Operations Research Letters 42 (2014) 429–431

Contents lists available at ScienceDirect

Operations Research Letters

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

Modified policy iteration algorithms are not strongly polynomial for discounted dynamic programming

Eugene A. Feinberg [a,*], Jefferson Huang [a], Bruno Scherrer [b,c]

[a] Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600, USA
[b] Inria, Villers-lès-Nancy, F-54600, France
[c] Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France

ARTICLE INFO

Article history:
Received 8 July 2014
Received in revised form
21 July 2014
Accepted 23 July 2014
Available online 1 August 2014

Keywords:
Markov decision process
Modified policy iteration
Strongly polynomial
Policy
Algorithm

ABSTRACT

This note shows that the number of arithmetic operations required by any member of a broad class of optimistic policy iteration algorithms to solve a deterministic discounted dynamic programming problem with three states and four actions may grow arbitrarily. Therefore any such algorithm is not strongly polynomial. In particular, the modified policy iteration and λ-policy iteration algorithms are not strongly polynomial.

1. Introduction

Value iteration (VI), policy iteration (PI), and linear programming algorithms are three major methods for computing optimal policies for Markov decision processes (MDPs) with expected total discounted rewards [8], [11, Ch. 6], also known under the name of discounted dynamic programming. As is well-known, PI can be viewed as an implementation of the simplex method applied to one of the two major linear programs used to solve MDPs; see e.g. [8], [11, Section 6.9]. Using these linear programs, Ye [16] proved that both Howard's [7] PI and the simplex method with Dantzig's pivoting rule are strongly polynomial when the discount factor is fixed; in other words, taking the discount factor to be a constant, the number of arithmetic operations needed by these two algorithms to return an optimal policy is bounded above by a polynomial function of the number of state–action pairs m. Post and Ye [10] subsequently showed that, if the MDP is deterministic, then the simplex method with Dantzig's rule is strongly polynomial regardless of the discount factor. In contrast, Feinberg and Huang [5] used a deterministic MDP to demonstrate that VI is not strongly polynomial even when the discount factor is fixed. As was proved by Tseng [15], the VI algorithm is weakly polynomial, that is, the number of required arithmetic operations can be bounded above by a polynomial in m and the total bit-size of the input data.

Each iteration of PI involves the solution of a system of linear equations, which may be time consuming if the number of states is large. Several methods have been proposed to deal with this issue by combining the advantages of PI and VI. One approach is modified policy iteration (MPI), where the exact solutions are replaced with estimates obtained via finite numbers of successive approximations; see Puterman and Shin [12]. Another approach is λ-policy iteration (λPI), also called temporal difference-based policy iteration; see Bertsekas and Tsitsiklis [2, Section 2.3.1]. Both of these algorithms include VI and PI as special cases. In studying performance bounds for approximate versions of λPI, Thiéry and Scherrer [14] considered a generalization of both MPI and λPI, which they refer to as optimistic policy iteration (OPI). In this note, we use a variant of Feinberg and Huang's [5] example to show that a generalization of OPI, which we call *generalized optimistic policy iteration (G-OPI)*, is not strongly polynomial (Theorem 1). In particular, our result implies that VI, MPI, λPI, and OPI are also not strongly polynomial (Corollary 2).

We remark that the results in Ye [16] have led to further developments. For instance, Hansen Miltersen and Zwick [6] improved the iteration bound for Howard's PI given in [16] by a factor of the number of states n and showed that it also applies to the strategy

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

http://dx.doi.org/10.1016/j.orl.2014.07.006
0167-6377/© 2014 Elsevier B.V. All rights reserved.

iteration algorithm for two-player turn-based zero-sum stochastic games. Scherrer [13] improved both the estimate in [6] for Howard's PI and the estimate in [16] for the simplex method by a factor of $\ln(n)$, showing that if the discount factor is fixed then Howard's PI needs at most a linear number of iterations in m and the simplex method with Dantzig's rule needs at most a linear number of iterations in mn. The results and analysis in Ye [16] have also been applied in both more general and different contexts. Kitahara and Mizuno [9] used the analysis in [16] to obtain a sufficient condition for the simplex method to be strongly polynomial for linear programs in general. In addition, Ye [16, Section 5] notes that the analysis of discounted MDPs can be extended to transient MDPs; Denardo [3] showed that with some modifications, the analysis given in [16, Section 5] can yield a bound improved by a factor of 2 for such MDPs. Finally, the results in [16] are relevant for certain MDPs under the average-reward criterion; see Feinberg and Huang [4] and Akian and Gaubert [1].

2. Generalized optimistic policy iteration

In Section 2.1 we describe the discounted-reward criterion. In Section 2.2, we formulate the G-OPI algorithm and state our results, namely Theorem 1 and Corollary 2, which are proved in Section 3.

2.1. Discounted-reward criterion

Consider a discrete-time MDP with finite state set $\mathbb{X}$, finite nonempty sets of actions $A(x)$ available at each $x \in \mathbb{X}$, transition probabilities $p(y|x, a)$ for each $x, y \in \mathbb{X}$ and $a \in A(x)$, and one-step rewards $r(x, a)$ for each $x \in \mathbb{X}$ and $a \in A(x)$. Let $m := \sum_{x\in\mathbb{X}} |A(x)|$ denote the total number of state–action pairs.

Here we are interested in maximizing expected infinite-horizon discounted rewards. In particular, a *policy* is a mapping $\phi : \mathbb{X} \to \bigcup_{x\in\mathbb{X}} A(x)$ such that $\phi(x) \in A(x)$ for each $x \in \mathbb{X}$. One may 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. [11, 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 $x_0a_0x_1a_1 \dots$ of the process under the policy ϕ with $x_0 = x$, and let $\mathbb{E}_x^\phi$ be the expectation operator associated with $\mathbb{P}_x^\phi$. Then, letting $\beta \in (0, 1)$ denote the discount factor, the expected total discounted reward earned when the policy ϕ is used starting from 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*, i.e. 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 $\bigcup_{x\in\mathbb{X}} A(x)$ are finite, then an optimal policy exists; see e.g. [11, p. 154]. To describe the G-OPI algorithm, it will be convenient to define the operators T and T_ϕ, $\phi \in F$, on functions $v : \mathbb{X} \to \mathbb{R}$ for each $x \in \mathbb{X}$ by

$$Tv(x) := \max_{a\in A(x)} \left\{ r(x, a) + \beta \sum_{y\in\mathbb{X}} p(y|x, a)v(y) \right\}$$

and

$$T_\phi v(x) := r(x, \phi(x)) + \beta \sum_{y\in\mathbb{X}} p(y|x, \phi(x))v(y),$$

where for $n = 1, 2, \dots, T_\phi^0 v(x) := v(x)$ and $T_\phi^n v(x) := T_\phi(T_\phi^{n-1} v)(x)$.

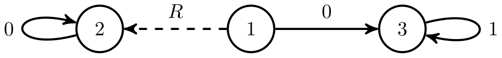

Fig. 1. The solid arcs correspond to transitions associated with action 0, and the dashed arc corresponds to action 1. The number next to each arc is the one-step reward that taking the corresponding action earns.

2.2. The algorithm

Algorithm 1 (*G-OPI*). Let $\mathbb{N}$ denote the set of positive integers, $\bar{\mathbb{N}} := \mathbb{N}\cup\{+\infty\}$, and let $\{N_j\}_{j=1}^\infty$ be an $\bar{\mathbb{N}}$-valued stochastic sequence with associated probability measure P and expectation operator E. Then given $V_0 : \mathbb{X} \to \mathbb{R}$, set $j = 1$ and choose any policy ϕ^j satisfying

$$T_{\phi^j} V_{j-1}(x) = TV_{j-1}(x) \quad \text{for each } x \in \mathbb{X}. \tag{1}$$

If $V_{j-1}(x) = TV_{j-1}(x)$ for all $x \in \mathbb{X}$, then ϕ^j is an optimal policy; otherwise, set

$$V_j(x) = E\left[T_{\phi^j}^{N_j} V_{j-1}(x)\right] \quad \text{for each } x \in \mathbb{X}, \tag{2}$$

increase j by 1, and repeat starting from (1).

In the sequel, we assume that a strongly polynomial algorithm exists for evaluating the expectation in (2) for each $j \in \mathbb{N}$; otherwise, it trivially follows that the G-OPI algorithm is not strongly polynomial. The following statement, which is proved in Section 3, is the main result of this note.

Theorem 1. *If*

$$P\{N_j < +\infty\} > 0 \quad \textit{for each } j \in \mathbb{N},$$

then the number of iterations needed by the generalized optimistic policy iteration algorithm to return the optimal policy may grow arbitrarily quickly as the number of state–action pairs m increases, which implies that the algorithm is not strongly polynomial.

The generalized optimistic policy iteration algorithm includes VI, MPI, λPI, OPI, and Howard's PI as special cases. In fact, we show in Section 3 that Theorem 1 implies

Corollary 2. *The value iteration, modified policy iteration, λ-policy iteration, and optimistic policy iteration algorithms are not strongly polynomial.*

Note that Theorem 1 does not apply to Howard's PI, under which $P\{N_j = +\infty\} = 1$ for each $j \in \mathbb{N}$, and which is strongly polynomial according to Ye [16].

3. Proofs

To prove Theorem 1, we shall consider the following example.

Example 1. Let the state set be $\mathbb{X} = \{1, 2, 3\}$, and given a positive integer k, let $A(1) = \{0, 1\}$, $A(2) = \{0\}$, and $A(3) = \{0\}$ be the sets of actions available at states 1, 2, and 3, respectively; hence the number of state–action pairs $m = 4$. The transition probabilities are $p(2|1, 1) = p(3|1, 0) = p(2|2, 0) = p(3|3, 0) = 1$. Finally, the one-step rewards are $r(1, 0) = r(2, 0) = 0$, $r(3, 0) = 1$, and $r(1, 1) = R < \beta/(1 - \beta)$. Fig. 1 illustrates this MDP.

For this MDP, each policy is characterized by the action selected at state 1. If action 1 is selected, then the total discounted reward starting from state 1 is $R < \beta/(1 - \beta)$; if action 0 is selected, the corresponding total discounted reward is $\beta/(1 - \beta)$. Hence action 0 is the optimal action at state 1. □

Proof of Theorem 1. Apply the G-OPI algorithm to the MDP in Example 1 with $V_0(1) = V_0(2) = V_0(3) = 0$. From (2),

$$\begin{aligned} V_1(3) &= E\left[1 + \beta + \cdots + \beta^{N_1 - 1} + \beta^{N_1} \cdot 0\right] \\ &= E\left[1 - \beta^{N_1}\right]/(1-\beta) \\ &= \frac{1}{1-\beta}(1 - E[\beta^{N_1}]) \end{aligned}$$

and

$$\begin{aligned} V_2(3) &= E[1 + \beta + \cdots + \beta^{N_2 - 1} + \beta^{N_2} V_1(3)] \\ &= \frac{1}{1-\beta}(1 - E[\beta^{N_2}] + E[\beta^{N_2}(1 - E[\beta^{N_1}])]) \\ &= \frac{1}{1-\beta}(1 - E[\beta^{N_2}]E[\beta^{N_1}]). \end{aligned}$$

Hence by induction,

$$\begin{aligned} V_j(3) &= E\left[1 + \beta + \cdots + \beta^{N_j - 1} + \beta^{N_j} V_{j-1}(3)\right] \\ &= \frac{1}{1-\beta}\left(1 - \prod_{\ell=1}^{j} E[\beta^{N_\ell}]\right) \quad \text{for each } j \in \mathbb{N}. \end{aligned}$$

This means the optimal action 0 at state 1 will be selected on iteration j^* only if

$$\beta V_{j^*-1}(3) = \frac{\beta}{1-\beta}\left(1 - \prod_{\ell=1}^{j^*-1} E[\beta^{N_\ell}]\right) \geq R.$$

Suppose $P\{N_j < +\infty\} = \sum_{n=1}^{\infty} P\{N_j = n\} > 0$ for each $j \in \mathbb{N}$. Then $P\{N_j = n_0\} > 0$ for some $n_0 \in \mathbb{N}$; since $\beta > 0$ and $\beta^{+\infty} = 0$ this implies

$$\begin{aligned} E[\beta^{N_j}] &= \sum_{n=1}^{\infty} \beta^n P\{N_j = n\} \\ &\geq \beta^{n_0} P\{N_j = n_0\} > 0 \quad \text{for each } j \in \mathbb{N}. \end{aligned}$$

It follows that

$$\begin{aligned} \beta V_j(3) &= \frac{\beta}{1-\beta} \cdot \left(1 - \prod_{\ell=1}^{j} E[\beta^{N_\ell}]\right) \\ &< \frac{\beta}{1-\beta} \quad \text{for each } j \in \mathbb{N}. \end{aligned}$$

Hence, for any $k < \infty$, R may be chosen such that for all $j \leq k$,

$$\frac{\beta}{1-\beta} > R > \beta V_j(3).$$

In other words, the number of iterations before the algorithm switches to the optimal action 0 can be arbitrarily large. □

Proof of Corollary 2. The VI, MPI, λPI, and OPI algorithms differ from G-OPI only in how V_j is computed in (2). For VI, (2) is replaced with

$$V_j(x) = T_{\phi^j} V_{j-1}(x) \quad \text{for each } x \in \mathbb{X},$$

so $P\{N_j < +\infty\} = P\{N_j = 1\} = 1$ for all $j \in \mathbb{N}$. For MPI, each N_j is simply a constant $n_j \in \mathbb{N}$, so (2) can be written as

$$V_j(x) = T_{\phi^j}^{n_j} V_{j-1}(x) \quad \text{for each } x \in \mathbb{X}$$

and $P\{N_j < +\infty\} = P\{N_j = n_j\} = 1$ for each $j \in \mathbb{N}$. For λPI, each N_j is an independent geometric random variable, i.e. for $j \in \mathbb{N}$

$$P\{N_j = n\} = (1 - \lambda_j)\lambda_j^{n-1}, \quad \lambda_j \in [0, 1), \; n \in \mathbb{N},$$

implying that $P\{N_j < +\infty\} = 1$ for each $j \in \mathbb{N}$. Finally, under the OPI algorithm the distribution of each N_j is defined by a sequence $\{\lambda_n^j\}_{n=1}^{\infty}$ of nonnegative numbers satisfying $\sum_{n=1}^{\infty} \lambda_n^j = 1$, where

$$P\{N_j = n\} = \lambda_n^j, \quad n \in \mathbb{N};$$

hence $P\{N_j < +\infty\} = \sum_{n=1}^{\infty} \lambda_n^j = 1$ for all $j \in \mathbb{N}$. Hence each of these algorithms is an instance of G-OPI where $P\{N_j < +\infty\} > 0$ for each $j \in \mathbb{N}$. □

Acknowledgment

The research of the first two authors was partially supported by NSF Grant CMMI-1335296.

References

[1] M. Akian, S. Gaubert, Policy iteration for perfect information stochastic mean payoff games with bounded first return times is strongly polynomial. Preprint, 2013. http://arxiv.org/abs/1310.4953v1.

[2] D.P. Bertsekas, J.N. Tsitsiklis, Neuro-Dynamic Programming, Athena Scientific, Belmont, MA, 1996.

[3] E.V. Denardo, Nearly strongly polynomial algorithms for transient Markov decision problems. Unpublished Manuscript, 2014.

[4] E.A. Feinberg, J. Huang, Strong polynomiality of policy iterations for average-cost MDPs modeling replacement and maintenance problems, Oper. Res. Lett. 41 (3) (2013) 249–251.

[5] E.A. Feinberg, J. Huang, The value iteration algorithm is not strongly polynomial for discounted dynamic programming, Oper. Res. Lett. 42 (2) (2014) 130–131.

[6] T.D. Hansen, P.B. Miltersen, U. Zwick, Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor, J. ACM 60 (1) (2013) Article 1, 16 pages.

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

[8] 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.

[9] T. Kitahara, S. Mizuno, A bound for the number of different basic solutions generated by the simplex method, Math. Program. A 137 (2013) 579–586.

[10] I. Post, Y. Ye, The simplex method is strongly polynomial for deterministic Markov decision processes, to appear in Math. Oper. Res..

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

[12] M.L. Puterman, M.C. Shin, Modified policy iteration algorithms for discounted Markov decision problems, Manag. Sci. 24 (11) (1978) 1127–1137.

[13] B. Scherrer, Improved and generalized upper bounds on the complexity of policy iteration, in: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26, NIPS Foundation, Inc., 2013, pp. 386–394.

[14] C. Thiéry, B. Scherrer, Least-squares policy iteration: bias–variance trade-off in control problems, in: Johannes Fürnkranz, Thorsten Joachims (Eds.), Proceedings of the 27th International Conference on Machine Learning, (ICML-10), Omnipress, Haifa, Israel, June, 2010, pp. 1071–1078.

[15] P. Tseng, Solving h-horizon, stationary Markov decision problems in time proportional to log(h), Oper. Res. Lett. 9 (5) (1990) 287–297.

[16] 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 (4) (2011) 593–603.