect any number of nodes. Then we can view (P) as a minimum cost flow problem with constraints in a staged hypernetwork with countably infinite number of stages and the same finite number of nodes in each stage. In the hypernetwork, each period-state pair (n,s) corresponds to a node and each action a at node (n,s) to a hyperarc which we denote as (n,s,a), and $x_n(s,a)$ is the amount of flow in the hyperarc. For $s'\in\mathcal{S}$ such that $p_n(s'|s,a)>0$, the flow from node (n,s) to node $(n+1,s')$ through the hyperarc (n,s,a) is $p_n(s'|s,a)x_n(s,a)$.

Notice that for a given policy π, there may be some node (n,s) such that $P_\beta^\pi\{S_n = s\} = 0$. For those nodes, it does not matter which actions are chosen for the purpose of solving (P). Thus, we do not distinguish those policies that differ only at those nodes with zero probability. Under the above convention, there exists a one-to-one correspondence between the set of policies and the set of nonnegative x satisfying the flow balance constraints. We refer to a nonnegative solution x satisfying the flow balance constraints as a *hyperarc frequency*. Note that for a policy π and the resulting hyperarc frequency x, $x_n(s,a)$ can be interpreted as the probability of encountering hyperarc (n,s,a) under policy π for the given initial state distribution β. There also exists an obvious one-to-one correspondence between $\mathcal{P}$ and Π. We refer to a (feasible) policy and the corresponding (feasible) hyperarc frequency interchangeably.

Theorem 2.1 *If (P) is feasible, then it has an extreme point optimal solution.*

Proof: It is easy to show that $\mathcal{P}$ is a closed and convex subset of $\mathbb{R}^\infty$. By Tychonoff's product theorem (see [1]) and (6), $\mathcal{P}$ is a subset of a compact set and thus, it is compact. Since the objective function is continuous and convex, by Bauer's Maximum Principle [1], (P) has an extreme point optimal solution. □

3 Duality Results

In this section, we define the dual of (P), introduce strong duality result, define complementary slackness, and prove its relation to optimality. We define the dual of (P) as

$$(D) \ \max \ g(y,\mu) = \sum_{s\in\mathcal{S}} \beta(s) y_1(s) - \sum_{k=1}^{K} V_k \mu_k \tag{7}$$

$$\text{s.t. } y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a) y_{n+1}(s') - \alpha^{n-1} \sum_{k=1}^{K} d_n^k(s,a)\mu_k \le \alpha^{n-1} c_n(s,a)$$

$$\text{for } n \in \mathbb{N}, s \in \mathcal{S}, a \in \mathcal{A} \tag{8}$$

$$\mu \ge 0 \tag{9}$$

$$y \in Y, \tag{10}$$

where Y is the subspace of all sequences $y \triangleq \{y_n(s)\}$ indexed by $(n,s) \in \mathbb{N}\times\mathcal{S}$ such that there exists τ_y that satisfies $|y_n(s)| \le \alpha^{n-1}\tau_y$ for all $(n,s) \in \mathbb{N}\times\mathcal{S}$. This formulation is more restrictive than the LP formulation of constrained MDPs given in [3] (Theorem 9.11) since (D) has the additional constraint (10). However, we make the following assumption which is also known as the Slater's condition, and it is easy to show that under the assumption the constraint (10) does not exclude an optimal solution of our interest. Note that this assumption is necessary only for the results in this section.

Assumption 1 (The Slater's condition) *(P) has a strict feasible solution, i.e., there exists a feasible solution x to (P) that satisfies all of the inequality constraints (4) strictly.*

To see that adding (10) does not exclude an optimal solution, consider the CILP obtained from (D) by removing the constraint (10). [3] shows that under the Slater's condition, there exists μ^* that achieves optimality of the CILP with some y^* . If we fix $\mu = \mu^*$ in (D), then it is equivalent to the CILP formulation of an unconstrained MDP. Then, y^* is the cost-to-go vector of an optimal policy of the unconstrained MDP and we can easily prove that y^* satisfies (10). Therefore, the constraint (10) does not rule out the optimal solution of our interest.

Note that we can derive (D) except (10) by applying the rules obtaining dual of finite LPs to (P). In general, if we apply the rules to CILPs, then the "dual" may not satisfy weak duality [17]. However, [3] showed that strong duality holds between primal and dual CILP formulation of constrained MDPs with countably infinite number of states, which also implies strong duality between (P) and (D).

Theorem 3.1 (Theorem 9.11 of [3]) *The optimal objective function values of (P) and (D) coincide.*

We now define complementary slackness of (P) and (D) and show that feasible solutions of (P) and (D) are optimal to their corresponding problems if and only if they satisfy the complementary slackness. Under the Slater's condition, [3] showed necessity of complementary slackness for optimality by using the concept of occupation measure and the interpretation of constrained MDPs as an inf-sup problem with Lagrangian multipliers. We provide an alternative proof for necessity, and moreover, establish sufficiency of complementary slackness for optimality.

Definition 3.2 (Complementary slackness) *Suppose x is feasible to (P). Then we say that x and*

(y, μ) *satisfy complementary slackness if*

$$x_n(s,a)\big[\alpha^{n-1}\big(c_n(s,a)+\sum_{k=1}^{K} d_n^k(s,a)\mu_k\big) - y_n(s) + \sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s')\big] = 0 \text{ for } n\in\mathbb{N}, s\in\mathcal{S}, a\in\mathcal{A}, \tag{11}$$

$$\mu_k\big[V_k - \sum_{n\in\mathbb{N}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a)x_n(s,a)\big] = 0 \text{ for } k = 1,2,\ldots,K. \tag{12}$$

Theorem 3.3 (Complementary slackness sufficiency) *Suppose x is feasible to (P) and satisfies complementary slackness with some (y,μ). Then $f(x) = g(y,\mu)$. If (y,μ) is feasible to (D), then x and (y,μ) are optimal to (P) and (D), respectively.*

Proof: From the complementary slackness condition (11), we have

$$\alpha^{n-1}c_n(s,a)x_n(s,a) = \big(y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s')\big)x_n(s,a) - \alpha^{n-1}\big(\sum_{k=1}^{K} d_n^k(s,a)\mu_k\big)x_n(s,a)$$

for $n\in\mathbb{N}, s\in\mathcal{S}, a\in\mathcal{A}$. By summing up both sides for $n = 1,2,\ldots,N$, $s\in\mathcal{S}, a\in\mathcal{A}$, we obtain

$$\begin{aligned}
&\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1}x_n(s,a)c_n(s,a)\\
&= \sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \big(y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s')\big)x_n(s,a) - \sum_{k=1}^{K}\mu_k\sum_{n=1}^{N}\alpha^{n-1}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_n^k(s,a)x_n(s,a).
\end{aligned} \tag{13}$$

We simplify the first sum of the right hand side as follows:

$$\begin{aligned}
&\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \big(y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s')\big)x_n(s,a)\\
&= \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_n(s)\sum_{a\in\mathcal{A}} x_n(s,a) - \sum_{n=1}^{N}\sum_{s'\in\mathcal{S}} y_{n+1}(s')\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_n(s'|s,a)x_n(s,a)\\
&= \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_n(s)\sum_{a\in\mathcal{A}} x_n(s,a) - \sum_{n=1}^{N}\sum_{s'\in\mathcal{S}} y_{n+1}(s')\sum_{a\in\mathcal{A}} x_{n+1}(s',a)\\
&= \sum_{s\in\mathcal{S}} y_1(s)\sum_{a\in\mathcal{A}} x_1(s,a) - \sum_{s\in\mathcal{S}} y_{N+1}(s)\sum_{a\in\mathcal{A}} x_{N+1}(s,a)\\
&= \sum_{s\in\mathcal{S}} \beta(s)y_1(s) - \sum_{s\in\mathcal{S}} y_{N+1}(s)\sum_{a\in\mathcal{A}} x_{N+1}(s,a).
\end{aligned} \tag{14}$$

By substituting (14) into (13), we have

$$\begin{aligned}
&\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1}c_n(s,a)x_n(s,a)\\
&= \sum_{s\in\mathcal{S}} \beta(s)y_1(s) - \sum_{s\in\mathcal{S}} y_{N+1}(s)\sum_{a\in\mathcal{A}} x_{N+1}(s,a) - \sum_{k=1}^{K}\mu_k\sum_{n=1}^{N}\alpha^{n-1}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_n^k(s,a)x_n(s,a).
\end{aligned} \tag{15}$$

The second term above goes to zero as N increases, because

$$-S\alpha^N\tau_y \le \sum_{s\in\mathcal{S}} y_{N+1}(s) \sum_{a\in\mathcal{A}} x_{N+1}(s,a) \le S\alpha^N\tau_y.$$

By the second condition (12) of complementary slackness, for $k = 1, 2, \ldots, K$,

$$\mu_k \sum_{n=1}^{N} \alpha^{n-1} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d_n^k(s,a) x_n(s,a) \to V_k\mu_k \text{ as } N \to \infty.$$

Thus, taking $N \to \infty$ on both sides of (15) gives $f(x) = g(y,\mu)$. The second statement of the theorem follows by weak duality. □

Theorem 3.4 (Complementary slackness necessity) *Suppose x and (y,μ) are optimal to (P) and (D), respectively. Then the complementary slackness conditions hold.*

Proof: We have $x_n(s,a) \ge 0$ and $\alpha^{n-1}\big(c_n(s,a)+\sum_{k=1}^K d_n^k(s,a)\mu_k\big) \ge y_n(s)-\sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s')$ for all $n \in \mathbb{N}, s \in \mathcal{S}, a \in \mathcal{A}$. Thus, the left hand side of the first condition (11) of complementary slackness is nonnegative for $n \in \mathbb{N}, s \in \mathcal{S}, a \in \mathcal{A}$. By summing up the left hand side of (11) for $n = 1, 2, \ldots, N$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$, we obtain

$$\begin{aligned}
0 \le & \sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} x_n(s,a)\big[\alpha^{n-1}\big(c_n(s,a)+\sum_{k=1}^{K} d_n^k(s,a)\mu_k\big) - y_n(s) + \sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s')\big] \\
= & \sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1}c_n(s,a)x_n(s,a) + \sum_{k=1}^{K}\mu_k \sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_n^k(s,a)x_n(s,a) \\
& - \sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \big[y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s')\big]x_n(s,a) \\
= & \sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1}c_n(s,a)x_n(s,a) + \sum_{k=1}^{K}\mu_k \sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_n^k(s,a)x_n(s,a) \\
& - \sum_{s\in\mathcal{S}} \beta(s)y_1(s) + \sum_{s\in\mathcal{S}} y_{N+1}(s) \sum_{a\in\mathcal{A}} x_{N+1}(s,a)
\end{aligned}$$

where we used (14) to get the last equality. By taking $N \to \infty$,

$$\begin{aligned}
0 \le & \lim_{N\to\infty} \sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} x_n(s,a)\big[\alpha^{n-1}\big(c_n(s,a)+\sum_{k=1}^{K} d_n^k(s,a)\mu_k\big) - y_n(s) + \sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s')\big] \\
\le & f(x) + \sum_{k=1}^{K} V_k\mu_k - \sum_{s\in\mathcal{S}} \beta(s)y_1(s) = f(x) - g(y,\mu) = 0,
\end{aligned}$$

owing to strong duality. This shows that the sum of the left hand side of (11) equals zero. However, we know the left hand side of (11) for $n \in \mathbb{N}, s \in \mathcal{S}, a \in \mathcal{A}$ is nonnegative, and thus, each of them equals zero. Therefore, (11) holds.
We now prove (12). We have $\mu_k \ge 0$ and

$$V_k \ge \sum_{n\in\mathbb{N}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a)x_n(s,a)$$

for $k = 1, 2, \ldots, K$ since x is feasible to (P). Therefore, the left hand side of (12) is nonnegative for $k = 1, 2, \ldots, K$. By summing up the left hand side of (12) for $k = 1, 2, \ldots, K$, we obtain

$$
\begin{aligned}
0 &\le \sum_{k=1}^{K} \mu_k \big[V_k - \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a) x_n(s,a) \big] \\
&= \sum_{s\in\mathcal{S}} \beta(s) y_1(s) - \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} c_n(s,a) x_n(s,a) - \sum_{k=1}^{K} \mu_k \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a) x_n(s,a) \qquad (16) \\
&= \lim_{N\to\infty} \big[\sum_{s\in\mathcal{S}} \beta(s) y_1(s) - \sum_{n=1}^{N} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} \big(c_n(s,a) + \sum_{k=1}^{K} d_n^k(s,a) \mu_k \big) x_n(s,a) \big] \qquad (17) \\
&\le \lim_{N\to\infty} \big[\sum_{s\in\mathcal{S}} \beta(s) y_1(s) - \sum_{n=1}^{N} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \big(y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a) y_{n+1}(s') \big) x_n(s,a) \big] \qquad (18) \\
&= \lim_{N\to\infty} \sum_{s\in\mathcal{S}} y_{N+1}(s) \sum_{a\in\mathcal{A}} x_{N+1}(s,a) \qquad (19) \\
&= 0
\end{aligned}
$$

where the equality (16) is strong duality; (17) is rearrangement of sums; the inequality (18) is the constraint (8) of (D); and (19) is due to (14). Therefore, (12) holds. Consequently, the complementary slackness is shown. □

4 Splitting Randomized Policies

One of the main objectives in this paper is to study extreme points of (P), the LP formulation of constrained nonstationary MDPs. The definition of an extreme point of a convex set is a point in the set that cannot be represented as a non-trivial convex combination of other points in the set. This section presents some preliminary results in the form of two different representations of a randomized policy as a convex combination of other policies, which will help us identify characteristics of extreme points. (Note, however, that the results in this section are not limited to constrained MDPs.) The first one represents a randomized policy as a convex combination of deterministic policies, a well-known one in literature. It will be needed in Section 6 to present a necessary and sufficient condition for a feasible solution of (P) to be an extreme point. The second one is needed in Section 5 to provide necessary conditions for an extreme point and has some specific characteristics required for that purpose.

We first introduce some definitions which will be helpful in describing these representations. We define a *submodel* of the MDP to be an MDP that is identical to the original one in all respects except that the action sets are limited to $B_n(s) \subset \mathcal{A}$ for $n \in \mathbb{N}, s \in \mathcal{S}$. For a given policy x, we also define a *submodel defined by* x as a submodel such that $B_n(s) = \{a \in \mathcal{A} \mid x_n(s,a) > 0\}$ for $n \in \mathbb{N}, s \in \mathcal{S}$. For a submodel B, we call the number $M = \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} (|B_n(s)| - 1)$ as the *index of the submodel*. A randomized policy that belongs to a submodel B with index M can be interpreted as using at most M "more" actions than a deterministic policy. Recall that in each period of the original MDP, there are S states and each state has A action choices. Thus, in each period a policy can use up to $|\mathcal{S}|(|\mathcal{A}| - 1)$ "more" actions than a deterministic policy.

Definition 4.1 *A randomized policy that belongs to a submodel with index M is called an* M-randomized policy. *An M-randomized policy that does not belong to any submodel with index less than M is called an* exactly M-randomized policy. *A randomized policy that does not belong to any submodel with a finite index is called an* ∞-randomized policy.

4.1 Splitting into deterministic policies

It has been shown that for a finite positive integer M, any M-randomized policy can be represented as a convex combination of $M+1$ *deterministic* policies [7].

Lemma 4.2 (cf. Theorem 5.1 in [7]) *For any finite positive integer M, any exactly M-randomized policy is a convex combination of $M+1$ 0-randomized (i.e., deterministic) policies.*

In addition, it was recently shown that for any finite positive integer M, it is possible to represent a M-randomized policy as a convex combination of $M+1$ deterministic policies that can be ordered so that each pair of consecutive policies differ at only one period-state pair [6]. They also provided an efficient algorithm to find the convex combination of deterministic policies.
Consider an exactly M-randomized policy x for a finite positive integer M. We introduce a set $\Lambda(x)$ which plays an important role in the necessary and sufficient condition for an extreme point which will be presented in Section 6. Let B be the submodel defined by x. Since M is finite, the number of deterministic policies in the submodel B is also finite, say N. Let $x^1, x^2, \ldots, x^N$ be these deterministic policies. Let

$$\Lambda(x) = \left\{ \lambda \in \mathbb{R}^N \mid x = \sum_{i=1}^{N} \lambda_i x^i, \sum_{i=1}^{N} \lambda_i = 1, \lambda \geq 0 \right\}. \tag{20}$$

That is, $\Lambda(x)$ is the set of nonnegative weights by which convex combination of $x^1, x^2, \ldots, x^N$ becomes x.
In (20), the definition of $\Lambda(x)$ is given by an infinite number of linear equations, but in fact, we can represent $\Lambda(x)$ by a finite number of linear equations as shown by the next theorem.

Theorem 4.3 *Let x be an exactly M-randomized policy and N be the number of deterministic policies in the submodel defined by x. Then there exists a matrix $A = \mathbb{R}^{M \times N}$ and $b \in \mathbb{R}^M$ such that $\Lambda(x) = \{\lambda \in \mathbb{R}^N \mid A\lambda = b, \mathbf{1}^T\lambda = 1, \lambda \geq 0\}$, and the matrix $\begin{bmatrix} A \\ \mathbf{1}^T \end{bmatrix}$ has a full row rank.*

Proof: For simplicity of illustration, we prove this theorem under an assumption that x does not allow any node that has zero incoming flow. The presented proof can be easily extended to the general case but it would detract us from the key idea.

Let $(n_1, s_1), (n_2, s_2), \ldots, (n_m, s_m)$ be the period-state pairs at which x randomizes and suppose that they are ordered so that the period index is nondecreasing. For $i = 1, 2, \ldots, m$, assume that x randomizes over $a^{i,1}, a^{i,2}, \ldots, a^{i,l_i}$ at (n_i, s_i). We have $\sum_{i=1}^{m}(l_i - 1) = M$ and $\prod_{i=1}^{m} l_i = N$. Let $\Lambda^0(x) = \{\lambda \in \mathbb{R}^N \mid \mathbf{1}^T\lambda = 1, \lambda \geq 0\}$. Since x does not randomize in periods 1 to $n_1 - 1$, the policies $x^1, x^2, \ldots, x^N$ does not randomize in those periods. Consequently, for any $\lambda \in \Lambda^0(x)$, $\sum_{i=1}^{N} \lambda_i x^i$ and x are the same in periods 1 to $n_1 - 1$. This implies that they also have the same flow on hyperarcs from the period-state pairs in period n_1 where x does not randomize. Let $\Lambda^1(x) = \{\lambda \in \mathbb{R}^N \mid \sum_{i=1}^{N} x^i_{n_1}(s_1, a^{1,1})\lambda_i = x_{n_1}(s_1, a^{1,1}), \mathbf{1}^T\lambda = 1, \lambda \geq 0\}$. Then, for any $\lambda \in \Lambda^1(x)$, $\sum_{i=1}^{N} \lambda_i x^i$ and x coincide in periods 1 to $n_1 - 1$ and on those hyperarcs from the period-state pairs in period n_1 where x does not randomize. Moreover, in period n_1, they have the same flow

on hyperarc $(n_1, s_1, a^{1,1})$. Let $\Lambda^{l_1-1}(x) = \{\lambda \in \mathbb{R}^N \mid \sum_{i=1}^N x^i_{n_1}(s_1, a^{1,j})\lambda_i = x_{n_1}(s_1, a^{1,j})$ for $j = 1, 2, \ldots, l_1 - 1, \mathbf{1}^T\lambda = 1, \lambda \geq 0\}$. For any $\lambda \in \Lambda^{l_1-1}(x)$, $\sum_{i=1}^N \lambda_i x^i$ and x coincide in periods 1 to $n_1 - 1$ and on those hyperarcs from the period-state pairs in period n_1 where x does not randomize, and additionally in period n_1, they have the same flow on hyperarc $(n_1, s_1, a^{1,j})$ for $j = 1, 2, \ldots, l_1 - 1$. Then, they also have the same flow on hyperarc (n_1, s_1, a^{1,l_1}), and thus, they coincide on all hyperarcs from (n_1, s_1). Note that x randomizes over l_1 actions at (n_1, s_1) and we added $l_1 - 1$ equations to obtain $\Lambda^{l_1-1}(x)$ from $\Lambda^0(x)$.

We can apply the same procedure to the other period-state pairs $(n_2, s_2), (n_3, s_3), \ldots, (n_m, s_m)$, in the order of nondecreasing period index. Then we obtain $\Lambda^M(x)$ such that for any $\lambda \in \Lambda^M(x)$, $\sum_{i=1}^N \lambda_i x^i$ and x coincide in periods 1 to $n_m - 1$ (which implies that they also have the same flow on all hyperarcs from any period-state pair in period n_m where x does not randomize), and additionally in period n_m, they have the same flow on all hyperarcs from any period-state pair where x randomizes. Thus, for any $\lambda \in \Lambda^M(x)$, $\sum_{i=1}^N \lambda_i x^i$ and x coincide. We showed that $\Lambda^M(x) \subset \Lambda(x)$. We can easily show that any $\lambda \in \Lambda(x)$ satisfies all of the equalities that define $\Lambda^M(x)$. Therefore, we showed $\Lambda^M(x) = \Lambda(x)$.

$\Lambda^M(x)$ has $M + 1$ equality constraints for λ and let A be the coefficient matrix of the M equalities added to obtain $\Lambda^M(x)$ from $\Lambda^0(x)$. Then, it is also easy to prove that the rows of the matrix $\begin{bmatrix} A \\ \mathbf{1}^T \end{bmatrix}$ are linearly independent. To see this, let us introduce some notations. The deterministic policies $x^1, x^2, \ldots, x^N$ in the submodel B defined by x differ only at the period-state pairs $(n_1, s_1), (n_2, s_2), \ldots, (n_m, s_m)$ and the submodel allows only the actions $a^{i,1}, a^{i,2}, \ldots, a^{i,l_i}$ at (n_i, s_i) for $i = 1, 2, \ldots, m$. Then, we can naturally correspond the deterministic policies $x^1, x^2, \ldots, x^N$ in the submodel B to an element of $\prod_{i=1}^m \{1, 2, \ldots, l_i\}$. That is, $(p_1, p_2, \ldots, p_m) \in \prod_{i=1}^m \{1, 2, \ldots, l_i\}$ corresponds to the deterministic policy in B that chooses a^{i,p_i} at (n_i, s_i) for $i = 1, 2, \ldots, m$. Without loss of generality, let $x^1, x^2, \ldots, x^N$ be ordered in lexicographic order of the representation. Recall that the columns of A correspond to the deterministic policies $x^1, x^2, \ldots, x^N$. Let the columns of A be sorted so that the kth column of A corresponds to the deterministic policy x^k for $k = 1, 2, \ldots, N$. For $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, l_i - 1$, let $A_{i,j}$ be the row of A corresponding to the equality $\sum_{k=1}^N x^k_{n_i}(s_i, a^{i,j})\lambda_k = x_{n_i}(s_i, a^{i,j})$. We also sort $A_{i,j}$'s, the rows of A, in lexicographic order of their subscripts. For $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, l_i - 1$, the row $A_{i,j}$ has nonzeros in those columns that correspond to the deterministic policies choosing action $a^{i,j}$ at (n_i, s_i). Especially, the row $A_{i,j}$ has a nonzero at the column corresponding to policy $(l_1, l_2, \ldots, l_{i-1}, j, 1, 1, \ldots, 1)$. However, this policy chooses a^{k,l_k} at (n_k, s_k) for $k = 1, 2, \ldots, i - 1$ and $a^{i,j}$ at (n_i, s_i). Then, by the construction of A, we can easily show that all rows of A that are above the row $A_{i,j}$ have zeros at the column. Therefore, each row of A has a nonzero in a column at which other rows above the row have zeros. Moreover, all the rows of A have zeros at the column of policy $(l_1, l_2, \ldots, l_m)$, and thus, the row with ones has a nonzero in the column at which all rows of A have zeros. Therefore, we proved that each row of $\begin{bmatrix} A \\ \mathbf{1}^T \end{bmatrix}$ has a nonzero in a column at which any other rows above the row have zeros, and thus, the matrix have a full row rank. □

This theorem provides a way to construct $\Lambda(x)$ for a given M-randomized policy x. Define $E(x)$ as the subset of $\Lambda(x)$ whose elements has at most $M + 1$ nonzeros, then we can easily show $\Lambda(x) = \text{conv}E(x)$. By Theorem 4.3, $\Lambda(x)$ is the set of feasible solutions of a standard form LP with $M + 1$ constraints. Thus, $E(x)$ contains all extreme points of $\Lambda(x)$ and therefore, we have $\Lambda(x) = \text{conv}E(x)$. One can construct $E(x)$ by finding every representation of x as a convex

combination of $M+1$ deterministic policies among $x^1, x^2, \ldots, x^N$, which can be done by applying the procedure described in the proof of Theorem 5.1 in Feinberg and Shwartz [7] or by Algorithm 1 in [6] in a straightforward way.

4.2 Splitting into "less" randomized policies

In this section, we introduce another representation of a randomized policy as a convex combination of "less" randomized policies. How to obtain the representation for a given randomized policy is illustrated in the proofs of the three lemmas and these lemmas prove that the representation satisfies the set of properties which will help us establish necessary conditions for an extreme point of $\mathcal{P}$ in Section 5. We start with the simplest case of exactly 1-randomized policy and this is a special case of Lemma 4.2. However, its proof introduces an important technique which is used repeatedly in this section.

Lemma 4.4 *Any exactly 1-randomized policy can be uniquely represented as a non-trivial convex combination of two 0-randomized (i.e., deterministic) policies, and moreover, the weights are positive.*

Proof: Let x be an exactly 1-randomized policy. There exists a unique period-state pair (n, s) and two actions a, b such that $x_n(s, a) = \delta > 0$ and $x_n(s, b) = \epsilon > 0$ and $x_n(s, a') = 0$ for $a' \in \mathcal{A} \setminus \{a, b\}$. We show that x is a convex combination of two 0-randomized policies, denoted as w and z. To construct them, we first define two flows, u and v. In this proof and the proofs of the following two lemmas, the steps to define flows (which are u and v in this proof) are similar to the proof of Theorem 4.3 of [10] and we also borrowed their notations.

In periods $k = n+1, n+2, \ldots$, x does not randomize, thus for $s' \in \mathcal{S}$, let $a_k(s')$ denote the chosen action at (k, s') by x, *i.e.*, $x_k(s', a_k(s')) > 0$. Let $\mathcal{S}_{n+1}(x) = \{s' \in \mathcal{S} \mid p_{n-1}(s'|s, a) > 0\}$. For $k = n+2, n+3, \ldots$, recursively define $\mathcal{S}_k(x) = \{s' \in \mathcal{S} \mid p_{k-1}(s'|\tilde{s}, a_{k-1}(\tilde{s})) > 0 \text{ for some } \tilde{s} \in \mathcal{S}_{k-1}(x)\}$. That is, $\mathcal{S}_k(x)$ is the set of states in period k that receive any portion of flow δ originating in hyperarc (n, s, a) under policy x. Let $\mathcal{F}(x)$ be the sub-hypernetwork formed by the node (n, s), hyperarc (n, s, a), nodes in $\cup_{k=n+1}^{\infty} \mathcal{S}_k(x)$ and hyperarcs $\cup_{k=n+1}^{\infty} \{(k, s_k, a_k(s_k)) \mid s_k \in \mathcal{S}_k(x)\}$. We construct a flow u in $\mathcal{F}(x)$ recursively in the following way. Let the node (n, s) be a source of supply 1 and all other nodes in sub-hypernetwork $\mathcal{F}_n(x)$ have no supply. Set $u_n(s, a) = 1$, then for each $s_{n+1} \in \mathcal{S}_{n+1}(x)$, set $u_{n+1}(s_{n+1}, a_{n+1}(s_{n+1})) = p_n(s_{n+1}|s, a)$. For $k = n+2, n+3, \ldots$ and for each $s_k \in \mathcal{S}_k(x)$, set

$$u_k(s_k, a_k(s_k)) = \sum_{s_{k-1} \in \mathcal{S}_{k-1}(x)} p_{k-1}(s_k|s_{k-1}, a_{k-1}(s_{k-1})) u_{k-1}(s_{k-1}, a_{k-1}(s_{k-1})).$$

By construction, $x_n(s, a) = \delta u_n(s, a)$. Note that for hyperarcs $(k, s_k, a_k(s_k))$ in $\mathcal{F}_n(x) \setminus \{(n, s, a)\}$, $u_k(s_k, a_k(s_k))$ can be interpreted as the conditional probability of encountering hyperarc $(k, s_k, a_k(s_k))$ by following policy x, given that we encountered hyperarc (n, s, a). Fix a hyperarc $(k, s_k, a_k(s_k))$ in $\mathcal{F}_n(x) \setminus \{(n, s, a)\}$. Let A be an event of encountering the hyperarc $(k, s_k, a_k(s_k))$ by following the policy x and let B be an event of encountering hyperarc (n, s, a) by following policy x. Then, $P(A|B) = u_k(s_k, a_k(s_k))$, $P(B) = \delta$, and $P(A) = x_k(s_k, a_k(s_k))$. Therefore, we have $x_k(s_k, a_k(s_k)) = P(A) \geq P(A \cap B) = P(A|B)P(B) = \delta u_k(s_k, a_k(s_k))$.

Similarly, for $k = n+1, n+2, \ldots$, let $\mathcal{T}_k(x) \subset \mathcal{S}$ be the set of states in period k receiving any portion of flow ϵ in hyperarc (n, s, b) under policy x. For any $t_k \in \mathcal{T}_k(x)$, there exists a unique action $b_k(t_k)$ such that $x_k(t_k, b_k(t_k)) > 0$. Let $\mathcal{G}(x)$ be the sub-hypernetwork similarly defined

as $\mathcal{F}(x)$, formed by the node (n,s), hyperarc (n,s,b), nodes in $\cup_{k=n+1}^{\infty}\mathcal{T}_k(x)$ and hyperarcs in $\cup_{k=n+1}^{\infty}\{(k,t_k,b_k(t_k)) \mid t_k \in \mathcal{T}_k(x)\}$. We construct a flow v in $\mathcal{G}_n(x)$ recursively in the following way. Let the node (n,s) be a source of supply 1 and all other nodes in sub-hypernetwork $\mathcal{G}_n(x)$ have no supply. Set $v_n(s,b)=1$, then for each $t_{n+1}\in\mathcal{T}_{n+1}(x)$, set $v_{n+1}(t_{n+1},b_{n+1}(t_{n+1}))=p_n(t_{n+1}|s,b)$. For $k=n+2,n+3,\ldots$ and for each $t_k\in\mathcal{T}_k(x)$, set

$$v_k(t_k,b_k(t_k)) = \sum_{t_{k-1}\in\mathcal{T}_{k-1}(x)} p_{k-1}(t_k|t_{k-1},b_{k-1}(t_{k-1}))v_{k-1}(t_{k-1},b_{k-1}(t_{k-1})). \tag{21}$$

By construction, $x_n(s,b)=\epsilon v_n(s,b)$ and by using the same interpretation as u, we obtain $x_k(t_k,b_k(t_k))\geq \epsilon v_k(t_k,b_k(t_k))$ for any other hyperarc $(k,t_k,b_k(t_k))$ in $\mathcal{G}(x)$.
We construct a new hyperarc frequency w as follows.

$$w_k(s_k,a_k)=\begin{cases} x_k(s_k,a_k) & \text{if } (k,s_k,a_k) \text{ not in } \mathcal{F}_n(x) \text{ or } \mathcal{G}_n(x)\\ x_k(s_k,a_k)-\delta u_k(s_k,a_k) & \text{if } (k,s_k,a_k) \text{ in } \mathcal{F}_n(x)\setminus\mathcal{G}_n(x)\\ x_k(s_k,a_k)+\delta v_k(s_k,a_k) & \text{if } (k,s_k,a_k) \text{ in } \mathcal{G}_n(x)\setminus\mathcal{F}_n(x)\\ x_k(s_k,a_k)+\delta(v_k(s_k,a_k)-u_k(s_k,a_k)) & \text{if } (k,s_k,a_k) \text{ in } \mathcal{F}_n(x)\cap\mathcal{G}_n(x).\end{cases} \tag{22}$$

Since we have $x_k(s_k,a_k)\geq \delta u_k(s_k,a_k)$ for hyperarcs (k,s_k,a_k) in $\mathcal{F}_n(x)$, w is nonnegative. Note that w is obtained from x by redirecting flow δ from $\mathcal{F}_n(x)$ to $\mathcal{G}_n(x)$. Thus, w satisfies the flow balance constraints and is 0-randomized.
z is constructed similarly, by redirecting flow ϵ from $\mathcal{G}_n(x)$ to $\mathcal{F}_n(x)$. More precisely,

$$z_k(s_k,a_k)=\begin{cases} x_k(s_k,a_k) & \text{if } (k,s_k,a_k) \text{ not in } \mathcal{F}_n(x) \text{ or } \mathcal{G}_n(x)\\ x_k(s_k,a_k)+\epsilon u_k(s_k,a_k) & \text{if } (k,s_k,a_k) \text{ in } \mathcal{F}_n(x)\setminus\mathcal{G}_n(x)\\ x_k(s_k,a_k)-\epsilon v_k(s_k,a_k) & \text{if } (k,s_k,a_k) \text{ in } \mathcal{G}_n(x)\setminus\mathcal{F}_n(x)\\ x_k(s_k,a_k)+\epsilon(u_k(s_k,a_k)-v_k(s_k,a_k)) & \text{if } (k,s_k,a_k) \text{ in } \mathcal{F}_n(x)\cap\mathcal{G}_n(x).\end{cases} \tag{23}$$

By construction, z also satisfies the flow balance constraints and are 0-randomized. Moreover, $x=\frac{\delta z+\epsilon w}{\delta+\epsilon}$, *i.e.*, x is a non-trivial convex combination of two 0-randomized hyperarc frequencies and the weights are uniquely determined. □

In the above lemma, an exactly 1-randomized policy x is represented as a convex combination of two 0-randomized (thus, "less" randomized) policies z and w. The properties of the representation are that the representation of x via convex combination of z and w is unique and that the weights are positive. Using an argument similar to the above proof, we establish a general result for a finite positive integer M.

Lemma 4.5 *For any exactly M-randomized policy x, there exist $M+1$ $(M-1)$-randomized policies $x^1,x^2,\ldots,x^{M+1}$ such that x can be uniquely represented as a convex combination of $x^1,x^2,\ldots,x^{M+1}$, and moreover, the weights of the representation are positive.*

Proof: We use induction on M. For $M=1$, Lemma 4.4 suffices.

Suppose the statement holds for $M=M'-1$. Let x be an exactly M'-randomized policy. There are finitely many period-state pairs at which x randomizes; among them, let (n,s) be the period-state pair with the biggest period index (in case of tie, choose any one of them). Then, at (n,s), there exist actions $a,b^1,b^2,\ldots,b^l$ on which x has a positive flow, say $x_n(s,a)=\delta>0$ and $x_n(s,b^i)=\epsilon_i>0$ for $i=1,2,\ldots,l$. Let $\epsilon=\sum_{i=1}^{l}\epsilon_i$. We show that x is a convex combination of two $(M'-1)$-randomized

hyperarc frequencies, denoted as w and z. To define w and z, we first introduce flows u and v^i for $i = 1, 2, \ldots, l$. Construction of flow u is exactly the same as in the proof of Lemma 4.4.

For $i = 1, 2, \ldots, l$, define v^i in the same way v was defined in the proof of Lemma 4.4 except that the starting hyperarc is (n, s, b^i). More precisely, For $k = n+1, n+2, \ldots$, let $\mathcal{T}_k^i(x) \subset \mathcal{S}$ be the set of states in period k that receive any portion of flow ϵ_i originating from hyperarc (n, s, b^i) under policy x. For any $t_k \in \mathcal{T}_k^i(x)$, there exists a unique action $b_k(t_k)$ such that $x_k(t_k, b_k(t_k)) > 0$. Let $\mathcal{G}^i(x)$ be the sub-hypernetwork formed by the node (n, s), hyperarc (n, s, b^i), nodes in $\cup_{k=n+1}^{\infty} \mathcal{T}_k^i(x)$ and hyperarcs $\cup_{k=n+1}^{\infty} \{(k, t_k, b_k(t_k)) \mid t_k \in \mathcal{T}_k^i(x)\}$. Construct flow v^i as follows. Let the node (n, s) have supply 1 and all other nodes in $\mathcal{G}^i(x)$ have 0. Set $v_n^i(s, b^i) = 1$ and for each $t_{n+1} \in \mathcal{T}_{n+1}^i(x)$, set

$$v_{n+1}^i(t_{n+1}, b_{n+1}(t_{n+1})) = \sum_{i=1}^{l} p_n(t_{n+1}|s, b^i) v_n(s, b^i). \tag{24}$$

For $k = n+2, n+3, \ldots$ and for each $t_k \in \mathcal{T}_k^i(x)$, set

$$v_k^i(t_k, b_k(t_k)) = \sum_{t_{k-1} \in \mathcal{T}_{k-1}^i(x)} p_{k-1}(t_k|t_{k-1}, b_{k-1}(t_{k-1})) v_{k-1}^i(t_{k-1}, b_{k-1}(t_{k-1})). \tag{25}$$

Set $v = \sum_{i=1}^{l} \epsilon_i v^i$. Then by using the same interpretation as in the proof of Lemma 4.4, we have $x_k(t_k, b_k(t_k)) \geq \sum_{i=1}^{l} \epsilon_i v_k^i(t_k, b_k(t_k))$ for hyperarcs $(k, t_k, b_k(t_k))$ in $\cup_{i=1}^{l} \mathcal{G}^i(x)$.
Then, we construct flow w and z using (22) and (23), respectively. That is, w is obtained from x by redirecting flow δ from $\mathcal{F}(x)$ to $\mathcal{G}^i(x)$'s while maintaining the original proportion of flows in $\mathcal{G}^i(x)$'s, and z is obtained from x by redirecting total flow ϵ from $\mathcal{G}^i(x)$'s to $\mathcal{F}(x)$. By construction, w and z are nonnegative and satisfy the flow balance constraints.

Moreover, note that except at (n, s), x does not have any randomization in either $\mathcal{F}(x)$ or $\mathcal{G}^i(x)$ for $i = 1, 2, \ldots, l$. Thus, w is exactly $(M'-1)$-randomized and z is exactly $(M'-l)$-randomized. By construction, $x = \frac{\delta z + \epsilon w}{\delta + \epsilon}$, *i.e.*, x is a convex combination of two $(M'-1)$-randomized hyperarc frequencies.

By the induction hypothesis, w is uniquely represented as a convex combination of M' $(M'-2)$-randomized hyperarc frequencies, say $w^1, w^2, \ldots, w^{M'}$ with weights $\lambda_1, \lambda_2, \ldots, \lambda_{M'}$ and the weights are all positive. Thus, x is a convex combination of z and $w^1, w^2, \ldots, w^{M'}$, *i.e.*, $M'+1$ M'-randomized policies. Now we have to show that the representation of x via convex combination of z and $w^1, w^2, \ldots, w^{M'}$ is unique and all of the weights are positive. Let

$$x = \lambda_z z + \sum_{i=1}^{M'} \lambda_i w^i \tag{26}$$

$$\lambda_z + \sum_{i=1}^{M'} \lambda_i = 1 \tag{27}$$

where $\lambda_z \in [0, 1]$ and $\lambda_i \in [0, 1]$ for $i = 1, 2, \ldots, M'$. By the definitions of w and z, and the fact that w is a convex combination of $w^1, w^2, \ldots, w^{M'}$, we have $z_n(s, a) = \delta + \epsilon > 0$ and $w_n^i(s, a) = 0$ for $i = 1, 2, \ldots, M'$. However, $x_n(s, a) = \delta > 0$. Therefore, we should have $\lambda_z = \frac{\delta}{\delta + \epsilon} > 0$. From (26), we obtain

$$\sum_{i=1}^{M'} \lambda_i w^i = x - \lambda_z z = \frac{\delta z + \epsilon w}{\delta + \epsilon} - \frac{\delta z}{\delta + \epsilon} = \frac{\epsilon w}{\delta + \epsilon}.$$

Since $\frac{\epsilon}{\delta+\epsilon} > 0$, by dividing the both sides by $\frac{\epsilon}{\delta+\epsilon}$ we obtain

$$w = \sum_{i=1}^{M'} \lambda'_i w^i, \tag{28}$$

where $\lambda'_i = \frac{\delta+\epsilon}{\epsilon}\lambda_i$. From (27), we also have $\sum_{i=1}^{M'} \lambda'_i = 1$. By the induction hypothesis, there exist unique λ'_i's for $i = 1, 2, \ldots, M'$ that satisfy (28) and they are positive. Thus, there exist positive and unique λ_i for $i = 1, 2, \ldots, M'$ that satisfy (26) and (27) along with $\lambda_z = \frac{\delta}{\delta+\epsilon}$. Therefore, when x is represented as a convex combination of z and $w^1, w^2, \ldots, w^{M'}$, all of the weights should be positive and the weights are uniquely determined. By induction, the lemma is proven. □

By the above lemma, for any finite positive integer M and any exactly M-randomized policy x, we can find $M+1$ $(M-1)$-randomized policies $x^1, x^2, \ldots, x^{M+1}$ that belong to the submodel defined by x such that we can uniquely represent x as a convex combination of $x^1, x^2, \ldots, x^{M+1}$ and the weights of convex combination are positive. For ∞-randomized policies, we prove a somewhat extended result with the same properties.

Lemma 4.6 *For any ∞-randomized policy x and for any positive integer L, there exist an integer $\bar{L} \geq L$ and policies $x^1, x^2, \ldots, x^{\bar{L}}$ that belong to the submodel defined by x such that x can be uniquely represented as a convex combination of $x^1, x^2, \ldots, x^{\bar{L}}$, and moreover, the weights are positive.*

Proof: Let $\mathcal{H}(x,n) \triangleq \{(n, s', a') \mid x_n(s', a') > 0, s' \in \mathcal{S}, a' \in \mathcal{A}\}$, that is, $\mathcal{H}(x,n)$ is the set of hyperarcs used by x in period n. In addition, let $r_n(x) = |\mathcal{H}(x,n)| - |S|$, that is, the number of "additional " actions used by x compared to a deterministic policy in period n. We use induction on L. For $L = 1$, we can let $\bar{L} = L = 1$ and $x^1 = x$, then this choice satisfies the statement.

Suppose the statement holds for $L = L' - 1 \geq 1$. Let x be an ∞-randomized policy, then $\sum_{n'=1}^{\infty} r_{n'}(x) = \infty$. Let $n = \min\{\bar{n} \mid \sum_{n'=1}^{\bar{n}} r_{n'}(x) \geq L'\}$. Choose a state $s \in \mathcal{S}$ such that at period-state pair (n, s), x randomizes over multiple actions, say $a, b^1, b^2, \ldots, b^l$. Let $x_n(s, a) = \delta > 0$ and $x_n(s, b^i) = \epsilon_i > 0$ and let $\epsilon = \sum_{i=1}^{l} \epsilon_i$. We will represent x as a convex combination of two hyperarc frequencies, w and z. Again, we define flows u and v^i for $i = 1, 2, \ldots, l$ to construct w and z.

Fix i and we first define v^i. For $k = n+1, n+2, \ldots$, let $\mathcal{T}^i_k(x) \subset \mathcal{S}$ be the set of states in period k that receive any portion of flows ϵ_i originating in hyperarc (n, s, b^i) under policy x. For any $t_k \in \mathcal{T}^i_k(x)$, let $\mathcal{B}^i_k(t_k)$ be the set of actions $b_k \in \mathcal{A}$ such that $x_k(t_k, b_k) > 0$. Let $\mathcal{G}^i(x)$ be the sub-hypernetwork formed by the node (n, s), hyperarc (n, s, b^i), nodes in $\cup_{k=n+1}^{\infty} \mathcal{T}^i_k(x)$ and hyperarcs in $\cup_{k=n+1}^{\infty} \cup_{t_k \in \mathcal{T}^i_k(x)} \mathcal{B}^i_k(t_k)$. For any $m \in \mathbb{N}$, $t \in \mathcal{S}$, $b \in \mathcal{A}$, let $\phi_m(t, b) = x_m(t, b) / \sum_{b' \in \mathcal{A}} x_m(t, b')$. Then, a flow v^i is defined in the following way. Let node (n, s) have supply 1 and all other nodes in $\mathcal{G}^i(x)$ have supply 0. Set $v^i_n(s, b^i) = 1$ and for each $t_{n+1} \in \mathcal{T}^i_{n+1}(x)$ and each $b_{n+1} \in \mathcal{B}^i_{n+1}(t_{n+1})$, set

$$v^i_{n+1}(t_{n+1}, b_{n+1}) = \phi_{n+1}(t_{n+1}, b_{n+1}) p_n(t_{n+1} | s, b^i). \tag{29}$$

For $k = n+2, n+3, \ldots$ and for each $t_k \in \mathcal{T}^i_k(x)$ and $b_k \in \mathcal{B}^i_k(t_k)$, set

$$v^i_k(t_k, b_k) = \phi_k(t_k, b_k) \sum_{t_{k-1} \in \mathcal{T}^i_{k-1}(x)} \sum_{b_{k-1} \in \mathcal{B}^i_{k-1}(t_{k-1})} p_{k-1}(t_k | t_{k-1}, b_{k-1}) v^i_{k-1}(t_{k-1}, b_{k-1}). \tag{30}$$

A flow u is defined similarly in the sub-hypernetwork consisting of the node (n, s), hyperarc (n, s, a) and the part of the hypernetwork receiving any portion of the flow δ.

As in the proof of Lemma 4.5, w is obtained from x by (22), redirecting flow δ from $\mathcal{F}(x)$ to $\mathcal{G}^i(x)$'s while maintaining the original proportion of flows in $\mathcal{G}^i(x)$'s, and z is obtained from x by (23), redirecting flow ϵ from $\mathcal{G}^i(x)$'s to $\mathcal{F}(x)$. By construction, w and z satisfy the flow balance constraints, and we have $x = \frac{\delta z + \epsilon w}{\delta + \epsilon}$.

In the construction of w, the hyperarc (n, s, a) is the only randomization removed from x in periods $1, 2, \ldots, n$. Since $\sum_{n'=1}^{n} r_{n'}(x) - 1 \geq L' - 1$, w is at least $(L'-1)$-randomized. We consider the following two cases regarding the randomization of w.

If w is exactly $\bar{N}$-randomized for some finite positive integer $\bar{N} \geq L' - 1$, then by Lemma 4.5, there exists $\bar{N}+1$ $(\bar{N}-1)$-randomized policies $w^1, w^2, \ldots, w^{\bar{N}+1}$ such that w is uniquely represented as a convex combination of $w^1, w^2, \ldots, w^{\bar{N}+1}$ and the weights are positive. By arguments in the proof of Lemma 4.5, we can show that z is necessary to represent x as a convex combination of z and $w^1, w^2, \ldots, w^{\bar{N}}$ and the weight of z is $\frac{\delta}{\delta+\epsilon} > 0$. Moreover, we can also prove that all of the $\bar{N} + 1 (\geq L')$ policies, z and $w^1, w^2, \ldots, w^{\bar{N}}$ are necessary to represent x as a convex combination of them, *i.e.*, positive weights and that the weights are uniquely determined.

If w is ∞-randomized, by the induction hypothesis, there exists a positive integer $N' \geq L' - 1$ and policies $w^1, w^2, \ldots, w^{N'}$ such that w is uniquely represented as a convex combination of $w^1, w^2, \ldots, w^{N'}$ and the weights are positive. Similarly, we can show that all of z and $w^1, w^2, \ldots, w^{N'}$ are necessary to represent x as a convex combination of them and the weights are uniquely determined.

Therefore, by induction, the lemma is proven. □

5 Necessary Conditions for an Extreme Point

We now return to constrained MDPs. In this section we provide necessary conditions for a feasible solution of (P) to be an extreme point, while the next section deals with a necessary and sufficient condition. Although many researchers have studied constrained MDPs, as far as we know, algebraic characterization of extreme points of CILPs that represent constrained MDPs with countably infinite number of states was not studied before. In this section, the existence of K-randomized optimal policy, which was also proven in [7] for a more general class of constrained MDPs, is given as a corollary of one of the necessary conditions. By combining Lemma 4.5 and Lemma 4.6, for any M-randomized policy x for $K + 1 \leq M \leq \infty$ we can find N policies $x^1, x^2, \ldots, x^N$ for some finite integer $N > K + 1$ such that x can be uniquely represented as a convex combination of the N policies and the weights are positive. Using this fact, we prove the following theorem. However, note that we are using the representation in Section 4.2, thus N may not be the number of deterministic policies in the submodel defined by x and $x^1, x^2, \ldots, x^N$ belong to the submodel defined by x but may not be deterministic.

Theorem 5.1 *Any extreme point of $\mathcal{P}$ is K-randomized.*

Proof: Let x be an extreme point of $\mathcal{P}$. Suppose that x is exactly M-randomized for some $K + 1 \leq M \leq \infty$. Then by Lemma 4.5 and Lemma 4.6, there exist a positive integer $N > K + 1$, policies $x^1, x^2, \ldots, x^N$ and positive weights $\lambda_1, \lambda_2, \ldots, \lambda_N$ whose sum is one such that $x = \sum_{i=1}^{N} \lambda_i x^i$ and the weights are uniquely determined by the N policies. Note that the N policies $x^1, x^2, \ldots, x^N$ may not be feasible to (P). Consider a feasibility problem (F_1) finding a convex combination of

$x^1, x^2, \ldots, x^N$ that is feasible to (P). That is, (F_1) finds a set of nonnegative weights $\nu_1, \nu_2, \ldots, \nu_N$ that sum up to one such that $x' = \sum_{i=1}^N \nu_i x^i \in \mathcal{P}$. We can easily show that any convex combination of hyperarc frequencies is a hyperarc frequency. Thus, in order for x' to belong to $\mathcal{P}$, it only has to satisfy the inequality constraints (4), *i.e.*, for $k = 1, 2, \ldots, K$,

$$V_k \geq \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a) x'_n(s,a) = \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a) \sum_{i=1}^N \nu_i x_n^i(s,a)$$
$$= \sum_{i=1}^N \nu_i \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a) x_n^i(s,a) \triangleq \sum_{i=1}^N \nu_i D^k(x^i). \tag{31}$$

The exchange of sums is justified because $x^1, x^2, \ldots, x^N$ are hyperarc frequencies, so they satisfy (6), and thus $D^k(x^i)$ exists for $k = 1, 2, \ldots K$ and $i = 1, 2, \ldots, N$. To use matrix notation, let $D = \{D_{k,i}\} \in \mathbb{R}^{K\times N}$ where $D_{k,i} \triangleq D^k(x^i)$, $\nu = (\nu_1, \nu_2, \ldots, \nu_N)^T \in \mathbb{R}^N$, and $v = (V_1, V_2, \ldots, V_K)^T \in \mathbb{R}^K$. Then the feasibility problem (F_1) is written as

$$(F_1) \quad \min \ \mathbf{0}^T \nu \tag{32}$$
$$\text{s.t. } D\nu + t = v \tag{33}$$
$$\mathbf{1}^T \nu = 1 \tag{34}$$
$$\nu \geq 0, t \geq 0. \tag{35}$$

(F_1) is a finite LP in standard form with $K+1$ equality constraints and $N+K$ variables. Note that $\nu = \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N)^T$ is feasible to (F_1) with some slack t_λ since x is feasible to (P), and we know $\lambda_i > 0$ for $i = 1, 2, \ldots, N$. Since $N > K+1$, (λ, t_λ) is not an extreme point of (F_1). Therefore, it is a convex combination of extreme points of (F_1), say $(\nu^1, t^1), (\nu^2, t^2), \ldots, (\nu^m, t^m)$ for some positive integer m. Set $z^j \triangleq \sum_{i=1}^N \nu_i^j x^i$ for $j = 1, 2, \ldots, m$. For $j = 1, 2, \ldots, m$, z^j is feasible to (P) because (ν^j, t^j) is feasible to (F_1). Since $x = \sum_{i=1}^N \lambda_i x^i$ and λ is a convex combination of $\nu^1, \nu^2, \ldots, \nu^m$, we can easily show that x is a convex combination of $z^1, z^2, \ldots, z^m$. To see this, let

$$\lambda = \sum_{j=1}^m a_j \nu^j$$

for some $a_j \in [0, 1]$ for $j = 1, 2, \ldots, m$ such that $\sum_{j=1}^m a_j = 1$. Then

$$x = \sum_{i=1}^N \lambda_i x^i = \sum_{i=1}^N \sum_{j=1}^m a_j \nu_i^j x^i = \sum_{j=1}^m a_j \sum_{i=1}^N \nu_i^j x^i = \sum_{j=1}^m a_j z^j.$$

Any extreme point of (F_1) has at most $K+1$ nonzeros. Thus $\nu^1, \nu^2, \ldots, \nu^m$ have at most $K+1$ nonzeros whereas λ has $N > K+1$ nonzeros. Since λ is the unique weight vector to represent x via convex combination of $x^1, x^2, \ldots, x^N$, z^j for $j = 1, 2, \ldots, m$ are different from x. That is, x is a convex combination of feasible solutions of (P) that are different from x, contradicting the assumption that x is an extreme point of $\mathcal{P}$. Therefore, the theorem is proven. □

Theorem 2.1 and Theorem 5.1 lead to the following corollary.

Corollary 5.2 *(Q) has a K-randomized optimal policy.*

The existence of K-randomized optimal policy for constrained stationary MDPs with countably infinite number of states, which covers the stationary MDP counterpart of constrained nonstationary MDPs with finite state space, was also proved in [7]. However, they used a different approach based on vector optimization and geometry of the performance set where performance set is defined as the set of vectors $(C(\beta,\pi), D^1(\beta,\pi), D^2(\beta,\pi), \ldots, D^K(\beta,\pi))$ for any $\pi \in \Pi$. Our proof is conceptually simpler and gives insights on geometry of the feasible region of the CILP representation of a class of constrained MDPs with countably infinite number of states.

The next theorem shows that at an extreme point x that uses M "more" actions than a deterministic policy at least M inequality constraints (4) are binding. To illustrate this, consider a feasible set $\{(x,s) \mid Ax + s = b, x \geq 0, s \geq 0\}$ of a finite LP. At an extreme point (x,s), the number of basic variables equals the number of equality constraints. That is, in order to make one more variable in x basic, one of the basic variables in the slack s should become nonbasic. The next theorem extends this condition to the CILP (P).

Theorem 5.3 *For any integer $M \leq K$, at an extreme point of $\mathcal{P}$ that is exactly M-randomized, at least M of the inequality constraints (4) are binding.*

Proof: Let x be an extreme point of (P) that is exactly M-randomized. Suppose that only $k < M$ inequalities of (4) are binding at x. Let $x^1, x^2, \ldots, x^{M+1}$ be the $M+1$ $(M-1)$-randomized policies and λ be the weight found by Lemma 4.5. Consider a feasibility problem (F_2) which finds a convex combination of $x^1, x^2, \ldots, x^{M+1}$ that is feasible to (P). Using similar notations, (F_2) is formulated as an LP (32) through (35), but it has $M+1+K$ variables and $K+1$ equality constraints. Since x is feasible to (P), λ is feasible to (F_2) with some slack variable t_λ. Since only k of the constraints (4) are binding at x, the slack t_λ has $K-k$ nonzeros. Therefore, (λ, t_λ) has $M+1+K-k$ nonzeros and since $k < M$, we have $M+1+K-k > K+1$. This implies that (λ, t_λ) is not an extreme point of (F_2). Then, (λ, t_λ) is a convex combination of extreme points of (F_2), say $(\nu^1, s^1), (\nu^2, s^2), \ldots, (\nu^m, s^m)$ for some positive integer m. Note that this convex combination is not a trivial one. Also, note that slack variables are determined by the weight variables, *i.e.*, an equality $\lambda = \nu_j$ for some j implies $(\lambda, t_\lambda) = (\nu^j, s^j)$. Thus, $\nu^1, \nu^2, \ldots, \nu^m$ are different from λ. Set $z^j \triangleq \sum_{i=1}^N \nu_i^j x^i$ for $j = 1, 2, \ldots, m$. Then, by Lemma 4.5, $z^1, z^2, \ldots, z^m$ are different from x. Similarly to the proof of Theorem 5.1, z^j for $j = 1, 2, \ldots, m$ is feasible to (P) and x is a convex combination of $z^1, z^2, \ldots, z^m$, contradicting that x is an extreme point of $\mathcal{P}$. $\square$

6 A Necessary and Sufficient Condition for an Extreme Point

From the previous section, Theorem 5.1 and Theorem 5.3 lead to the following necessary condition for $x \in \mathcal{P}$ to be an extreme point: it should be exactly M-randomized for some $M \leq K$ and at least M of the inequality constraints should be binding at x. In this section, we establish a necessary and sufficient condition for a feasible solution to (P) to be an extreme point. We first introduce a definition of an extreme set [21].

Definition 6.1 *A convex subset E of a convex set D is called* extreme *if any representation $x = \lambda z + (1-\lambda)w$ for $0 < \lambda < 1$, with $z, w \in D$ of a point $x \in E$ implies $z, w \in E$.*

For example, a face of a polyhedron in a finite dimensional space is an extreme set of the polyhedron. A subset E of a convex set D is called *exposed* if there is a hyperplane H supporting E such that

$E = H \cap D$. In general, an exposed subset of a convex set is extreme but the converse may not hold [7].

Guided by the necessary conditions from the previous section, we consider an exactly M-randomized feasible policy x where $M \leq K$ at which M of the inequality constraints (4) are binding. Let B be the submodel defined by x. Let N be the number of deterministic policies in the submodel B and let $x^1, x^2, \ldots, x^N$ be these deterministic policies. Notice that the definition of x^i for $i = 1, 2, \ldots, N$ is different from the one in Section 5. Consider a feasibility problem (G) which finds a convex combination of $x^1, x^2, \ldots, x^N$ that is feasible to (P). Explicitly, (G) finds a set of nonnegative weights $\nu_1, \nu_2, \ldots, \nu_N$ that sum up to one such that $\sum_{i=1}^N \nu_i x^i \in \mathcal{P}$. By using the same notations of (F_1), (G) is formulated as:

$$\begin{aligned} (G)\ \min\ & \mathbf{0}^T \nu \\ \text{s.t. } & D\nu + t = v \\ & \mathbf{1}^T \nu = 1 \\ & \nu \geq 0, t \geq 0. \end{aligned}$$

We emphasize that unlike (F_1) and (F_2), (G) finds a feasible convex combination of *deterministic* policies and N is the number of *deterministic* policies in the submodel B.

By Lemma 4.2, there exists a (possibly multiple) nonnegative weight vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N)^T$ that sum up to one such that $x = \sum_{i=1}^N \lambda_i x^i$. Since x is feasible to (P), $\nu = \lambda$ is feasible to (G) with some slack variables. We can easily prove that the slack variables depend not on the weight λ but only on the policy x itself, thus let t_x denote the vector of slack variables corresponding to x. Let

$$\tilde{\Lambda}(x) = \left\{ (\lambda, t_x) \in \mathbb{R}^N \times \mathbb{R}^K_+ \mid \lambda \in \Lambda(x), t_x = v - D\lambda \right\}.$$

Note that the last K components of elements of $\tilde{\Lambda}(x)$ (the slack part) are fixed at t_x. Since x is feasible to (P), $\tilde{\Lambda}(x)$ is contained in the feasible region of (G). We state the following theorem, a proof of which will be provided later in this section.

Theorem 6.2 *A feasible exactly M-randomized policy x for some $M \leq K$ at which at least M of the inequality constraints (4) are binding is an extreme point of $\mathcal{P}$ if and only if $\tilde{\Lambda}(x)$ is an extreme set of the feasible region of (G).*

Theorem 5.1, Theorem 5.3, and Theorem 6.2 lead to the following corollary, a necessary and sufficient condition for a feasible solution of (P) to be an extreme point and the condition can be checked using the finite LP (G).

Corollary 6.3 *A feasible solution x to (P) is an extreme point of (P) if and only if it is an exactly M-randomized policy for some $M \leq K$ at which at least M of the inequality constraints (4) are binding and $\tilde{\Lambda}(x)$ is an extreme set of the feasible region of (G).*

We first illustrate the above corollary for the case of $K = 1$. For $K = 1$, the only candidates to be considered are feasible deterministic policies and feasible exactly 1-randomized policies for which the constraint (4) is binding. Let x be a feasible deterministic policy. Let t_x be the slack variable for x. Then, $\tilde{\Lambda}(x) = \{(1, t_x)\}$ and (G) is formulated as

$$\begin{aligned} \min\ & \mathbf{0}^T \nu \\ \text{s.t. } & D(x)\nu + t = V \\ & \nu = 1 \\ & \nu \geq 0, t \geq 0, \end{aligned}$$

thus (G) has a feasible region that consists of one point, $(1, t_x)$. Therefore, $\tilde{\Lambda}(x)$ is an extreme set of the feasible region of (G). Now, let x be a feasible 1-randomized policy with $t_x = 0$. There exists $\lambda \in (0,1)$ and deterministic policies x^1, x^2 such that $x = \lambda x^1 + (1-\lambda)x^2$, and we have $\tilde{\Lambda}(x) = \{(\lambda, 1-\lambda, 0)^T\}$. Since $\tilde{\Lambda}(x)$ is a singleton, it is an extreme set if and only if the point $(\lambda, 1-\lambda, 0)^T$ is an extreme point. Since $K = 1$, by dropping the constraint index k, (G) can be written as

$$\begin{aligned} \min\ & \mathbf{0}^T \nu \\ \text{s.t.}\ & D(x^1)\nu_1 + D(x^2)\nu_2 + t = V \\ & \nu_1 + \nu_2 = 1 \\ & \nu \geq 0, t \geq 0. \end{aligned}$$

Since $(\lambda, 1-\lambda, 0)$ should be feasible to the above (G), we have either $D(x^1) < V < D(x^2)$ or $D(x^2) < V < D(x^1)$ or $D(x^1) = D(x^2) = V$. The point $(\lambda, 1-\lambda, 0)$ is an extreme point if and only if the corresponding basis matrix is nonsingular, which is equivalent to $D(x^1) \neq D(x^2)$. Consequently, $\tilde{\Lambda}(x)$ is an extreme set of the feasible region of (G) if and only if either $D(x^1) < V < D(x^2)$ or $D(x^2) < V < D(x^1)$. Therefore, according to Corollary 6.3, for $K = 1$, a feasible solution x of (P) is an extreme point if and only if x is either a feasible deterministic policy or a feasible exactly 1-randomized policy such that the inequality constraint is binding at x and it is a non-trivial convex combination of two deterministic policies x^1 and x^2 for which either $D(x^1) < V < D(x^2)$ or $D(x^2) < V < D(x^1)$ holds.

To gain intuition, consider the intersection of a polyhedron and a halfspace in a finite dimensional space (Figure 1). Extreme points of the intersection of the polyhedron P and the halfspace defined by an additional constraint $\{x : d^T x \leq v\}$ are either extreme points of P that belong to the halfspace (such as x_3 in Figure 1) or points where an edge of P intersects the hyperplane defined by the halfspace (such as x' in Figure 1, which is a convex combination of adjacent extreme points x^1 and x^2 of P). Consider now an unconstrained MDP obtained by excluding the linear inequality constraint (4) from (P). A feasible solution to the unconstrained MDP is an extreme point if and only if it is a deterministic policy (Theorem 4.3 of [10]). Then, the necessary and sufficient condition for $K = 1$ shows that the characterization of extreme points of the intersection of a polyhedron and a halfspace in finite dimensional space naturally extends to $\mathcal{P}$, the intersection of the infinite dimensional feasible region of the unconstrained MDP and the set satisfying the (linear) inequality constraint.

Proof of Theorem 6.2: Suppose that $\tilde{\Lambda}(x)$ is not an extreme set of (G). Then there exist (σ, t_1) and (τ, t_2) that are feasible to (G) such that $(\theta\sigma + (1-\theta)\tau, \theta t_1 + (1-\theta)t_2) \in \tilde{\Lambda}(x)$ for some $\theta \in (0,1)$ but either $(\sigma, t_1) \notin \tilde{\Lambda}(x)$ or $(\tau, t_2) \notin \tilde{\Lambda}(x)$. Without loss of generality, suppose $(\sigma, t_1) \notin \tilde{\Lambda}(x)$. Let $z \triangleq \sum_{i=1}^N \sigma_i x^i$ and $w \triangleq \sum_{i=1}^N \tau_i x^i$, then z and w are feasible to (P). If $z = x$, then it implies $(\sigma, t_1) \in \tilde{\Lambda}(x)$ since the slack of z, t_1, should equal the slack of x. Thus, z is not equal to x. However, $x = \sum_{i=1}^N [\theta\sigma_i + (1-\theta)\tau_i] x^i = \theta z + (1-\theta)w$ and we have $\theta \in (0,1)$ and $z \neq x$. Since σ and τ are feasible to (G), z and w are feasible to (P). Therefore, x is not an extreme point of (P). We showed that if x is an extreme point, then $\tilde{\Lambda}(x)$ is an extreme set of (G).

Suppose x is not an extreme point of (P). Then there exist z and w feasible to (P) such that $x = \theta z + (1-\theta)w$ for some $\theta \in (0,1)$. Since z and w belong to the submodel defined by x, let $z = \sum_{i=1}^N \sigma_i x^i$ and $w = \sum_{i=1}^N \tau_i x^i$ for nonnegative weight vectors σ and τ such that each of them sums up to be one. Since z and w are feasible to (P), σ and τ are feasible to (G) with slack variables t_z and t_w, respectively. Since z and w are different from x, (σ, t_z) and (τ, t_w) are not in

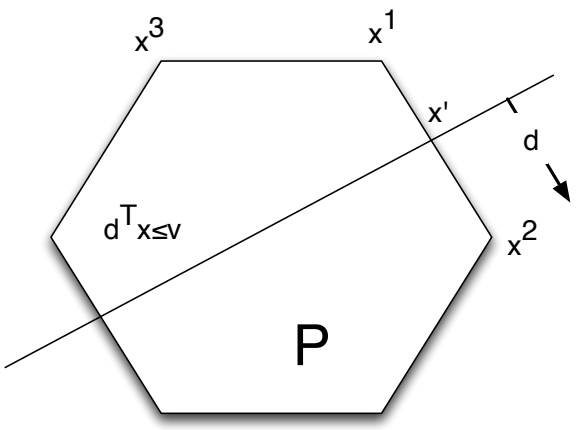

Figure 1: Extreme points for $K = 1$

$\tilde{\Lambda}(x)$. However,

$$\sum_{i=1}^{N}[\theta\sigma_i + (1-\theta)\tau_i]x^i = \theta\sum_{i=1}^{N}\sigma_i x^i + (1-\theta)\sum_{i=1}^{N}\tau_i x^i = \theta z + (1-\theta)w = x,$$

and moreover,

$$\theta t_z + (1-\theta)t_w = \theta(v - D\sigma) + (1-\theta)(v - D\tau) = v - D(\theta\sigma + (1-\theta)\tau) = t_x.$$

Therefore, $(\theta\sigma+(1-\theta)\tau, \theta t_z+(1-\theta)t_w) \in \tilde{\Lambda}(x)$ and it is a convex combination of (σ, t_z) and (τ, t_w) which are not in $\tilde{\Lambda}(x)$. That is, $\tilde{\Lambda}(x)$ is not an extreme set of the feasible region of (G). Therefore, if $\tilde{\Lambda}(x)$ is an extreme set of (G), then x is an extreme point of (P).

By combining the above arguments, the theorem is proven. □

The next example illustrates Theorem 6.2 for a kind of 2-randomized policy where $K = 2$.

Example 1 *Let $K = 2$ and consider an exactly 2-randomized policy x such that the two inequality constraints (4) are binding at x and x randomizes only at a period-state pair (n, s) over three actions, say a^1, a^2, and a^3. Then in the submodel defined by x, there are three deterministic policies, say x^1, x^2, x^3 where x^i chooses a^i at (n, s) for $i = 1, 2, 3$. Let $x = \sum_{i=1}^{3}\lambda_i x^i$ and λ has positive entries whose sum is one. Since the two inequality constraints are binding at x, its corresponding slack variables form a zero vector. We can check that $\tilde{\Lambda}(x) = \{(\lambda_1, \lambda_2, \lambda_3, 0, 0)\}$. Then, Theorem 6.2 implies that x is an extreme point of (P) if and only if $(\lambda_1, \lambda_2, \lambda_3, 0, 0)$ is an extreme point of the finite LP (G), which is equivalent to the following basis matrix being nonsingular:*

$$D_B = \begin{bmatrix} D^1(x^1) & D^1(x^2) & D^1(x^3) \\ D^2(x^1) & D^2(x^2) & D^2(x^3) \\ 1 & 1 & 1 \end{bmatrix}.$$

We can consider x^i for $i = 1, 2, 3$ as a vector in $\mathbb{R}^\infty$. Consider the subspace S of $\mathbb{R}^\infty$ spanned by x^1, x^2, x^3. We can easily show that x^1, x^2, x^3 are linearly independent, so the dimension of S is three. Define an isomorphism linear operator $T : S \to \mathbb{R}^3$ as $T(\nu_1 x^1 + \nu_2 x^2 + \nu_3 x^3) = (\nu_1, \nu_2, \nu_3)$.

Since x is a convex combination of x^1, x^2, x^3, Tx belongs to the hyperplane $\nu_1 + \nu_2 + \nu_3 = 1$ in $\mathbb{R}^3$ which we denote by P. $D^1(\cdot)$ is a linear functional on $\mathbb{R}^\infty$ and $D^1(x') = v_1$ defines a hyperplane in $\mathbb{R}^\infty$. The image of the intersection of the hyperplane and S by T can be written as $D^1(\nu_1 x^1 + \nu_2 x^2 + \nu_3 x^3) = D^1(x^1)\nu_1 + D^1(x^2)\nu_2 + D^1(x^3)\nu_3 = v_1$, which also defines a plane in $\mathbb{R}^3$ and we denote the plane in $\mathbb{R}^3$ as P^1. In the same way, we can define another plane P^2 in $\mathbb{R}^3$ which is the image of the intersection of the hyperplane $D^2(x') = v_2$ and S by T. Then, we can see that the nonsingularity of D_B is equivalent to that the planes P, P^1, and P^2 meet at one point in $\mathbb{R}^3$. However, $Tx = (\lambda_1, \lambda_2, \lambda_3)$ is on P and x also satisfies $D^1(x) = v_1$ and $D^2(x) = v_2$. Thus, if the planes P, P^1, and P^2 meet at one point, then Tx is the point. Therefore, the necessary and sufficient condition given by Theorem 6.2 is equivalent to that the planes P, P^1, and P^2 meet only at Tx.

References

[1] C. Aliprantis and K. Border. *Infinite-dimensional analysis: a hitchhiker's guide*. Springer-Verlag, Berlin, Germany, 1994.

[2] E. Altman. Denumerable constrained Markov decision processes and finite approximations. *Mathematics of Operations Research*, 19:169–191, 1994.

[3] E. Altman. *Constrained Markov decision processes*. Chapman and Hall, CRC, 1998.

[4] E. Altman and A. Shwartz. Optimal priority assignment: a time sharing approach. *IEEE Trans. on Auto. Control*, AC-34:1089–1102, 1989.

[5] E. J. Anderson and P. Nash. *Linear programming in infinite-dimensional spaces: theory and applications*. John Wiley and Sons, Chichester, UK, 1987.

[6] E. A. Feinberg and U. G. Rothblum. Splitting randomized stationary policies in total-reward Markov decision processes. *Mathematics of Operations Research*, 37:129–153, 2012.

[7] E. A. Feinberg and A. Shwartz. Constrained discounted dynamic programming. *Mathematics of Operations Research*, 21:922–945, 1996.

[8] E. B. Frid. On optimal strategies in control problems with constraints. *Theory of Probability and Its Applications*, 17:188–192, 1972.

[9] A. Ghate and R. L. Smith. Characterizing extreme points as basic feasible solutions in infinite linear programs. *Operations Research Letters*, 33:7–10, 2009.

[10] A. Ghate and R. L. Smith. A linear programming approach to nonstationary Markov decision processes. 2013. accepted to *Operations Research*.

[11] K. Golabi, R. B. Kulkarni, and G. B. Way. A statewide pavement management system. *Interfaces*, 12:5–21, 1982.

[12] D. P. Heyman and M. J. Sobel. *Stochastic models in operations research. Vol 2: Stochastic optimization*. McGraw-Hill, N.Y., 1984.

[13] A. Hordijk and F. Spieksma. Constrained admission control to a queueing system. *Adv. Appl. Probab.*, 21:409–431, 1989.

[14] L. C. M. Kallenberg. Linear programming and finite Markovian control problems. *Mathematical Centre Tracts*, 148:1–245, 1983.

[15] A. Lazar. Optimal flow control of a class of queueing networks in equilibrium. *IEEE Trans. on Auto. Control*, 28:1001–1007, 1983.

[16] P. Nain and K. W. Ross. Optimal priority assignment with hard constraint. *IEEE Trans. on Auto. Control*, 31:883–888, 1986.

[17] H. E. Romeijn, R. L. Smith, and J. Bean. Duality in infinite dimensional linear programming. *Mathematical Programming*, 53:79–97, 1992.

[18] K. W. Ross. Randomized and past-dependent policies for Markov decision processes with multiple constraints. *Operations Research*, 37:474–477, 1989.

[19] K. W. Ross and B. Chen. Optimal scheduling of interactive and noninteractive traffic in telecommunications systems. *IEEE Trans. on Auto. Control*, 33:261–267, 1988.

[20] L. I. Sennott. Constrained discounted Markov decision chains. *Probability in the Engineering and Informational Sciences*, 5:463–475, 1991.

[21] J. Stoer and C. Witzgall. *Convexity and Optimization in Finite Dimensions 1.* Springer-Verlag, New York, 1970.

[22] C. V. Winden and R. Dekker. Markov decision models for building maintenance: A feasibility study. *Journal of the Operations Research Society*, 49:928–935, 1998.