Computers & Industrial Engineering 87 (2015) 311–316

Contents lists available at ScienceDirect

Computers & Industrial Engineering

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

Optimally solving Markov decision processes with total expected discounted reward function: Linear programming revisited

Oguzhan Alagoz [a,*], Mehmet U.S. Ayvaci [b], Jeffrey T. Linderoth [a]

[a] Department of Industrial and Systems Engineering, University of Wisconsin, Madison, WI, United States
[b] Jindal School of Management, University of Texas at Dallas, Dallas, TX, United States

ARTICLE INFO

Article history:
Received 16 February 2015
Received in revised form 22 May 2015
Accepted 23 May 2015
Available online 29 May 2015

Keywords:
Markov decision process
MDP
Linear programming
Policy iteration
Total expected discounted reward
Treatment optimization

ABSTRACT

We compare the computational performance of linear programming (LP) and the policy iteration algorithm (PIA) for solving discrete-time infinite-horizon Markov decision process (MDP) models with total expected discounted reward. We use randomly generated test problems as well as a real-life health-care problem to empirically show that, unlike previously reported, barrier methods for LP provide a viable tool for optimally solving such MDPs. The dimensions of comparison include transition probability matrix structure, state and action size, and the LP solution method.

1. Introduction

Discrete-time Markov decision processes (MDPs) provide a natural framework for modeling sequential decision-making problems under uncertainty. MDPs have been applied to many areas of application including finance, logistics, manufacturing, and recently in health-care. While MDPs can be analyzed to prove the existence of certain structured policies, closed-form solutions typically exist under very restrictive conditions, therefore, most researchers solve them numerically.

There are three common methods for optimally solving MDPs with total expected discounted rewards, the most commonly used MDPs. These methods are linear programming (LP), the policy iteration algorithm (PIA), and the value iteration algorithm (VIA). Littman, Dean, and Kaelbling (1995) show that the MDP problem is polynomial in the number of states, the size of action space and the maximum number of bits required to encode immediate rewards and state-transition probabilities as rational numbers. More recently, Ye (2015) proves that PIA is a strongly polynomial-time algorithm for solving MDPs with total expected discounted reward.

It has been proven that PIA is faster than VIA. The convergence rate of VIA is linear, whereas the convergence rate of PIA is superlinear (quadratic in most cases) (Puterman, 1994). As Puterman (1994) notes "one should never use the value iteration algorithm" unless there exists a structured optimal policy that increases the speed of the VIA (such as a lower and upper bound on the optimal value function). While modified policy iteration algorithm (a variant of VIA) (Morton, 1971; Puterman & Shin, 1978) has been noted to have superior performance over PIA and hence VIA for some test problems, a search of the literature shows that the most commonly used methods for solving MDPs are PIA and VIA. Note that unlike PIA and LP, both VIA and the modified policy iteration algorithms generate an ϵ-optimal policy, therefore, ϵ should be selected sufficiently small to ensure that the resulting policy is exactly optimal. Surprisingly, LP has not been very popular for solving MDPs despite recent research in improving the efficiency of the LPs and commercially available software.

For instance, we searched all articles published in the EI-Compendex® database from 2005 until 2013 and identified all articles with a keyword "Markov decision processes." We found that among studies that solved a MDP model using total expected discounted reward criterion and reported the solution method, 19 of them used VIA, 8 used PIA and only 6 used LP as the solution method. Table 1 lists these studies.

The purpose of this paper is to compare the computational performance of LP and PIA and empirically show that LP is a viable tool to optimally solve MDPs. It is known that PIA is equivalent to the simplex method for solving linear programs with full pivotal operations (Kallenberg, 1983). Previous literature on comparing LP to

* Corresponding author.
E-mail addresses: alagoz@engr.wisc.edu (O. Alagoz), Mehmet.Ayvaci@utdallas.edu (M.U.S. Ayvaci), linderoth@cae.wisc.edu (J.T. Linderoth).

http://dx.doi.org/10.1016/j.cie.2015.05.031
0360-8352/

Table 1
Solution methods used in infinite-horizon total expected discounted MDP articles published in INSPEC database between 2005 and 2012.

Author	Year	Method
Buongiorno and Zhou (2011)	2011	LP
Wang and Schonfeld (2010)	2010	LP
Farran and Zayed (2009)	2009	LP
Grizzle et al. (2008)	2008	LP
Bello and Riano (2006)	2006	LP
Le Ny and Feron (2006)	2006	LP
Sun et al. (2011)	2011	PIA
Sandıkçı et al. (2008)	2008	PIA
Alagoz et al. (2007a)	2007	PIA
Alagoz et al. (2007b)	2007	PIA
Idoumghar and Schott (2006)	2006	PIA
Kuppuswamy and Lee (2005)	2005	PIA
Chang and Chong (2005)	2005	PIA
Mosharaf et al. (2005)	2005	PIA
Flapper et al. (2012)	2012	VIA
Rezaei Yousefi et al. (2012)	2012	VIA
Viet et al. (2012)	2012	VIA
Arruda et al. (2011)	2011	VIA
Chen and Liu (2011)	2011	VIA
Sharna et al. (2011)	2011	VIA
Al-Zubaidy et al. (2010)	2010	VIA
Asadian et al. (2010)	2010	VIA
Kurt and Kharoufeh (2010)	2010	VIA
Min and Yih (2010)	2010	VIA
Farrokh et al. (2009)	2009	VIA
Akselrod and Kirubarajan (2008)	2008	VIA
Stevens-Navarro et al. (2008)	2008	VIA
Al-Zubaidy et al. (2007)	2007	VIA
Chen and Cheng (2007)	2007	VIA
Agrawal et al. (2007)	2007	VIA
Chamberland et al. (2007)	2007	VIA
Glazebrook et al. (2005)	2005	VIA
Zobel and Scherer (2005)	2005	VIA

PIA notes that LP is slower than PIA. In particular, in one of the few studies comparing the two methods, Littman et al. (1995) note that "While progress has been made on speeding up linear programming algorithms, MDP-specific algorithms (such as PIA) hold more promise for efficient solution." They further note that "more empirical study is needed" to determine which algorithm is better for optimally solving MDPs.

On the other hand, it is well-known that advances in computational LP over the past 15 years have resulted in extremely fast and robust packages. Bob Bixby, the main architect of the CPLEX linear programming software, has performed careful studies that conclusively demonstrate that for many large-scale and sparse problems (like those LPs arising from solving MDPs), LP algorithms often run *thousands of times faster* than a decade ago (Bixby, 2002). This multiple order of magnitude difference is *solely* from improved algorithms, as Bixby compared different versions of the CPLEX code on the same (modern) computing hardward. As Puterman (1994) noted "At present, LP has not been proven to be an efficient method for solving large discounted MDPs; however, innovations in LP algorithms in the past decade might change this." Similarly, and more recently, Powell states that "given the dramatic strides in the speed of linear programming solvers over the past decade, the relative performance of value iteration over the linear programming method is an unresolved question." (Powell, 2007). Our work provides evideence for answering this unresolved question, performing a careful computational experiment comparing PIA (typically superior to VIA) to LP.

We focus on comparing PIA and LP particularly for MDPs with transition probability matrices populated around the diagonal entries, a common structure observed in MDPs developed for treatment optimization problems (Alagoz, Maillart, Schaefer, & Roberts, 2004; Alagoz, Maillart, Schaefer, & Roberts, 2007a, Alagoz, Maillart, Schaefer, & Roberts, 2007b; Kurt, Denton, Schaefer, Shah, & Smith, 2011; Sandıkçı, Maillart, Schaefer, Alagoz, & Roberts, 2008; Schaefer, Bailey, Shechter, & Roberts, 2004; Shechter, Bailey, Schaefer, & Roberts, 2008). In these MDP models, the state space typically represents the health state of the patient and the transition probabilities determine how the patient health changes over time. Because patient health does not change dramatically in a single decision epoch, such MDPs have sparse transition probability matrices with nonzero entries densely populated around the diagonals. For such problems, LPs may provide a useful tool since commercially available software typically takes the advantage of the structure of the constraint matrix.

In this study, we use many randomly generated test problems to empirically compare the performance of LP and PIA. We focus our attention to PIA instead of VIA since it has been shown to converge to the optimal solution faster than VIA. Furthermore, the selection of ϵ is critical for comparing the performance of VIA and LP. Unlike previous researchers who note that LPs are slower than PI, we find that the performance of LP is superior to PIA in most of the test problems. Furthermore, we test the effects of transition probability matrix structure (such as matrices with a banded-diagonal structure) on the performance of LPs. We conclude with a comparison of the performance of PIA and LP on a real-life MDP model that optimizes colonoscopy screening decisions for early diagnosis of colorectal cancer.

The remainder of this paper is organized as follows: In Section 2, we present a formal definition of the MDPs under consideration and describe LP and PIA for optimally solving them. We provide our computational experiments and results in Section 3, and in Section 4, we discuss our findings and conclusions.

2. MDPs and solution algorithms

We refer to the collection of objects, $(T, S, A_s, P(.|s,a), r(s,a))$ as a discrete-time infinite-horizon MDP with stationary rewards and transition probabilities, where $T = \{1, 2, \ldots\}$ represents the decision epochs, S represents the *state space*, A_s represents the *action space* for state $s \in S$; $P(.|s,a)$ represents the *transition probabilities* for a given state s and action $a \in A_s$; and $r(s,a)$ is the immediate *reward function* for state s and action a.

A *decision rule* specifies the action selection from A_s for each state, i.e., $d(s) \in A_s$. We can drop the index s from this expression and use $d \in A$ to represent a decision rule specifying the actions to be taken at all states, where $A = \cup_{s \in S} A_s$ is the set of all actions. A *policy* δ is a sequence of decision rules to be used at each decision epoch.

We consider an infinite-horizon discounted MDP with a discount factor λ, where $0 \leqslant \lambda < 1$. The optimal policy of such an MDP can be found by solving the following optimality equations (Puterman, 1994):

$$V(s) = \sup_{a \in A_s} \left\{ r(s,a) + \sum_{j \in S} \lambda P(j|s,a) V(j) \right\} \text{ for } s \in S, \tag{1}$$

where $V(s)$ is the optimal value of the MDP at state s. The policy maximizing (1) is the optimal policy. Note that some MDPs optimize average expected reward (White, 1993) or total expected reward (Puterman, 1994) whereas we focus on MDPs that optimize total expected discounted reward.

The PIA proceeds as follows:

Step 1. Set $n = 0$, select an arbitrary decision rule $d_0 \in A$.
Step 2. (Policy evaluation) Obtain V^n by solving

$$(I - \lambda P_{d_n}) V^n = r_{d_n}, \tag{2}$$