Operations Research Letters 45 (2017) 570–574

Contents lists available at ScienceDirect

Operations Research Letters

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

Linear programming formulation for non-stationary, finite-horizon Markov decision process models

Arnab Bhattacharya, Jeffrey P. Kharoufeh *

Department of Industrial Engineering, University of Pittsburgh, 1025 Benedum Hall, 3700 O'Hara Street, Pittsburgh, PA 15261, USA

ARTICLE INFO

Article history:
Received 2 March 2017
Received in revised form 22 July 2017
Accepted 5 September 2017
Available online 18 September 2017

Keywords:
Non-stationary MDP model
Linear programming

ABSTRACT

Linear programming (LP) formulations are often employed to solve stationary, infinite-horizon Markov decision process (MDP) models. We present an LP approach to solving non-stationary, finite-horizon MDP models that can potentially overcome the computational challenges of standard MDP solution procedures. Specifically, we establish the existence of an LP formulation for risk-neutral MDP models whose states and transition probabilities are temporally heterogeneous. This formulation can be recast as an approximate linear programming formulation with significantly fewer decision variables.

1. Introduction

It is well known that stationary Markov decision process (MDP) models can be reformulated as linear programs and solved efficiently using linear programming (LP) algorithms [2,17–19]. The appeal of the LP formulation stems from the fact that it allows for the inclusion of additional model constraints and facilitates sensitivity analysis in sequential decision making problems. Furthermore, duality theory allows one to characterize the optimal decisions in a MDP model via the optimal solution of the associated dual problem [18,19]. Owing to recent advances in the computational speed of LP solvers, the LP approach has been successfully employed to solve large-scale, stationary MDP models (see [1,4,5,15,17,21]).

While the LP reformulation has been most prevalent in the case of stationary, infinite-horizon MDP models [6,18], by contrast, this formulation is seldom used as a solution strategy for non-stationary, finite-horizon MDP models. This is due to the ease of implementation of the backward dynamic programming (BDP) procedure to solve finite-horizon problems as a sequence of simpler single-stage problems using the optimality equations. It is well known that, for an N-stage MDP model with K states and L actions in each stage, the BDP procedure requires $(N-1)LK^2$ multiplicative operations to determine an optimal policy [18]. The BDP procedure is computationally viable for models with low-dimensional state and action spaces, as the number of such operations is relatively small. However, for models with multidimensional state and action spaces, BDP becomes computationally intractable due to the curses of dimensionality [2,17]. The problem is further exacerbated for models with a large number of decision stages. For example, solving a 10-stage model comprised of four state and three action variables, each with five feasible values in each stage, requires more than 2.2×10^9 multiplicative operations, which is prohibitively large. The computational burden may increase further for *non-stationary* MDP models that include temporal heterogeneity in the states and transition probabilities.

In this paper, we present a linear programming formulation for non-stationary, finite-horizon MDP models as a viable approach to overcome these computational challenges. Specifically, we prove the existence of a general LP formulation for such models with countable state and action spaces under a risk-neutral objective. We establish lower and upper bounds of the value functions, which are used to formulate the primal LP model. The solution of this model is the value function of the MDP model, while the solution of its dual problem recovers the optimal policy. Although the LP approach does not (in and of itself) overcome the curses of dimensionality, it lays the groundwork for implementing approximate linear programming (ALP) procedures [4] to solve computationally intractable finite-horizon models. Specifically, we suggest an ALP formulation that utilizes parametric basis functions to approximate the value functions at each stage. In light of recent advances in LP solvers, the ALP approach offers computational advantages over traditional MDP solution procedures (such as the value and policy iteration algorithms) for solving high-dimensional finite-horizon problems.

The remainder of the paper is organized as follows. Section 2 introduces some preliminaries of the non-stationary, finite-horizon MDP model, while Section 3 presents our main results which establish existence of the LP formulation. In Section 4, we discuss the computational advantages of using the LP approach as compared to

* Corresponding author.
E-mail addresses: arb141@pitt.edu (A. Bhattacharya), jkharouf@pitt.edu (J.P. Kharoufeh).

http://dx.doi.org/10.1016/j.orl.2017.09.001
0167-6377/

standard MDP solution procedures. Some concluding remarks are provided in Section 5.

2. Preliminaries

Consider a finite planning horizon $T = \{1, 2, \ldots, N\}$ with N decision stages (or decision epochs) and let $t \in T$ be the index of the tth decision epoch. For convenience, define $T' \equiv T \setminus \{N\}$. In what follows, all random variables are defined on a common, complete probability space $(\Omega, \mathscr{A}, \mathbb{P})$, where Ω is a sample space, $\mathscr{A}$ is a σ-field of subsets of Ω and $\mathbb{P}$ is a probability measure on $(\Omega, \mathscr{A})$. In what follows, all vectors are assumed to be column vectors, unless otherwise noted. The state of the process at the start of stage t is denoted by the random vector $\boldsymbol{S}_t$ whose state space is a countable set $\mathcal{S}_t \subset \mathbb{R}^n$. A realization of $\boldsymbol{S}_t$ is denoted by $\boldsymbol{s} \in \mathcal{S}_t$. When the state of the process is $\boldsymbol{s}$, the set of feasible decisions (or action space) is denoted by a countable set $\mathcal{X}_t(\boldsymbol{s}) \subset \mathbb{R}^m$. For notational convenience, we suppress the dependence of this set on $\boldsymbol{s}$ and simply write $\mathcal{X}_t$. A decision rule is a vector-valued mapping $\boldsymbol{x}_t : \mathcal{S}_t \to \mathcal{X}_t$ that prescribes feasible actions for each $\boldsymbol{s} \in \mathcal{S}_t$. A set of decision rules, one for each stage $t \in T'$, is called a policy and is denoted by $\pi = \{\boldsymbol{x}_t : t \in T'\} \in \Pi$, where Π is the collection of all feasible Markov deterministic (MD) policies. It is noted that no decisions are made in the terminal stage $t = N$. For a given decision rule $\boldsymbol{x}_{t-1}$, the temporally-heterogeneous transition probabilities are denoted by $\mathbb{P}_t(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{x}_{t-1}(\boldsymbol{s}))$, where $(\boldsymbol{s}', \boldsymbol{s}) \in \mathcal{S}_t \times \mathcal{S}_{t-1}$ and $\boldsymbol{x}_{t-1}(\boldsymbol{s}) \in \mathcal{X}_{t-1}$. Let $p : \mathcal{S}_1 \to [0, 1]$ be the probability mass function of the initial state $\boldsymbol{S}_1$ such that $0 \leq p(\boldsymbol{s}) \leq 1$ for all $\boldsymbol{s} \in \mathcal{S}_1$ and $\sum_{\boldsymbol{s} \in \mathcal{S}_1} p(\boldsymbol{s}) = 1$. For a given policy π, the transition probability matrix at stage t, denoted by $\boldsymbol{Q}_t^\pi$, is defined as

$$\boldsymbol{Q}_t^\pi = \left(\mathbb{P}_t(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{x}_{t-1}(\boldsymbol{s})) : (\boldsymbol{s}', \boldsymbol{s}) \in \mathcal{S}_t \times \mathcal{S}_{t-1}, \boldsymbol{x}_{t-1} \in \pi\right), \quad t = 2, \ldots, N,$$

where $\sum_{\boldsymbol{s}' \in \mathcal{S}_t} \mathbb{P}_t(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{x}_{t-1}(\boldsymbol{s})) = 1$ for all $\boldsymbol{s} \in \mathcal{S}_{t-1}$. The random one-step cost incurred in stage $t \in T'$ is denoted by $c_t(\boldsymbol{S}_t, \boldsymbol{x}_t(\boldsymbol{S}_t))$, while the terminal cost is $c_N(\boldsymbol{S}_N)$. For a given policy π, the vector of one-step costs at stage t is

$$\boldsymbol{c}_t^\pi = \begin{cases} (c_t(\boldsymbol{s}, \boldsymbol{x}_t(\boldsymbol{s})) : \boldsymbol{s} \in \mathcal{S}_t, \boldsymbol{x}_t \in \pi), & t \in T', \\ (c_N(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}_N), & t = N, \end{cases}$$

where we assume that $|c_t(\boldsymbol{s}, \boldsymbol{x}_t(\boldsymbol{s}))| < \infty$ and $|c_N(\boldsymbol{s})| < \infty$. Consider a mapping $V_t^\pi : \mathcal{S}_t \to \mathbb{R}$, where $V_t^\pi(\boldsymbol{s})$ denotes the expected future cost incurred under policy π starting in state $\boldsymbol{s}$ at stage t, and let $\boldsymbol{V}_t^\pi \equiv (V_t^\pi(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}_t)$. By definition,

$$\boldsymbol{V}_t^\pi = \boldsymbol{c}_t^\pi + \delta \boldsymbol{Q}_{t+1}^\pi \boldsymbol{V}_{t+1}^\pi, \quad t \in T',$$
$$\boldsymbol{V}_N^\pi = \boldsymbol{c}_N,$$

where $\delta \in (0, 1]$ is a discount factor. Given an initial state $\boldsymbol{s}$, the risk-neutral objective is to minimize the expected total discounted costs incurred over the planning horizon as follows:

$$z^*(\boldsymbol{s}) = \inf_{\pi \in \Pi} \left\{\boldsymbol{V}_1^\pi(\boldsymbol{s})\right\} = \inf_{\pi \in \Pi} \left\{\mathbb{E}_\pi \left(\sum_{t \in T'} \delta^{t-1} c_t(\boldsymbol{S}_t, \boldsymbol{x}_t(\boldsymbol{S}_t)) + \delta^{N-1} c_N(\boldsymbol{S}_N) \middle| \boldsymbol{S}_1 = \boldsymbol{s} \right)\right\}. \quad (1)$$

We denote an optimal policy of (1) by π^* and the corresponding value function at stage t by $V_t^* \equiv V_t^{\pi^*}$. Let $\boldsymbol{V}_t^* = (V_t^*(\boldsymbol{s}_t) : \boldsymbol{s}_t \in \mathcal{S}_t)$ be the vector of optimal values in stage t. Then, the optimality equations (in vector form) are given by

$$\boldsymbol{V}_t^* = \inf_{\pi \in \Pi} \left\{\boldsymbol{c}_t^\pi + \delta \boldsymbol{Q}_{t+1}^\pi \boldsymbol{V}_{t+1}^*\right\}, \quad t \in T', \quad (2a)$$
$$\boldsymbol{V}_N^* = \boldsymbol{c}_N. \quad (2b)$$

3. Linear programming formulation

In this section, we establish the existence of an LP formulation for the model in (1) whose optimal solutions are the value functions defined in (2). Additionally, we present an associated dual LP formulation whose optimal solutions can be used to obtain an optimal policy π^*.

Let $\mathbb{V}$ denote the set of all real-valued, bounded functions on $\mathcal{S}_t$. For each $t \in T$, consider a complete, normed linear space $(\mathbb{V}, \|\cdot\|_\infty)$ of bounded functions on $\mathcal{S}_t$ that is equipped with the supremum norm $\|\cdot\|_\infty$ and component-wise partial order $\preceq$. Let $J_t : \mathcal{S}_t \to \mathbb{R}$ be a function that belongs to $\mathbb{V}$, and let $\boldsymbol{J}_t = (J_t(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}_t)$ be the vector form of J_t so that its supremum norm is $\|\boldsymbol{J}_t\|_\infty = \sup_{\boldsymbol{s} \in \mathcal{S}_t} \{|J_t(\boldsymbol{s})|\}$. Moreover, for any two functions $J_t^1, J_t^2 \in \mathbb{V}$, the relation $J_t^1 \preceq J_t^2$ implies that $J_t^1(\boldsymbol{s}) \leq J_t^2(\boldsymbol{s})$ for all $\boldsymbol{s} \in \mathcal{S}_t$, or simply that $\boldsymbol{J}_t^1 \leq \boldsymbol{J}_t^2$. For each $t \in T'$, define a nonlinear operator $\Lambda_t : \mathbb{V} \to \mathbb{V}$, such that for any $J_{t+1} \in \mathbb{V}$,

$$\Lambda_t \boldsymbol{J}_{t+1} = \inf_{\pi \in \Pi} \left\{\boldsymbol{c}_t^\pi + \delta \boldsymbol{Q}_{t+1}^\pi \boldsymbol{J}_{t+1}\right\}. \quad (3)$$

For stage N, define another operator, $\Psi : \mathbb{V} \to \mathbb{V}$, such that for all $\boldsymbol{J}_N \in \mathbb{V}$,

$$\Psi \boldsymbol{J}_N = \boldsymbol{c}_N. \quad (4)$$

Thus, the operator Ψ maps any bounded function in stage N to the terminal cost function in stage N so that $\Psi \boldsymbol{J}_N(\boldsymbol{s}) = \boldsymbol{c}_N(\boldsymbol{s})$ for each $\boldsymbol{s} \in \mathcal{S}_N$. Next, denote an arbitrary vector of functions, one for each stage $t \in T$, by $\boldsymbol{J} = (\boldsymbol{J}_t : t \in T) \in \mathbb{V}^N$ and define the operator $\Lambda : \mathbb{V}^N \to \mathbb{V}^N$ such that for any $\boldsymbol{J} \in \mathbb{V}^N$,

$$\Lambda \boldsymbol{J} = (\Lambda_1 \boldsymbol{J}_2, \Lambda_2 \boldsymbol{J}_3, \ldots, \Lambda_{N-1} \boldsymbol{J}_N, \Psi \boldsymbol{J}_N). \quad (5)$$

A fixed point of the operator Λ is any vector $\boldsymbol{J}^* \in \mathbb{V}^N$ satisfying the equality

$$\Lambda \boldsymbol{J}^* = \boldsymbol{J}^*. \quad (6)$$

Any $\boldsymbol{J} \in \mathbb{V}^N$ that satisfies the inequality $\boldsymbol{J} \leq \Lambda \boldsymbol{J}$ is called a sub-solution of (6), while a super-solution of (6) satisfies $\boldsymbol{J} \geq \Lambda \boldsymbol{J}$. Proposition 1 shows that sub- and super-solutions of (6) are, respectively, lower and upper bounds of the value function vector $\boldsymbol{V}^* = (\boldsymbol{V}_t^* : t \in T)$.

Proposition 1. *For any $\boldsymbol{J} \in \mathbb{V}^N$, if $\boldsymbol{J} \leq \Lambda \boldsymbol{J}$ ($\boldsymbol{J} \geq \Lambda \boldsymbol{J}$), then $\boldsymbol{J} \leq \boldsymbol{V}^*$ ($\boldsymbol{J} \geq \boldsymbol{V}^*$).*

Proof. Let $\bar{\pi}$ be a feasible policy of (1). First consider the case $\boldsymbol{J} \leq \Lambda \boldsymbol{J}$ for some $\boldsymbol{J} \in \mathbb{V}^N$. In this case, $\boldsymbol{J}_t \leq \Lambda_t \boldsymbol{J}_{t+1}$, for each $t \in T'$ and $\boldsymbol{J}_N \leq \Psi \boldsymbol{J}_N = \boldsymbol{c}_N$. Using equalities (3) and (4), respectively, we obtain the following system of inequalities:

$$\boldsymbol{J}_t \leq \inf_{\pi \in \Pi} \{\boldsymbol{c}_t^\pi + \delta \boldsymbol{Q}_{t+1}^\pi \boldsymbol{J}_{t+1}\} \leq \boldsymbol{c}_t^{\bar{\pi}} + \delta \boldsymbol{Q}_{t+1}^{\bar{\pi}} \boldsymbol{J}_{t+1}, \quad t \in T', \quad (7a)$$
$$\boldsymbol{J}_N \leq \boldsymbol{c}_N = \boldsymbol{V}_N^*. \quad (7b)$$

The right-most inequality in (7a) holds because $\bar{\pi}$ is feasible, but not necessarily optimal, for $\boldsymbol{J}_{t+1}$ in (3). Starting in stage 1 and sequentially applying constraints (7) for stages $t = 2, \ldots, N$, we have

$$\begin{aligned} \boldsymbol{J}_1 &\leq \boldsymbol{c}_1^{\bar{\pi}} + \delta \boldsymbol{Q}_2^{\bar{\pi}} \boldsymbol{J}_2, \\ &\leq \boldsymbol{c}_1^{\bar{\pi}} + \delta \boldsymbol{Q}_2^{\bar{\pi}} (\boldsymbol{c}_2^{\bar{\pi}} + \delta \boldsymbol{Q}_3^{\bar{\pi}} \boldsymbol{J}_3) = \boldsymbol{c}_1^{\bar{\pi}} + \delta \boldsymbol{Q}_2^{\bar{\pi}} \boldsymbol{c}_2^{\bar{\pi}} + \delta^2 \boldsymbol{Q}_2^{\bar{\pi}} \boldsymbol{Q}_3^{\bar{\pi}} \boldsymbol{J}_3, \\ &\vdots \\ &\leq \boldsymbol{c}_1^{\bar{\pi}} + \delta \boldsymbol{Q}_2^{\bar{\pi}} \boldsymbol{c}_2^{\bar{\pi}} + \cdots + \delta^{N-1} \left(\prod_{t \in T'} \boldsymbol{Q}_{t+1}^{\bar{\pi}} \right) \boldsymbol{c}_N \\ &= \boldsymbol{J}_1^{\bar{\pi}}, \end{aligned}$$