Operations Research Letters 42 (2014) 238–245

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

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

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

Industrial and Operations Engineering, University of Michigan, 1205 Beal Avenue, Ann Arbor, MI 48109-2117, United States

ARTICLE INFO

Article history:
Received 11 March 2013
Received in revised form
28 February 2014
Accepted 14 March 2014
Available online 25 March 2014

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

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.

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 others, 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 [13,15]. Priority queueing systems with a fixed service rate are another example [4,17,20]. Here, one optimizes the queueing time of non-interactive traffic while satisfying a constraint on the average end-to-end delay of interactive traffic. For these problems, [21] 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 [23]. 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 [2]. 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], 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 related literature, 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 this CILP was developed and shown to achieve optimality in the limit. For constrained MDPs, duality results were provided in [2]. Definition of complementary slackness for constrained MDPs and its

* Corresponding author.
E-mail addresses: ilbinlee@umich.edu (I. Lee), mepelman@umich.edu (M.A. Epelman), romeijn@umich.edu (H.E. Romeijn), rlsmith@umich.edu (R.L. Smith).

http://dx.doi.org/10.1016/j.orl.2014.03.001
0167-6377/

relation to optimality were established in [16]. In general, a simplex-type algorithm is expected to navigate through extreme points, so a complete characterization of extreme points is essential. 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, setting a foundation for developing a simplex-type algorithm for constrained nonstationary MDPs.

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, in a problem with K constraints, we are interested in the existence of a K-randomized optimal policy, i.e., a policy that 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. 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 the corresponding finite LP formulation [12,14,19]. For constrained MDPs with a countably infinite number of states, a K-randomized optimal policy is proven to exist for $K = 1$ using the Lagrangian multiplier approach in [21] 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 $0 \le c_n(s, a) \le c$, $0 \le d_n^k(s, a) \le d^k$ 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 by E_β^π. Let

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

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

and let $\Pi \triangleq \{\pi \mid D^k(\beta, \pi) \le V_k \text{ for } k = 1, \ldots, K\}$. 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 [2] 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 [3,2], which can be written as:

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

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

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

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

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

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

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

From (5) 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 c- and d-costs 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; here, a hyperarc (n, s, a) has (n, s) as its tail node, and all nodes $(n + 1, s')$ such that $p_n(s'|s, a) > 0$ 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 $x_n(s, a) = \pi_n(a|s) \cdot P_\beta^\pi(S_n = s)$ for $n \in \mathbb{N}$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, i.e., $x_n(s, a)$ is