Extreme point characterization of constrained nonstationary infinite-horizon Markov decision processes with finite state space

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

[a] *University of Michigan, 1205 Beal Avenue, Ann Arbor, MI 48109-2117*

Abstract

We study infinite-horizon nonstationary Markov decision processes with discounted cost criterion, finite state space, and side constraints. This problem can equivalently be formulated as a countably infinite linear program (CILP), a linear program with countably infinite number of variables and constraints. We provide a complete algebraic characterization of extreme points of the CILP formulation and illustrate the characterization for special cases. The existence of a K-randomized optimal policy for a problem with K side constraints also follows from this characterization.

Keywords: Extreme point, Markov decision process, Constrained optimization, Countably infinite linear program

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, etc. 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. One setting where such problems often arise is data communications. In queueing systems with service rate control, the average throughput is maximized with constraints on the average delay [13, 16]. Priority queueing systems with a fixed service rate are another example [4, 18, 21]. 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, [22] 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. Examples are finding an optimal maintenance policy for each mile of a network of highways [11] and a problem in building management [24]. 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 a discounted cost criterion with nonstationary problem data and finite state space. This problem is obtained from a constrained stationary MDP with finite state space by relaxing the stationarity assumption on the problem data, which is often violated in practice. It is less obvious but still well-known that constrained nonstationary MDPs with finite state space in turn form a subclass of constrained MDPs with stationary data and countably infinite state space. A constrained nonstationary MDP with finite state space 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 [3]. Unlike finite LPs, CILPs lack a general solution method and may fail useful

*Corresponding Author: Ilbin Lee; 1205 Beal Avenue, Ann Arbor, MI 48109-2117; Phone, 1-734-679-9844

Email addresses: ilbinlee@umich.edu (Ilbin Lee), mepelman@umich.edu (Marina A. Epelman), romeijn@umich.edu (H. Edwin Romeijn), rlsmith@umich.edu (Robert L. Smith)

theoretical properties such as duality, which make them hard to analyze [5]. By Bauer's Maximum Principle [1], there exists an extreme point optimal solution for finite LPs, 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 in general [9]. In this paper we provide algebraic necessary conditions for a feasible solution of the CILP formulation of a 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, which can be checked by considering a familiar finite dimensional polyhedron. This yields a complete algebraic characterization of extreme points for CILPs representing constrained nonstationary MDPs with finite state space, and thereby provides an important building block towards the development of a simplex-type algorithm for solving 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, an often pursued goal in literature is to prove existence of an optimal policy that is as close to deterministic as possible. In particular, this means that we are interested in the existence of a 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 3). It is well-known that extreme points of LP formulations of unconstrained MDPs with a 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 point of the LP formulation of the constrained MDP is a convex combination of extreme points of the LP formulation of the unconstrained MDP, i.e., deterministic policies, and this explains how randomization is introduced. The existence of $(K+1)$-randomized optimal policy was shown for constrained MDPs with Borel state space and stationary problem data in [14] using the Carathéodory's theorem. For constrained 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 the corresponding finite LP formulation [12, 15, 20]. For constrained MDPs with a countably infinite number of states, a K-randomized optimal policy is proven to exist for the single constraint case using the Lagrangian multiplier approach in [22] and for the general case in [7] by studying the Pareto frontier of the performance set. In this paper, we obtain the existence of a K-randomized optimal policy for constrained nonstationary MDPs with finite state space as a byproduct of characterizing extreme points of the CILP formulation.

2. Problem Formulation

Consider a dynamic system operating in discrete time periods on a finite state space. In period $n \in \mathbb{N} = \{1, 2, \ldots\}$, 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, denoted by $c_n(s,a)$ and $d_n^k(s,a)$ for $k = 1, \ldots, K$, are incurred, 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 $k = 1, \ldots, K$ 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 $k = 1, \ldots, K$. The goal is to minimize the expected discounted "c-cost" satisfying K constraints on the expected discounted "d^k-costs" for $k = 1, \ldots, K$, 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^{π}_{β}. Let

$$C(\beta,\pi) \triangleq E^{\pi}_{\beta}\big[\sum_{n=1}^{\infty}\alpha^{n-1}c_n(S_n,A_n)\big],$$
$$D^k(\beta,\pi) \triangleq E^{\pi}_{\beta}\big[\sum_{n=1}^{\infty}\alpha^{n-1}d^k_n(S_n,A_n)\big] \text{ for } k=1,\ldots,K,$$

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

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

In [8] it was shown that an optimal policy for a constrained MDP may depend on the initial state; more generally, we formulate (Q) with a fixed initial state distribution β. This problem can be reformulated as a constrained stationary MDP with a 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 our 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:

$$\text{(P)} \ \min \ f(x) = \sum_{n\in\mathbb{N}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\alpha^{n-1}c_n(s,a)x_n(s,a) \tag{1}$$
$$\text{s.t.} \ \sum_{a\in\mathcal{A}} x_1(s,a) = \beta(s) \text{ for } s\in\mathcal{S} \tag{2}$$
$$\sum_{a\in\mathcal{A}} x_n(s,a) - \sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_{n-1}(s|s',a)x_{n-1}(s',a) = 0 \text{ for } n\in\mathbb{N}\setminus\{1\}, s\in\mathcal{S} \tag{3}$$
$$\sum_{n\in\mathbb{N}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\alpha^{n-1}d^k_n(s,a)x_n(s,a) \le V_k \text{ for } k=1,\ldots,K \tag{4}$$
$$x \ge 0. \tag{5}$$

If $\mathcal{P}$ denotes the feasible region of (P), constraints (2) and (3) imply that for any $x\in\mathcal{P}$,

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

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.

To gain intuition, it is convenient to interpret solutions of (P) as flows in a directed staged *hypernetwork* with infinite stages (cf. [9]). Stage n in the hypernetwork corresponds to period n of the MDP, and each stage includes S nodes, one for each state in $\mathcal{S}$. There are A directed *hyperarcs* emanating from each node, one for each action in $\mathcal{A}$; thus, a hyperarc (n,s,a) corresponds to action a in state s in stage n. A hyperarc (in a hypernetwork) can connect its "tail" node to multiple "head" nodes. In our setting, a hyperarc (n,s,a) has (n,s) as its tail node, and all nodes $(n+1,s')$ such that $p_n(s'|s,a) > 0$ (i.e., state s' in period $n+1$ is reachable on choosing action a in state s in period n) as its head nodes. If the nodes $(1,s)$ have supply of $\beta(s)$ units for $s\in\mathcal{S}$, while all other nodes have no supply or demand, any x satisfying (2), (3) and (5) can be visualized as a flow in this hypernetwork. Specifically, $x_n(s,a)$ is the flow in the hyperarc (n,s,a), and the flow reaching from node (n,s) to node $(n+1,s')$ through this hyperarc equals $p_n(s'|s,a)x_n(s,a)$. Moreover, constraints (2) and (3) ensure *flow balance* at each node. We will refer to any x satisfying (2), (3) and (5) as a *flow* in the corresponding hypernetwork. This interpretation provides particularly helpful intuition for proofs in Section 3.2.

For any Markov policy π for the nonstationary MDP with finite state space, the corresponding flow x can be found as

$$x_n(s,a) = \pi_n(a|s) \cdot P_\beta^\pi(S_n = s),\ n \in \mathbb{N},\ s \in \mathcal{S},\ a \in \mathcal{A},$$

i.e., $x_n(s,a)$ is proportional to the probability, under π, of using action a in state s in period n, scaled by the probability of reaching this state under the probability measure induced by π and β. Thus, $x_n(s,a)$ can also be interpreted as the probability of encountering hyperarc (n,s,a) under policy π for the given initial state distribution β, while the total inflow into node (n,s) is precisely $P_\beta^\pi(S_n = s)$. In light of this interpretation, a Markov policy corresponding to any flow x is also easy to identify, with the following caveat: for a given flow x, there may be some nodes (n,s) that receive no incoming flow, and thus have $\sum_{a\in\mathcal{A}} x_n(s,a) = 0$ and no outgoing flow. For those nodes, we can define $\pi_n(s)$ arbitrarily, i.e., we do not distinguish between policies that have the same corresponding flows. Under this convention, there exists a one-to-one correspondence between the set of policies and the set of flows. There also exists an obvious one-to-one correspondence between $\mathcal{P}$ and Π. We refer to a (feasible) policy and the corresponding (feasible) flow interchangeably.

Finally, the quantity $\alpha^{n-1}x_n(s,a),\ n \in \mathbb{N},\ s \in \mathcal{S},\ a \in \mathcal{A}$ can be interpreted as the occupation measure studied in [6]. The next result was shown to hold for a more general setting (*e.g.* Theorem 9 in [19]), but to make this paper self-contained, we provide the theorem and its proof.

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. Splitting Randomized Policies

Our main objective in this paper is to study extreme points of the LP formulation of constrained nonstationary MDPs with finite state space. An extreme point of a convex set is defined as a point in the set that cannot be represented as a non-trivial convex combination of other points in the set. This section provides preliminary results in the form of two different representations of a randomized MDP policy as a convex combination of other policies (the results in this section are not limited to constrained problems). These two representations will help us identify characteristics of extreme points. The first one, which describes a convenient characterization of representations of a randomized policy as a convex combination of deterministic policies, will be needed in Section 5 to derive a necessary and sufficient condition for a feasible solution of (P) to be an extreme point. The second one is needed in Section 4 to provide necessary conditions for an extreme point.

We first introduce some definitions which will be helpful in describing the two representations. Following [7], 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 $B_n(s) \subseteq \mathcal{A}$ for $n \in \mathbb{N}, s \in \mathcal{S}$. For a given policy x, we 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 $n \in \mathbb{N}, s \in \mathcal{S}$. We also say that a policy x *belongs to a submodel* if $\{a \in \mathcal{A} \mid x_n(s,a) > 0\} \subseteq B_n(s)$. For a submodel, we refer to 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 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 $S(A-1)$ "more" actions than a deterministic policy.

Definition 3.1. *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.

3.1. Splitting into deterministic policies

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

Lemma 3.2 (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 in [6] that, for any finite positive integer M, it is possible to represent an 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, and an efficient algorithm to find such a convex combination of deterministic policies was introduced.

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 5. 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 $x^1, \ldots, x^N$ be these deterministic policies. Let

$$\Lambda(x) = \left\{ \lambda \in \mathbb{R}^N \mid x = \sum_{i=1}^{N} \lambda_i x^i, \sum_{i=1}^{N} \lambda_i = 1, \lambda \geq 0 \right\}. \tag{7}$$

That is, $\Lambda(x)$ is the set of all weights of convex combinations of $x^1, \ldots, x^N$ that equal x.

In (7) $\Lambda(x)$ is defined as the set of solutions of an infinite number of linear equations, but the following theorem shows that in fact we can represent $\Lambda(x)$ with a finite number of linear equations.

Theorem 3.3. *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 $A \in \mathbb{R}^{M \times N}$ and a vector $b \in \mathbb{R}^M$ such that $\Lambda(x) = \{\lambda \in \mathbb{R}^N \mid A\lambda = b, \mathbf{1}^T\lambda = 1, \lambda \geq 0\}$, and matrix $\begin{bmatrix} A \\ \mathbf{1}^T \end{bmatrix}$ has 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.

Let $\Lambda^0(x) = \{\lambda \in \mathbb{R}^N \mid \mathbf{1}^T\lambda = 1,\ \lambda \geq 0\}$. Let (n, s) be the period-state pair with the smallest period index where x randomizes, say, over $a^1, \ldots, a^l$ for some $l \geq 2$ (in case of ties, choose one arbitrarily). Then, for any $\lambda \in \Lambda^0(x)$, $\sum_{i=1}^N \lambda_i x^i$ and x have the same flow in periods 1 to $n-1$. This implies that they also have the same flow on hyperarcs from the period-state pairs in period n where x does not randomize. Let $\Lambda^1(x) = \{\lambda \in \mathbb{R}^N \mid \sum_{i=1}^N x_n^i(s, a^1)\lambda_i = x_n(s, a^1),\ \mathbf{1}^T\lambda = 1,\ \lambda \geq 0\}$. Then, for any $\lambda \in \Lambda^1(x)$, $\sum_{i=1}^N \lambda_i x^i$ and x additionally have the same flow on the hyperarc (n, s, a^1) in period n. Let $\Lambda^{l-1}(x) = \{\lambda \in \mathbb{R}^N \mid \sum_{i=1}^N x_n^i(s, a^j)\lambda_i = x_n(s, a^j)$ for $j = 1, \ldots, l-1,\ \mathbf{1}^T\lambda = 1,\ \lambda \geq 0\}$. For any $\lambda \in \Lambda^{l-1}(x)$, $\sum_{i=1}^N \lambda_i x^i$ and x coincide in periods 1 to $n-1$ and on those hyperarcs from the period-state pairs in period n where x does not randomize, and additionally, in period n they have the same flow on hyperarcs (n, s, a^j) for $j = 1, \ldots, l-1$. Then, they should also have the same flow on hyperarc (n, s, a^l), and thus, they coincide on all hyperarcs from the node (n, s). Note that x randomizes over l actions at (n, s) and we added $l-1$ equations to obtain $\Lambda^{l-1}(x)$ from $\Lambda^0(x)$.

We can apply the same procedure repeatedly to the other period-state pairs where x randomizes in order of nondecreasing period index. Let m be the last period where x randomizes. 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 $m-1$ (which implies that they also have the same flow on all hyperarcs from any period-state pair in period m where x does not randomize). Moreover, in period m, they have the same flow on all hyperarcs from any period-state pair where x randomizes due to the added equality constraints on λ. Thus, for any $\lambda \in \Lambda^M(x)$, $\sum_{i=1}^N \lambda_i x^i$ and x coincide in all periods. We showed that $\Lambda^M(x) \subset \Lambda(x)$. We can easily see that any $\lambda \in \Lambda(x)$ satisfies all of the equalities that define $\Lambda^M(x)$. Therefore, we showed $\Lambda^M(x) = \Lambda(x)$. $\Lambda^M(x)$ is defined by $M+1$ equality constraints, and it is also easy to prove that they are linearly independent. This proves the desired result (for a more detailed proof, see Theorem 4.3 in [17]). □

This theorem provides a way to construct $\Lambda(x)$ for a given exactly M-randomized policy x. Define $E(x)$ as the subset of $\Lambda(x)$ whose elements have at most $M+1$ nonzeros; then we can easily show $\Lambda(x) = \text{conv}E(x)$. Indeed, by Theorem 3.3, $\Lambda(x)$ 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 $x^1, \ldots, x^N$, which can be done by applying the procedure described in the proof of Theorem 5.1 in [7] or Algorithm 1 in [6] in a straightforward way.

3.2. Splitting into "less randomized" policies

In this section, we introduce another representation of a randomized policy as a convex combination of "less randomized" policies. The particular representation is described in the following three lemmas, and these lemmas prove that the representation satisfies a set of properties which will help us establish necessary conditions for an extreme point of $\mathcal{P}$ in Section 4. We start in Lemma 3.4 with the simplest case of exactly 1-randomized policy. This is a special case of Lemma 3.2; however, its proof introduces an important technique which is used repeatedly in this section.

Lemma 3.4. *Any exactly 1-randomized policy can be represented as a non-trivial convex combination of two 0-randomized (i.e., deterministic) policies such that the weights of the representation are uniquely determined (by the policies) and positive.*

Proof: Let x be an exactly 1-randomized policy. There exists a unique period-state pair (n, s) and two actions a and 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 by w and z. To construct them, we first define two *sub-flows*, u and v. In this proof and the proofs of the following two lemmas, the steps to define sub-flows (here, u and v) are similar to the proof of Theorem 4.3 of [10], and we also borrowed their notation.

For $k = n+1, n+2, \ldots$, since x does not randomize in those periods, let $a_k(s')$ for $s' \in \mathcal{S}$ denote the action chosen by x at (k, s'), i.e., $x_k(s', a_k(s')) > 0$. Let $\mathcal{S}_{n+1}(x) = \{s' \in \mathcal{S} \mid p_n(s'|s, a) > 0\}$. 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), the 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 sub-flow u in $\mathcal{F}(x)$ recursively in the following way. Node (n, s) is the only source node in the sub-network, with supply of 1. Set $u_n(s, a) = 1$, and 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, we can easily see that $x_n(s, a) = \delta u_n(s, a)$ and $x_k(s_k, a_k(s_k)) \geq \delta u_k(s_k, a_k(s_k))$ for any other hyperarc $(k, s_k, a_k(s_k))$ in $\mathcal{F}(x)$.

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 and let $\mathcal{G}(x)$ be the sub-hypernetwork defined similarly to $\mathcal{F}(x)$, formed by the node (n, s), the hyperarc (n, s, b), nodes in $\cup_{k=n+1}^{\infty} \mathcal{T}_k(x)$ and the corresponding hyperarcs on which x has a positive flow. Let the node (n, s) be a source with supply 1 and construct a sub-flow v in $\mathcal{G}(x)$ similarly to construction of u. Then, it is also easy to show $x_n(s, b) = \epsilon v_n(s, b)$ and $x_k(t_k, b_k(t_k)) \geq \epsilon v_k(t_k, b_k(t_k))$ for any other hyperarc $(k, t_k, b_k(t_k))$ in $\mathcal{G}(x)$.

We construct a new flow w from x by subtracting δu in $\mathcal{F}(x)$ and adding δv in $\mathcal{G}(x)$, that is, redirecting flow δ from $\mathcal{F}(x)$ to $\mathcal{G}(x)$. z is constructed similarly, by redirecting flow ϵ from $\mathcal{G}(x)$ to $\mathcal{F}(x)$. By construction, w and z satisfy 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 flows. Note that the weights in the convex combination are both positive, and are uniquely determined (by w and z). For a more detailed proof, see [17], Lemma 4.4. □

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 in the representation of x via a convex combination of z and w, the weights are positive and uniquely determined. Using a similar argument, we establish a general result for a finite positive integer M.

Lemma 3.5. *Any exactly M-randomized policy x can be represented as a convex combination of $M+1$ $(M-1)$-randomized policies $x^1, x^2, \ldots, x^{M+1}$ such that the weights of the representation are uniquely determined and positive.*

Proof: We use induction on M. For $M = 1$, Lemma 3.4 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 largest period index (in case of tie, choose any one of them). Then, at (n, s), x randomizes over actions $a, b^1, \ldots, b^l$ for some $l \geq 1$; let $x_n(s, a) = \delta > 0$ and $x_n(s, b^i) = \epsilon_i > 0$ for $i = 1, \ldots, l$, with $\epsilon = \sum_{i=1}^{l} \epsilon_i$. We show that x is a convex combination of two $(M' - 1)$-randomized flows, denoted as w and z. To define w and z, we first introduce sub-flows u and v^i for $i = 1, \ldots, l$. Construction of sub-flow u is the same as in the proof of Lemma 3.4. For $i = 1, \ldots, l$, define v^i in the same way as v was defined in the proof of Lemma 3.4 except that the starting hyperarc is (n, s, b^i). Let $\mathcal{G}^i(x)$ denote the corresponding sub-hypernetwork. Then, we construct flow w from x by subtracting δu in $\mathcal{F}(x)$ and adding $(\delta\epsilon_i/\epsilon)v^i$ in $\mathcal{G}^i(x)$ for $i = 1, \ldots, l$. That is, w is obtained from x by redirecting flow δ from $\mathcal{F}(x)$ to $\mathcal{G}^i(x)$'s, maintaining the original proportion of flows in $\mathcal{G}^i(x)$'s. We construct flow z from x by adding ϵu in $\mathcal{F}(x)$ and subtracting $\epsilon_i v^i$ in $\mathcal{G}^i(x)$ for $i = 1, \ldots, l$. That is, 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 $i = 1, \ldots, l$. Therefore, 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 nontrivial convex combination of the two $(M' - 1)$-randomized flows.

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

$$x = \lambda_z z + \sum_{i=1}^{M'} \lambda_i w^i, \tag{8}$$

$$\lambda_z + \sum_{i=1}^{M'} \lambda_i = 1, \tag{9}$$

where $\lambda_z \in [0, 1]$ and $\lambda_i \in [0, 1]$ for $i = 1, \ldots, M'$. By the definitions of w and z and the fact that w is a convex combination of $w^1, \ldots, w^{M'}$, we have $z_n(s, a) = \delta + \epsilon > 0$ and $w^i_n(s, a) = 0$ for $i = 1, \ldots, M'$. However, $x_n(s, a) = \delta > 0$. Therefore, we should have $\lambda_z = \frac{\delta}{\delta+\epsilon} > 0$. From (8), 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

$$w = \sum_{i=1}^{M'} \lambda'_i w^i, \tag{10}$$

where $\lambda'_i = \frac{\delta+\epsilon}{\epsilon}\lambda_i$. From (9), we also have $\sum_{i=1}^{M'} \lambda'_i = 1$. By the induction hypothesis, the representation in (10) is unique and has positive weights. Thus, there exist unique and positive λ_i's for $i = 1, \ldots, M'$ that satisfy (8) and (9) along with $\lambda_z = \frac{\delta}{\delta+\epsilon}$. Therefore, representation of x as a convex combination of z and

$w^1, \ldots, w^{M'}$ is unique, and all of the weights are positive. 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 $x^1, \ldots, x^{M+1}$ that, by construction, belong to the submodel defined by x such that we can uniquely represent x as a convex combination of $x^1, \ldots, x^{M+1}$ and the weights of the convex combination are positive. Note that we can also find $M+1$ deterministic policies to represent x via a convex combination, but there may not exist $M+1$ deterministic policies such that the convex combination representation of x is *unique*, and *all the weights are positive* [6, Example 6.1]. For ∞-randomized policies, we prove a somewhat extended result with the same properties.

Lemma 3.6. *For any ∞-randomized policy x and for any positive integer L, there exist an integer $\bar{L} \geq L$ such that x can be represented as a convex combination of policies $x^1, \ldots, x^{\bar{L}}$ that belong to the submodel defined by x, such that the weights of the representation are uniquely determined and 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. Also, for any $m \in \mathbb{N}$, $t \in \mathcal{S}$, $b \in \mathcal{A}$, let $\phi_m(t,b) = x_m(t,b)/\sum_{b' \in \mathcal{A}} x_m(t,b')$, i.e., $\phi_m(t,b)$ is the probability that x will select action b in state t in period m.

We prove the lemma by induction on L. For $L = 1$, we can let $\bar{L} = L = 1$ and $x^1 = x$, and this choice satisfies the statement.

Suppose the statement holds for $L = L' - 1 \geq 1$. Since x is an ∞-randomized policy, $\sum_{n'=1}^{\infty} r_{n'}(x) = \infty$. 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 $x_n(s,a) = \delta > 0$, $x_n(s,b^i) = \epsilon_i > 0$, $i = 1, \ldots, l$, and let $\epsilon = \sum_{i=1}^{l} \epsilon_i$. We will represent x as a convex combination of two flows, w and z. Again, we first define sub-flows u and v^i for $i = 1, \ldots, l$ to construct w and z.

We first define v^i. 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 in hyperarc (n,s,b^i) under policy x. For any $t_k \in \mathcal{T}_k^i(x)$, let $\mathcal{B}_k^i(t_k)$ 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 $\cup_{k=n+1}^{\infty} \mathcal{T}_k^i(x)$, and hyperarcs in $\cup_{k=n+1}^{\infty} \cup_{t_k \in \mathcal{T}_k^i(x)} \mathcal{B}_k^i(t_k)$. Then, a sub-flow v^i is defined in the following way. Let node (n,s) be the source node with supply 1. Set $v_n^i(s,b^i) = 1$ and for each $t_{n+1} \in \mathcal{T}_{n+1}^i(x)$ and each $b_{n+1} \in \mathcal{B}_{n+1}^i(t_{n+1})$,

$$v_{n+1}^i(t_{n+1}, b_{n+1}) = \phi_{n+1}(t_{n+1}, b_{n+1}) p_n(t_{n+1}|s, b^i). \tag{11}$$

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

$$v_k^i(t_k, b_k) = \phi_k(t_k, b_k) \sum_{t_{k-1} \in \mathcal{T}_{k-1}^i(x)} \sum_{b_{k-1} \in \mathcal{B}_{k-1}^i(t_{k-1})} p_{k-1}(t_k|t_{k-1}, b_{k-1}) v_{k-1}^i(t_{k-1}, b_{k-1}). \tag{12}$$

The sub-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 3.5, w is obtained from x by redirecting flow δ from $\mathcal{F}(x)$ to $\mathcal{G}^i(x)$'s, maintaining the original proportion of flows in $\mathcal{G}^i(x)$'s, and z is obtained from x by 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 3.5, there exists $\bar{N}+1$ $(\bar{N}-1)$-randomized policies $w^1, \ldots, w^{\bar{N}+1}$ such that w is uniquely represented as a convex combination of $w^1, \ldots, w^{\bar{N}+1}$ and the weights are positive. By arguments in the proof of Lemma 3.5, we can show that

z is necessary to represent x as a convex combination of z and $w^1, \ldots, w^{\bar{N}}$ 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 $w^1, \ldots, w^{\bar{N}}$ are necessary to represent x as their convex combination, i.e., the representation has uniquely determined positive weights.

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

Therefore, by induction, the lemma is proven. □

Remark 3.7. *Since we have argued that nonstationary MDPs with finite state space can be seen as a special case of stationary MDPs with countably infinite state space, it is natural to consider extending the results presented here to the more general problem class. At present, we do not know whether the generalization is possible. The proofs in this section, which provides the foundation for the following results, have relied on the staged structure of the hypernetwork corresponding to nonstationary MDPs (e.g., by finding the smallest or the largest period index of a state at which randomization occurs), and thus are not directly extendable to the stationary case. For example, to generalize results of subsection 3.2 to stationary MDPs with countably infinite state space, for a given M- or ∞-randomized policy x, we may follow steps similar to the proofs of Lemmas 3.5 and 3.6 to obtain the split into two policies, w and z. For nonstationary MDPs with finite state space, it is clear by construction that w is $(M-1)$-randomized if x is M-randomized and that we can find w that has any number of randomizations as needed if x is ∞-randomized. However, for stationary MDPs with countably infinite state space, the randomization of w is unknown. To generalize our results, the randomization of w should be studied.*

4. 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 characterizations of extreme points of CILPs that represent constrained MDPs with countably infinite number of states have not been studied before. In this section, the existence of a K-randomized optimal policy, which was proven in [7] for a more general class of constrained MDPs, is given as a corollary of Theorem 4.1.

Combining Lemma 3.5 and Lemma 3.6 of Section 3.2, we can state that for any M-randomized policy x for $K+1 \leq M \leq \infty$ there exist a (finite) integer $N > K+1$ and N policies $x^1, \ldots, x^N$ such that $x^1, \ldots, x^N$ belong to the submodel defined by x, x can be uniquely represented as their convex combination, and the weights are positive. This fact is instrumental in proving the following theorem.

Theorem 4.1. *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 3.5 and Lemma 3.6, there exist a positive integer $N > K+1$, policies $x^1, \ldots, x^N$ and positive weights $\lambda_1, \ldots, \lambda_N$ 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 $x^1, \ldots, x^N$ may not be feasible to (P). Consider a feasibility problem (F_1) finding a convex combination of $x^1, \ldots, x^N$ that is feasible to (P). That is, (F_1) finds a set of nonnegative weights $\nu_1, \ldots, \nu_N$ 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 flows is a flow. Thus, in order for x' to belong to $\mathcal{P}$, it only has to satisfy the inequality

constraints (4), i.e., for $k = 1, \ldots, K$,

$$V_k \geq \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a) x'_n(s,a) = \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \left(\alpha^{n-1} d_n^k(s,a) \sum_{i=1}^{N} \nu_i x_n^i(s,a) \right)$$
$$= \sum_{i=1}^{N} \nu_i \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a) x_n^i(s,a) \triangleq \sum_{i=1}^{N} \nu_i D^k(x^i). \tag{13}$$

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

$$(F_1) \quad \min \mathbf{0}^T \nu \tag{14}$$
$$\text{s.t. } D\nu + t = v \tag{15}$$
$$\mathbf{1}^T \nu = 1 \tag{16}$$
$$\nu \geq 0, \ t \geq 0. \tag{17}$$

(F_1) is a finite LP in standard form with $K + 1$ equality constraints and $N + K$ variables. Note that $\nu = \lambda = (\lambda_1, \ldots, \lambda_N)^T$ is feasible to (F_1) with some slack t_λ since x is feasible to (P), and we know $\lambda_i > 0$ for $i = 1, \ldots, N$. 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 $j = 1, \ldots, m$. For $j = 1, \ldots, m$, 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 $\nu^1, \ldots, \nu^m$, we can easily show that x is a convex combination of $z^1, \ldots, z^m$.

Any extreme point of (F_1) has at most $K + 1$ nonzeros. Thus $\nu^1, \ldots, \nu^m$ have at most $K + 1$ nonzeros whereas λ has $N > K + 1$ nonzeros. Since $\lambda > 0$ is the unique weight vector to represent x via a convex combination of $x^1, \ldots, x^N$, z^j for $j = 1, \ldots, m$ 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 4.1 lead to the following corollary.

Corollary 4.2. *(Q) has a K-randomized optimal policy.*

The existence of a 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 number of states, was proven in [7]. However, the approach in [7] was based on vector optimization and geometry of the performance set, 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 into geometry of the feasible region of the CILP representation of a subclass 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, the number of non-basic (and hence 0) slack variables is greater than or equal to the number of positive components of x. Theorem 4.3 extends this condition to the CILP (P).

Theorem 4.3. *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 $\mathcal{P}$ that is exactly M-randomized. Suppose that only $k < M$ inequalities of (4) are binding at x. Let $x^1, \ldots, x^{M+1}$ be the $M + 1$ $(M - 1)$-randomized policies and λ be the weight found by Lemma 3.5. Consider a feasibility problem (F_2) which finds a convex combination of $x^1, \ldots, x^{M+1}$ that is feasible to (P). Using similar notation, (F_2) is formulated as an LP (14) through (17), 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, $(\nu^1, s^1), \ldots, (\nu^m, s^m)$ 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, $\nu^1, \ldots, \nu^m$ are different from λ. Set $z^j \triangleq \sum_{i=1}^N \nu_i^j x^i$ for $j = 1, \ldots, m$. Then, by Lemma 3.5, $z^1, \ldots, z^m$ are different from x. Similarly to the proof of Theorem 4.1, z^j for $j = 1, \ldots, m$ are feasible to (P) and x is a convex combination of $z^1, \ldots, z^m$, contradicting that x is an extreme point of $\mathcal{P}$. □

5. A Necessary and Sufficient Condition for an Extreme Point

From the previous section, Theorem 4.1 and Theorem 4.3 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 [23].

Definition 5.1. *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 with $M \leq K$ 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 submodel B and let $x^1, \ldots, x^N$ be these deterministic policies. (Notice that the policies x^i for $i = 1, \ldots, N$ considered here are different from those used in the proof of Theorem 4.1; rather, they correspond to the decomposition discussed in subsection 3.1.)

Consider a feasibility problem (G) which finds a convex combination of $x^1, \ldots, x^N$ that is feasible to (P). Explicitly, (G) finds nonnegative weights $\nu_1, \ldots, \nu_N$ that sum up to one such that $\sum_{i=1}^N \nu_i x^i \in \mathcal{P}$. By using the same notation as (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 again 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 3.2, there exists a (possibly non-unique) nonnegative weight vector $\lambda = (\lambda_1, \ldots, \lambda_N)^T \in \Lambda(x)$, where $\Lambda(x)$ is as defined in (7). Since $x = \sum_{i=1}^N \lambda_i x^i$ is feasible to (P), $\nu = \lambda$ is feasible to (G) together with some slack variables. Since the values of the slack variables are determined by x (and do not depend on a specific weights λ), 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 will be provided later in this section).

Theorem 5.2. *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 4.1, Theorem 4.3, and Theorem 5.2 lead to the following corollary, a necessary and sufficient condition for a feasible solution of (P) to be an extreme point that can be checked using the finite LP (G).

Corollary 5.3. *A feasible solution x to (P) is an extreme point of $\mathcal{P}$ 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, and let t_x be the corresponding slack variable. Then $\tilde{\Lambda}(x) = \{(1, t_x)\}$. It is easy to show that the corresponding feasibility problem (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 5.3, 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. Extreme points of the intersection are either extreme points of the polyhedron that belong to the halfspace, or points where an edge of the polyhedron intersects the hyperplane defined by the halfspace. 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 $\mathcal{P}$, which is the intersection of the infinite dimensional feasible region of the unconstrained MDP and the set satisfying the (linear) inequality constraint.

Proof of Theorem 5.2: Suppose that $\tilde{\Lambda}(x)$ is not an extreme set of the feasible region 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 $(\tau, t_2) \notin \tilde{\Lambda}(x)$, or both. Without loss of generality, suppose $(\sigma, t_1) \notin \tilde{\Lambda}(x)$. Let $z \triangleq \sum_{i=1}^{N} \sigma_i x^i$ and $w \triangleq \sum_{i=1}^{N} \tau_i x^i$; since σ and τ are feasible to (G), z and w are feasible to (P). If $z = x$, then $(\sigma, t_1) \in \tilde{\Lambda}(x)$ since the slack of z, t_1, should equal the slack of x. Thus, z is not equal to x. However, $x = \sum_{i=1}^{N} [\theta\sigma_i + (1 - \theta)\tau_i]x^i = \theta z + (1 - \theta)w$, where $\theta \in (0, 1)$ and $z \neq x$. Since both z and w are feasible to (P), x is not an extreme point of $\mathcal{P}$. We showed that if x is an extreme point, then $\tilde{\Lambda}(x)$ is an extreme set of the feasible region of (G).

For the converse, suppose x is not an extreme point of $\mathcal{P}$. Then there exist z and w feasible to (P) such that $x = \theta z + (1-\theta)w$ for some $\theta \in (0,1)$. It can be shown that z and w belong to the submodel defined by x and thus expressed as convex combinations of $x^1, \ldots, x^N$. Let these convex combinations be $z = \sum_{i=1}^{N} \sigma_i x^i$ and $w = \sum_{i=1}^{N} \tau_i x^i$. 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 $\tilde{\Lambda}(x)$. However, we can easily show that $\sum_{i=1}^{N} [\theta\sigma_i + (1-\theta)\tau_i] x^i = x$, and $\theta t_z + (1-\theta) t_w = t_x$. 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 the feasible region of (G), then x is an extreme point of $\mathcal{P}$. □

The next example illustrates Theorem 5.2 for one possible type of a 2-randomized policy and $K = 2$.

Example 1. Let $K = 2$. 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 $x = \sum_{i=1}^{3} \lambda_i x^i$ where $\lambda > 0$ and $\mathbf{1}^T \lambda = 1$. Since the two inequality constraints are binding at x, its corresponding slack variables are both 0. We can check that $\tilde{\Lambda}(x) = \{(\lambda_1, \lambda_2, \lambda_3, 0, 0)\}$. Then, Theorem 5.2 implies that x is an extreme point of $\mathcal{P}$ 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, and x^3, Tx belongs to the hyperplane $\nu_1 + \nu_2 + \nu_3 = 1$ in $\mathbb{R}^3$ which we denote by H. $D^1(\cdot)$ 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 hyperplane in $\mathbb{R}^3$ which we denote by H^1. In the same way, we can define another hyperplane H^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 the hyperplanes H, H^1 and H^2 intersecting at a single point in $\mathbb{R}^3$. However, $Tx = (\lambda_1, \lambda_2, \lambda_3)$ is in H and x also satisfies $D^1(x) = v_1$ and $D^2(x) = v_2$. Thus, if the intersection of H, H^1 and H^2 is a single point, then Tx is that point. Therefore, the necessary and sufficient condition given by Theorem 5.2 is equivalent to $H \cap H^1 \cap H^2 = Tx$.

6. Concluding Remarks

Duality, complementary slackness, and characterization of extreme points are the basic building blocks for the simplex method for finite LPs. Moreover, the CILP representation of unconstrained nonstationary MDPs, along with duality results and an algebraic characterization of extreme points, was recently studied in [10]. Based on these, a simplex algorithm for the CILP was developed and shown to achieve optimality in the limit. For constrained MDPs, duality results were provided in [3]. Definition of complementary slackness for constrained MDPs and its relation to optimality were established in [17]. In general, a simplex-type algorithm is expected to navigate through extreme points, so a complete characterization of extreme points is essential. Thus, this paper sets a foundation for developing a simplex-type algorithm for constrained nonstationary MDPs.

Acknowledgements

This work has been supported in part by the National Science Foundation grants CMMI-1333260 and CMMI-0926508. The authors thank the anonymous referee for valuable suggestions.

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. *Operations Research*, 61:413–425, 2013.

[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] G.-H. J. and O. Hernández-Lerma. Extreme points of sets of randomized strategies in constrained optimization and control problems. *SIAM Journal on Optimization*, 15:1085–1104, 2005.

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

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

[17] I. Lee, M. A. Epelman, H. E. Romeijn, and R. L. Smith. A linear programming approach to constrained nonstationary infinite-horizon Markov decision processes. IOE Technical Report TR13-01, University of Michigan, 2013.

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

[19] A. Piunovskiy. Controlled random sequences: methods of convex analysis and problems with functional constraints. *Russian Mathematical Surveys*, 53:1233–1293, 1998.

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

[21] 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.

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

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

[24] 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.