Operations Research Letters 41 (2013) 249–251

Contents lists available at SciVerse ScienceDirect

Operations Research Letters

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

Strong polynomiality of policy iterations for average-cost MDPs modeling replacement and maintenance problems

Eugene A. Feinberg *, Jefferson Huang

Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600, USA

ARTICLE INFO

Article history:
Received 18 January 2013
Accepted 9 February 2013
Available online 18 February 2013

Keywords:
Markov Decision Process
Average cost
Policy iteration algorithm
Linear program
Simplex method
Strongly polynomial

ABSTRACT

This note considers an average-cost Markov Decision Process (MDP) with finite state and action sets and satisfying the additional condition that there is a state to which the system jumps from any state and under any action with a positive probability. The main result is that the policy iteration algorithm is strongly polynomial for such MDPs, which are often used to model replacement and maintenance problems.

1. Introduction

Linear programming is one of the major and most efficient methods for solving discounted and average-cost Markov Decision Processes (MDPs) with finite state and action sets [5,8]. Therefore, these problems can be solved in (weakly) polynomial time [7,6]. For each of these two criteria, another solution method, policy iteration algorithms, can be viewed as a version of the simplex method applied to one of the two major linear programs used to solve MDPs; see, for example, [4,5], or [8].

Typically minimization of average costs per unit time is a more difficult problem than minimization of total discounted costs. Ye [11] provided a strongly polynomial algorithm for finding discount-optimal policies. More recently, Ye [12] proved that policy iterations find discount-optimal policies in strongly polynomial time with the bound depending on the discount factor. For average costs per unit time Zadorojniy et al. [13] constructed a strongly polynomial algorithm for finding optimal average-cost policies for certain birth and death processes.

In this note, we consider an average-cost MDP with a special state to which the process moves with a fixed positive probability from each state under each action. Such problems arise in replacement and preventive maintenance problems, where the special state corresponds to machine breaks, replacements, or repairs. For many particular replacement and maintenance problems, there are explicit formulas for optimal policies; see e.g., [1]. For a fixed failure probability, the results of this paper imply that all such problems can be solved in strongly polynomial time.

Ross [10,9] studied such problems in the context of infinite-state MDPs and showed that they can be reduced to discounted problems. Gubenko & Štatland [2] and Dynkin & Yushkevich [1] considered for infinite state spaces a generalization of such problems, which were called problems with minorants in Dynkin & Yushkevich [1].

In this note, we show that Howard's policy iteration algorithm and the simple policy iteration algorithm are strongly polynomial for such average-cost MDPs. In fact this is true because, under the reduction introduced by Ross [10], a policy iteration algorithm for average-cost problems becomes a policy iteration algorithm for discounted-cost problems.

2. Definitions

Consider a discrete-time MDP with a finite state space I and finite nonempty action sets $A(i)$, $i \in I$. If the process is in a state i and an action $a \in A(i)$ is chosen then a one-step cost $c(i,a)$ is incurred and the process transitions to a state $j \in I$ with probability $p(j|i,a)$. Let $m = |I|$ denote the total number of states, and let $n = \sum_{i\in I} |A(i)|$ denote the total number of actions. Of course, $m \le n$.

A *stationary policy* is a mapping $\phi : I \to \bigcup_{i\in I} A(i)$; under the stationary policy ϕ the controller always selects action $\phi(i)$ when the system is in state $i \in I$. Let F denote the set of all such policies.

Given a policy $\phi \in F$ and an initial state $i \in I$, let $\mathbb{P}_i^{\phi}$ denote the probability distribution on the set of possible histories $i_0a_0i_1$

* Corresponding author. Tel.: +1 6316327189.
E-mail address: eugene.feinberg@stonybrook.edu (E.A. Feinberg).

0167-6377/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.orl.2013.02.002

$a_1 \ldots$ of the process under the policy ϕ with $i_0 = i$. Let $\mathbb{E}_i^\phi$ denote the expectation operator associated with $\mathbb{P}_i^\phi$.

In this note, we study MDPs satisfying the following condition.

Assumption 1. There is a state $i^* \in I$ such that $p(i^*|i, a) > 0$ for all $a \in A(i), i \in I$.

If Assumption 1 holds, then $p(i^*|i, a) \geq \gamma$ for all $i \in I$ and $a \in A(i)$ for some $\gamma > 0$. Note that it takes $O(mn)$ operations to check whether the MDP satisfies Assumption 1 and to find the largest possible γ.

Any stationary policy defines a Markov chain with the state space I. Assumption 1 implies that under any stationary policy this Markov chain has one recurrent class. Such MDPs are called *unichain.*

Average-cost criterion. For this criterion, the goal is to minimize the long-run expected average costs per unit time. In particular, for a given initial state $i \in I$ and stationary policy ϕ, let

$$v^\phi(i) = \limsup_{n\to\infty} \mathbb{E}_i^\phi \left[\frac{1}{n} \sum_{t=0}^{n-1} c(i_t, a_t) \right].$$

If the MDP is unichain, then $v^\phi(i) = v^\phi(j)$ for all $i, j \in I$ and for all stationary policies ϕ. Therefore, we shall write v^ϕ instead of $v^\phi(i)$. Let $v = \inf_{\phi\in F} v^\phi$. We would like to find a policy ϕ^* that is *average-cost optimal*, i.e. such that $v^{\phi^*} = v$.

It is well known for unichain MDPs that, if we fix $b(\ell) = 0$ for one arbitrarily chosen state $\ell \in I$, the equation

$$g + b(i) = c(i, \phi(i)) + \sum_{j\in I} p(j|i, \phi(i))b(j), \tag{1}$$

$i \in I$, has a unique solution (g^ϕ, b^ϕ), and $g^\phi = v^\phi$ [3]. In addition, there exists a stationary policy ϕ such that

$$v^\phi + b^\phi(i) = \min_{a\in A(i)} \left\{ c(i, a) + \sum_{j\in I} p(j|i, a)b^\phi(j) \right\} \tag{2}$$

for all $i \in I$. Such a policy is average-cost optimal.

Polynomial algorithms. An algorithm for finding an optimal policy for an MDP is *polynomial* if the number of arithmetic operations needed to return an optimal policy is bounded by a polynomial in the number of actions n and the bit-size of the (rational) input data. If the number of arithmetic operations needed is bounded by a polynomial only in n, then the algorithm is *strongly polynomial.* A polynomial algorithm that is not strongly polynomial is called *weakly polynomial.*

Both Khachiyan's ellipsoid method [7] and Karmarkar's interior-point algorithm [6] can be used to solve linear programs to find stationary optimal policies for discounted- and average-cost MDPs. However, as the number of arithmetic operations required by these algorithms is bounded by a polynomial dependent on the bit-size of the input data, these algorithms are weakly polynomial.

3. Policy iteration for average-cost MDPs

Howard's policy iteration algorithm for unichain average-cost MDPs [3] proceeds as follows.

1. Start with any stationary policy ϕ.
2. Fix any state $\ell \in I$, set $b^\phi(\ell) = 0$, and then determine $g^\phi = v^\phi$ and $b^\phi(i)$ by solving Eq. (1).
3. For each $i \in I$, set $\psi(i) = \phi(i)$ if Eq. (2) holds for state i; otherwise, let $\psi(i) \in A(i)$ be such that
$$c(i, \psi(i)) + \sum_{j\in I} p(j|i, \psi(i))b^\phi(j) = \min_{a\in A(i)} \left\{ c(i, a) + \sum_{j\in I} p(j|i, a)b^\phi(j) \right\}. \tag{3}$$
4. If $\psi(i) = \phi(i)$ for all $i \in I$, the policy ϕ is optimal; otherwise, replace ϕ with ψ and return to step 2.

We remark that the policy iteration algorithm is the simplex method with block pivoting applied to the linear program [4, Chapter 4]

$$\begin{aligned}
\text{minimize} \quad & \sum_{i\in I} \sum_{a\in A(i)} c(i, a)x_{i,a} \\
\text{subject to} \quad & \sum_{a\in A(j)} x_{j,a} - \sum_{i\in I} \sum_{a\in A(i)} p(j|i, a)x_{i,a} = 0, \quad j \in I, \\
& \sum_{i\in I} \sum_{a\in A(i)} x_{i,a} = 1, \\
& x_{i,a} \geq 0 \quad \text{for all } a \in A(i),\ i \in I.
\end{aligned}$$

A variant of the policy iteration algorithm described above, called *simple policy iteration*, only updates the policy at a single state. In particular, the policy is updated at a state i^* for which the difference Δ_i^ϕ between the right-hand side of (3) and $(v^\phi + b^\phi(i))$ is minimal among all $i \in I$. This corresponds to the simplex method with Dantzig's pivoting rule applied to the above LP; see Kallenberg [4, Chapter 4].

Theorem 1. *If Assumption 1 holds, then Howard's and simple policy iteration algorithms for average-cost MDPs are strongly polynomial. In particular, they both terminate in $O(m(n - m)\gamma^{-1} \log(m^2\gamma^{-1}))$ iterations, where the number of arithmetic operations in each iteration is bounded by a polynomial in the total number of actions n.*

4. Discounted-cost criterion

To prove Theorem 1, consider the discounted-cost criterion. Given an initial state i, a policy ϕ, and a *discount factor* $\beta \in [0, 1)$, let

$$v_\beta^\phi(i) = \mathbb{E}_i^\phi \left[\sum_{t=0}^{\infty} \beta^t c(i_t, a_t) \right],$$

and let $v_\beta(i) = \inf_{\phi\in F} v_\beta^\phi(i)$, $i \in I$. A policy ϕ^* is called β*-optimal* if $v_\beta^{\phi^*}(i) = v_\beta(i)$ for all $i \in I$.

It is well known that the equation

$$u(i) = c(i, \phi(i)) + \beta \sum_{j\in I} p(j|i, \phi(i))u(j), \tag{4}$$

$i \in I$, has the unique finite solution $u(i) = v_\beta^\phi(i)$. In addition, there exists a stationary policy ϕ such that

$$v_\beta^\phi(i) = \min_{a\in A(i)} \left\{ c(i, a) + \beta \sum_{j\in I} p(j|i, a)v_\beta^\phi(j) \right\}, \tag{5}$$

and it is well-known that a stationary policy ϕ is β-optimal if and only if (5) holds.

We remark that it is also possible to consider more general policies than stationary ones. However, for discounted and average-cost MDPs with finite state and action sets, stationary policies are optimal within the set of all policies [8].

For discounted MDPs, Howard's policy iteration algorithm [3] proceeds as follows.

1. Start with any stationary policy ϕ.
2. Determine $u(i) = v_\beta^\phi(i)$, $i \in I$, by solving Eq. (4).
3. For each $i \in I$, set $\psi(i) = \phi(i)$ if Eq. (5) holds for state i; otherwise, let $\psi(i) \in A(i)$ be such that
$$c(i, \psi(i)) + \beta \sum_{j\in I} p(j|i, \psi(i)) v_\beta^\phi(j) = \min_{a\in A(i)} \left\{ c(i, a) + \beta \sum_{j\in I} p(j|i, a) v_\beta^\phi(j) \right\}. \quad (6)$$
4. If $\psi(i) = \phi(i)$ for all $i \in I$, the policy ϕ is optimal; otherwise, replace ϕ with ψ and return to step 2.

Simple policy iteration for the discounted-cost criterion is defined analogously to the average-cost criterion version, with $\Delta_i^\phi(\beta)$ defined as the difference between the right-hand side of (6) and $v_\beta^\phi(i)$. Ye [12] showed that both simple policy iteration and Howard's policy iteration algorithm are strongly polynomial with the upper bound on the number of iterations depending on the discount factor β. In particular, both algorithms terminate after $O(m(n-m)(1-\beta)^{-1}\log(m^2(1-\beta)^{-1}))$ iterations, where each iteration requires a number of arithmetic operations bounded by a polynomial in n.

5. Ross's reduction and proof of Theorem 1

Following Ross [10], set $\beta = 1 - \gamma$ and
$$\tilde{p}(j|i, a) = \begin{cases} p(j|i, a)/(1-\gamma), & \text{if } j \neq i^*, \\ (p(i^*|i, a) - \gamma)/(1-\gamma), & \text{if } j = i^*. \end{cases}$$

Consider a discounted-cost MDP with discount factor β, state space I, action sets $A(i)$, costs $c(i, a)$, and transition probabilities $\tilde{p}(j|i, a)$, where $i, j \in I$ and $a \in A(i)$.

Let $i \in I$ and $\phi \in F$. Recall that for the average-cost MDP,
$$\Delta_i^\phi = \min_{a\in A(i)} \left\{ c(i, a) + \sum_{j\in I} p(j|i, a) b^\phi(j) \right\} - (v^\phi + b^\phi(i)).$$

For the discounted MDP obtained from Ross's reduction, let
$$\tilde{\Delta}_i^\phi(\beta) = \min_{a\in A(i)} \left\{ c(i, a) + \beta \sum_{j\in I} \tilde{p}(j|i, a) v_\beta^\phi(j) \right\} - v_\beta^\phi(i).$$

Lemma 2. *Let Assumption 1 hold. Then for each stationary policy ϕ and for each state $i \in I$, the following statements hold:*

(i) *the set of actions minimizing the right-hand side of (3) coincides with the set of actions minimizing the right-hand side of (6) for the discounted MDP obtained via Ross's reduction of the original MDP;*

(ii) *$\Delta_i^\phi = \tilde{\Delta}_i^\phi(\beta)$ for all $\phi \in F$, $i \in I$, where $\beta = 1 - \gamma$.*

Proof. The solution to (4) obtained in step 2 of Howard's policy iteration algorithm for the discounted MDP is $v_\beta^\phi(i)$, $i \in I$. Since $u = v_\beta^\phi$, (4) is equivalent to
$$\gamma v_\beta^\phi(i^*) + v_\beta^\phi(i) = c(i, \phi(i)) + \sum_{j\in I} p(j|i, \phi(i)) v_\beta^\phi(j),$$
$i \in I$. Thus $v^\phi = \gamma v_\beta^\phi(i^*)$ and $b^\phi(i) = v_\beta^\phi(i) - v_\beta^\phi(\ell)$, where ℓ is any fixed state from I, satisfies (2) with $b^\phi(\ell) = 0$; by uniqueness, this is the solution obtained in step 2 of Howard's policy iteration algorithm for the original average-cost MDP.

Observe that $\sum_{j\in I} p(j|i, a) b^\phi(j)$ and $\beta \sum_{j\in I} \tilde{p}(j|i, a) v_\beta^\phi(j)$ differ by $\gamma v_\beta^\phi(i^*) - v_\beta^\phi(l)$, which does not depend on i. This implies (i). Statement (ii) holds because
$$\begin{aligned} \Delta_i^\phi &= \min_{a\in A(i)} \left\{ c(i, a) + \beta \sum_{j\in I} \tilde{p}(j|i, a) v_\beta^\phi(j) \right\} - v_\beta^\phi(i) \\ &= \min_{a\in A(i)} \left\{ c(i, a) + \sum_{j\in I} p(j|i, a) v_\beta^\phi(j) \right\} - \gamma v_\beta^\phi(i^*) - v_\beta^\phi(i) \\ &= \min_{a\in A(i)} \left\{ c(i, a) + \sum_{j\in I} p(j|i, a) b^\phi(j) \right\} - v^\phi - b^\phi(i) \\ &= \tilde{\Delta}_i^\phi(\beta). \quad \square \end{aligned}$$

Proof of Theorem 1. Lemma 2(i) implies that, given an average-cost MDP satisfying Assumption 1, there is a one-to-one correspondence between sequences of policies generated by Howard's policy iteration algorithm and sequences generated by Howard's policy iteration algorithm for the discounted MDP obtained via Ross's reduction. In view of Lemma 2(ii), this is also true for the simple policy iteration algorithms. Hence the iteration bound established by Ye [12, Theorem 4.2 & Corollary 4.1] also applies to Howard's and the simple policy iteration algorithms applied to an average-cost MDP satisfying Assumption 1.

Acknowledgment

This research was partially supported by NSF grant CMMI-0928490.

References

[1] E.B. Dynkin, A.A. Yushkevich, Controlled Markov Processes, Springer-Verlag, Berlin, 1979.

[2] L.G. Gubenko, E.S. Štatland, On controlled, discrete-time Markov decision processes, Theory of Probability and Mathematical Statistics 7 (1975) 47–61.

[3] R.A. Howard, Dynamic Programming and Markov Processes, MIT Press, Cambridge, MA, 1960.

[4] L.C.M. Kallenberg, Linear programming and finite Markovian control problems, in: Mathematical Centre Tracts, Mathematisch Centrum, Amsterdam, 1983.

[5] L.C.M. Kallenberg, Finite state and action MDPs, in: E.A. Feinberg, A. Schwartz (Eds.), Handbook of Markov Decision Processes, Kluwer, Boston, 2002, pp. 21–87.

[6] N. Karmarkar, A new polynomial-time algorithm for linear programming, Combinatorica 4 (1984) 373–395.

[7] L.G. Khachiyan, A polynomial algorithm in linear programming, Doklady Akademiia Nauk SSSR 244 (1979) 1086–1093.

[8] M. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons, Inc., New York, 1994.

[9] S.M. Ross, Arbitrary state Markovian decision processes, The Annals of Mathematical Statistics 39 (1968) 2118–2122.

[10] S.M. Ross, Non-discounted denumerable Markovian decision models, The Annals of Mathematical Statistics 39 (1968) 412–423.

[11] Y. Ye, A new complexity result on solving the Markov decision problem, Mathematics of Operations Research 30 (2005) 733–749.

[12] Y. Ye, The simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate, Mathematics of Operations Research 36 (2011) 593–603.

[13] A. Zadorojniy, G. Even, A. Shwartz, A strongly polynomial algorithm for controlled queues, Mathematics of Operations Research 34 (2009) 992–1007.