Operations Research Letters 42 (2014) 130–131

Contents lists available at ScienceDirect

Operations Research Letters

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

The value iteration algorithm is not strongly polynomial for discounted dynamic programming

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 20 December 2013
Accepted 31 December 2013
Available online 17 January 2014

Keywords:
Markov decision process
Value iteration
Strongly polynomial
Policy
Algorithm

ABSTRACT

This note provides a simple example demonstrating that, if exact computations are allowed, the number of iterations required for the value iteration algorithm to find an optimal policy for discounted dynamic programming problems may grow arbitrarily quickly with the size of the problem. In particular, the number of iterations can be exponential in the number of actions. Thus, unlike policy iterations, the value iteration algorithm is not strongly polynomial for discounted dynamic programming.

1. Introduction

Value iterations, policy iterations, and linear programming are three major methods for computing optimal policies for Markov Decision Processes (MDPs) with expected total discounted rewards [1], [3, Chapter 6], also known under the name of discounted dynamic programming. As is well-known, policy iterations can be viewed as implementations of the simplex method applied to one of the two major linear programs used to solve MDPs; see e.g., [1], [3, Section 6.9]. Ye [5] proved that policy iterations are strongly polynomial when the discount factor is fixed. This note shows that value iterations may not be strongly polynomial.

For value iteration, the best known upper bound on the required number of iterations was obtained by Tseng [4] (see also Littman [2] and Ye [5]) and is a polynomial in the number of states n, the number of actions m, the number of bits B needed to write down the problem data, and $(1-\beta)^{-1}$, where $\beta \in (0,1)$ is the discount factor. Since the number of arithmetic operations needed per iteration is at most a constant times n^2m, this means that the value iteration algorithm is weakly polynomial if the discount factor is fixed.

This note provides a simple example that demonstrates that, if exact computations are allowed, the number of operations performed by the value iteration algorithm can grow arbitrarily quickly as a function of the total number of available actions. In particular, the running time can be exponential with respect to the total number of actions m. Thus, unlike policy iterations, value iterations are not strongly polynomial.

2. Example

Consider an arbitrary increasing sequence $\{M_i\}_{i=1}^{\infty}$ of natural numbers. Let the state space be $\mathbb{X} = \{1,2,3\}$, and for a natural number k let the action space be $\mathbb{A} = \{0,1,\dots,k\}$. Let $\mathbb{A}(1) = \mathbb{A}$, $\mathbb{A}(2) = \{0\}$, and $\mathbb{A}(3) = \{0\}$ be the sets of actions available at states 1, 2, and 3, respectively. The transition probabilities are $p(2|1,i) = p(3|1,0) = p(2|2,0) = p(3|3,0) = 1$ for $i = 1,\dots,k$. Finally, the one-step rewards are $r(1,0) = r(2,0) = 0$, $r(3,0) = 1$, and

$$r(1,i) = \frac{\beta}{1-\beta}(1-\exp(-M_i)), \quad i = 1,\dots,k.$$

Fig. 1 illustrates such an MDP for $k = 2$.

2.1. Discounted-reward criterion

Here we are interested in maximizing expected infinite-horizon discounted rewards. In particular, a *policy* is a mapping $\phi : \mathbb{X} \to \mathbb{A}$ such that $\phi(x) \in \mathbb{A}(x)$ for each $x \in \mathbb{X}$. It is possible to consider more general policies, but for infinite-horizon discounted MDPs with finite state and action sets it is sufficient to consider only policies of this form; see e.g., [3, p. 154]. Let F denote the set of all policies. Also, given an initial state $x \in \mathbb{X}$, let $\mathbb{P}_x^{\phi}$ denote the probability distribution on the set of possible histories $x_0a_0x_1a_1\cdots$ of the process under the policy ϕ with $x_0 = x$, and let $\mathbb{E}_x^{\phi}$ be the

[*] Corresponding author.
E-mail address: eugene.feinberg@stonybrook.edu (E.A. Feinberg).

0167-6377/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.orl.2013.12.011

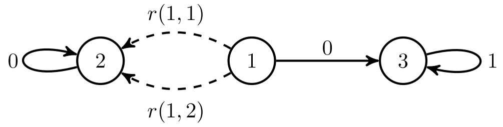

Fig. 1. Diagram of the MDP for $k = 2$. The solid arcs correspond to transitions associated with action 0, and dashed arcs correspond to the remaining actions. The number next to each arc is the reward associated with the corresponding action.

expectation operator associated with $\mathbb{P}_x^\phi$. Then the expected total discounted reward earned when the policy ϕ is used starting in state $x \in \mathbb{X}$ is

$$v_\beta(x, \phi) = \mathbb{E}_x^\phi \sum_{t=0}^{\infty} \beta^t r(x_t, a_t).$$

The goal is to find an *optimal policy*, that is, a policy ϕ^* such that $v_\beta(x, \phi^*) = \sup_{\phi \in F} v_\beta(x, \phi)$ for all $x \in \mathbb{X}$. It is well-known that if $\mathbb{X}$ and $\mathbb{A}$ are finite, then an optimal policy exists; see e.g., [3, p. 154].

For the above described MDP each policy is defined by an action selected at state 1. Note that if action $i \in \{1, \dots, k\}$ is selected, then the total discounted reward starting from state 1 is $r(1, i)$; if action 0 is selected, the corresponding total discounted reward is $\beta/(1 - \beta)$. Since

$$r(1, i) = \frac{\beta}{1 - \beta}(1 - \exp(-M_i)) < \frac{\beta}{1 - \beta}$$

for each $i = 1, \dots, k$, action 0 is the unique optimal action at state 1.

2.2. Running time of value iterations

We are interested in obtaining the optimal policy using *value iteration*. In particular, set $V_0 \equiv 0$, and for each $x \in \mathbb{X}$ and $j = 0, 1, \dots$ let

$$V_{j+1}(x) = \max_{a \in \mathbb{A}(x)} \left\{ r(x, a) + \beta \sum_{y \in \mathbb{X}} p(y|x, a) V_j(y) \right\},$$

and

$$\phi^{j+1}(x) \in \arg\max_{a \in \mathbb{A}(x)} \left\{ r(x, a) + \beta \sum_{y \in \mathbb{X}} p(y|x, a) V_j(y) \right\}.$$

Since the numbers M_i increase in i, for $j = 0, 1, \dots$

$$V_{j+1}(1) = \max\left\{ \frac{\beta(1 - \beta^j)}{1 - \beta}, \frac{\beta(1 - \exp(-M_k))}{1 - \beta} \right\},$$

$$V_{j+1}(2) = 0,$$

$$V_{j+1}(3) = \frac{1 - \beta^{j+1}}{1 - \beta},$$

which implies that

$$\phi^{j+1}(1) = \begin{cases} k, & \text{if } j < M_k/(-\ln\beta), \\ 0, & \text{if } j > M_k/(-\ln\beta). \end{cases}$$

Hence at least $M_k/(-\ln\beta)$ iterations are needed to select the optimal action 0 at state 1. Let $M_k = 2^k$. Since the total number of actions $m = k + 3$, at least $C2^m/(-\ln\beta)$ iterations are required to obtain the optimal policy, where $C = 2^{-3}$.

Acknowledgment

This research was partially supported by NSF grant CMMI-1335296.

References

[1] 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.
[2] M.L. Littman, T.L. Dean, L.P. Kaelbling, On the complexity of solving Markov decision problems, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, 1995, pp. 394–402.
[3] M. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons, Inc., New York, 1994.
[4] P. Tseng, Solving h-horizon, stationary Markov Decision Problems in time proportional to $\log(h)$, Oper. Res. Lett. 9 (1990) 287–297.
[5] Y. Ye, The simplex and policy-iteration methods are strongly polynomial for the Markov Decision Problem with a fixed discount rate, Math. Oper. Res. 36 (2011) 593–603.