Strongly Polynomial Algorithms for Transient and Average-Cost MDPs

Eugene A. Feinberg
Dept. Applied Mathematics and Statistics
Stony Brook University
Stony Brook, NY 11794
eugene.feinberg@stonybrook.edu

Jefferson Huang
School of ORIE
Cornell University
Ithaca, NY 14853
jefferson.huang@cornell.edu

ABSTRACT

This paper considers transient total-cost MDPs with transition rates whose values may be greater than one, and average-cost MDPs satisfying the condition that the expected time to hit a certain state from any initial state and under any stationary policy is bounded above by a constant. Linear programming formulations for such MDPs are provided that are solvable in strongly polynomial time.

Keywords

Markov decision process, algorithm, linear program, transient, total cost, average cost

1. INTRODUCTION

Markov decision processes (MDPs) provide an important framework for the optimization of controlled stochastic systems. For examples of modern applications of MDPs in healthcare, transportation, production systems, communications, and finance, see [2].

It is well-known that there is a close relation between MDPs and linear programming; see e.g., [11]. This relation was used in [19] to develop a combinatorial interior-point algorithm for discounted MDPs and to show, for the first time, that such MDPs can be solved in *strongly polynomial* time when the discount factor is fixed. This means that the required number of iterations can be bounded above by a polynomial in the number of state-action pairs only. The linear programming formulation of discounted MDPs was again used in [20] to prove that two classic algorithms, the policy iteration method proposed in [10] and the simplex method with Dantzig's pivoting rule, are also strongly polynomial when the discount factor is fixed. In fact, the complexity estimates for these two algorithms provided in [20] are superior to the estimate for the interior-point algorithm in [19]. Improvements on the complexity estimates in [20] were subsequently provided in [8, 1, 15, 3]. In addition, the estimates for Howard's policy iteration method were generalized to two-player zero-sum stochastic games in [8, 1], and the analysis in [20] was applied to general linear programs (LPs) in [12]. We remark that, in constrast to policy iteration, any member of a broad class of modified policy iteration algorithms, which includes the classic value iteration algorithm, is not strongly polynomial for MDPs with a fixed discount factor [6], and policy iteration may require exponential time when the discount factor is not fixed [9]. On the other hand, certain discounted MDPs with special structure are solvable in strongly polynomial time regardless of the discount factor [21, 13].

The results for discounted MDPs in [20] have also led to complexity estimates for MDPs under other optimality criteria. In [20], the analysis of discounted MDPs presented there is used to show that for transient total-cost MDPs where the spectral radius of every transition matrix is bounded above by a constant strictly less than one, both the simplex and policy iteration methods are strongly polynomial. In [3], the analysis in [20] is improved, and complexity estimates are provided in terms of the lifetime of the process under any stationary policy. For average costs, in [4] the results in [20] are used to show that the simplex and policy iteration methods are strongly polynomial for such MDPs with a state to which the system transitions under any action with probability at least $\alpha > 0$. For two-player zero-sum mean-payoff stochastic games, it is shown in [1] that a generalization of Howard's policy iteration method to this context is strongly polynomial when, for any initial state and under any pair of stationary strategies, the expected hitting time to a certain state is bounded above by a constant. We remark that MDPs satisfying this hitting time assumption are unichain, and that it is not known whether a strongly polynomial algorithm exists for unichain average-cost MDPs in general.

In this paper, we provide alternative linear programming formulations for transient total-cost MDPs, and average-cost MDPs satisfying a hitting time assumption, that are solvable in strongly polynomial time. In Section 2, the model and assumptions are presented. Section 3 deals with the total-cost criterion, and Section 4 deals with average costs per unit time.

2. MODEL DESCRIPTION

Consider a discrete-time MDP with finite *state set* $\mathbb{X}$ and finite *action set* $\mathbb{A}$. For each $x \in \mathbb{X}$, the *set of available actions* $A(x)$ is a nonempty subset of $\mathbb{A}$. Let $m := \sum_{x\in\mathbb{X}} |A(x)|$ and $n := |\mathbb{X}|$. The *one-step costs* are denoted by $c(x, a)$ for $x \in \mathbb{X}$ and $a \in A(x)$. Finally, to each $x, y \in \mathbb{X}$ and $a \in A(x)$ is associated a number $q(y|x, a) \geq 0$ called the *transition rate* to y given that the current state is x and action a is performed. For the transient MDPs considered in this paper, the case where $\sum_{y\in\mathbb{X}} q(y|x, a) > 1$ for some $x \in \mathbb{X}$ and $a \in A(x)$ is allowed. Such models are relevant to the control of branching processes; see e.g., [14]. For average-cost MDPs, we will only consider the case where q is *stochastic*, i.e., $\sum_{y\in\mathbb{X}} q(y|x,a) = 1$ for all $x \in \mathbb{X}$ and $a \in A(x)$, in which case $q(y|x,a)$ is interpreted as the probability that the system transitions to state y given that the current state is x and action a is performed.

A *stationary policy* is a mapping $\phi : \mathbb{X} \to \mathbb{A}$ satisfying $\phi(x) \in A(x)$ for each $x \in \mathbb{X}$; let $\mathbb{F}$ denote the set of all such policies. It can be shown that it suffices to consider stationary policies for the optimality criteria considered in this paper. Under $\phi \in \mathbb{F}$, the decision-maker always selects the action $\phi(x)$ when the current state is x. For $\phi \in \mathbb{F}$, consider the matrix of one-step transition rates Q_ϕ with elements $q(y|x,\phi(x))$, $x, y \in \mathbb{X}$. For a matrix B with elements $B(x,y)$ for $x, y \in \mathbb{X}$, let $\|B\| := \max_{x\in\mathbb{X}} \sum_{y\in\mathbb{X}} |B(x,y)|$.

For undiscounted total costs, which are considered in Section 3, the following transience condition [17] is assumed to hold.

Assumption T. *The MDP is* transient*, that is, there is a constant $K \geq 1$ that satisfies $\|\sum_{n=0}^{\infty} Q_\phi^n\| \leq K < \infty$ for all $\phi \in \mathbb{F}$.*

Assumption T can be checked in strongly polynomial time using the procedure described in [18, proof of Theorem 1].

For $\phi \in \mathbb{F}$, let $c_\phi(x) := c(x,\phi(x))$ for $x \in \mathbb{X}$. Under Assumption T, the *total cost* incurred under $\phi \in \mathbb{F}$, when the initial state is $x \in \mathbb{X}$, is $v^\phi(x) := \sum_{n=0}^{\infty} Q_\phi^n c_\phi(x)$. A policy ϕ_* is *total-cost optimal* if $v^{\phi_*}(x) = \inf_{\phi\in\mathbb{F}} v^\phi(x)$ for all $x \in \mathbb{X}$. The following characterization of Assumption T [5, Proposition 1] will be used to define the linear programs given in Sections 3 and 4.

Proposition 1. *Assumption T holds if and only if there is a function $\mu : \mathbb{X} \to [1,\infty)$ that is bounded above by K and satisfies*

$$\mu(x) \geq 1 + \sum_{y\in\mathbb{X}} q(y|x,a)\mu(y), \quad x \in \mathbb{X},\ a \in A(x). \qquad (1)$$

For average costs, which are dealt with in Section 4, Assumption HT on hitting times formulated below is assumed to hold. To state it, for $z \in \mathbb{X}$ and $\phi \in \mathbb{F}$ consider the matrix ${}_zQ_\phi$ with elements ${}_zQ_\phi(x,y) := q(y|x,\phi(x))$ if $x \in \mathbb{X}$ and $y \neq z$, and ${}_zQ_\phi(x,z) := 0$ for $x \in \mathbb{X}$.

Assumption HT. *There is a state $\ell \in \mathbb{X}$ and a constant K^* satisfying $\|\sum_{n=0}^{\infty} {}_\ell Q_\phi^n\| \leq K^* < \infty$ for all $\phi \in \mathbb{F}$.*

Assumption HT is equivalent to state ℓ being recurrent under all stationary policies, which according to [7] implies that Assumption HT can be checked in strongly polynomial time. We remark that any MDP satisfying Assumption HT is unichain, and that in general the problem of checking if an MDP is unichain is NP-hard [16].

For the initial state $x \in \mathbb{X}$, the *average cost* incurred under $\phi \in \mathbb{F}$ is $w^\phi(x) := \limsup_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} Q_\phi^n c_\phi(x)$. A policy ϕ_* is *average-cost optimal* if $w^{\phi_*}(x) = \inf_{\phi\in\mathbb{F}} w^\phi(x)$ for all $x \in \mathbb{X}$.

3. UNDISCOUNTED TOTAL COSTS

Let μ be a function satisfying the conditions in Proposition 1. When the constant K in Assumption T is fixed, a total-cost optimal policy can be computed in strongly polynomial time by using the transformation in [5, Sec. 3.1] of the original problem into a discounted MDP, with discount factor $(K-1)/K$, whose transition rates are stochastic. It follows from [5, Prop. 2] that a total-cost optimal policy for the original transient MDP can be computed by solving the following LP.

$$\begin{aligned}
\text{minimize} \quad & \sum_{x\in\mathbb{X}} \sum_{a\in A(x)} \mu(x)^{-1} c(x,a) z_{x,a} \\
\text{such that} \quad & \sum_{a\in A(x)} z_{x,a} - \sum_{y\in\mathbb{X}} \sum_{a\in A(y)} \frac{q(x|y,a)\mu(x)}{\mu(y)} z_{y,a} = 1, \\
& \qquad\qquad\qquad\qquad\qquad\qquad x \in \mathbb{X}, \\
& z_{x,a} \geq 0,\ x \in \mathbb{X},\ a \in A(x).
\end{aligned}$$

The estimates for discounted MDPs in [15] imply the following complexity estimates for this LP.

Proposition 2. *The block-pivoting simplex method that corresponds to Howard's policy iteration algorithm for discounted MDPs needs $O((m-n)K \log K)$ iterations to solve the above LP. In addition, the simplex method with Dantzig's rule needs at most $O(n(m-n)K \log K)$ iterations to solve the above LP.*

We remark that the above estimate for Howard's policy iteration algorithm matches the one in [3] for transient MDPs, which was obtained without reducing the original problem to a discounted one. In fact, it follows from the definition of the transformation in [5, Sec. 3.1] that Howard's policy iteration for the constructed discounted MDP, which is equivalent to a block-pivoting simplex method for the above LP, corresponds to Howard's policy iteration for the original transient MDP. On the other hand, it can be shown that applying the simplex method with Dantzig's rule to the above LP is different than applying this version of the simplex method to the LP formulation for transient MDPs considered in [3]. In addition, observe that, when K is fixed, the above estimate for the simplex method with Dantzig's rule is better than the one for Dantzig's rule in [3].

Using the estimates in [3], it can be shown that a suitable function μ can be computed using $O((m-n)K \log K)$ iterations of Howard's policy iteration algorithm. Since each iteration of both the simplex and policy iteration methods can be completed using $O(n^3 + mn)$ arithmetic operations, the preceding implies the following theorem.

Theorem 3. *Suppose the constant K in Assumption T is fixed. Then both the block-pivoting simplex method corresponding to Howard's policy iteration algorithm for discounted MDPs, as well as the simplex method with Dantzig's rule, can be used to compute a total-cost optimal policy in strongly polynomial time.*

4. AVERAGE COSTS PER UNIT TIME

In this section, we assume that the transition rates q are stochastic. According to Proposition 1, there is a function $\mu^* : \mathbb{X} \to [1,\infty)$ that satisfies $\mu^* \leq K^*$ and $\mu^*(x) \geq 1 + \sum_{y\in\mathbb{X}\setminus\{\ell\}} q(y|x,a)\mu^*(y)$ for all $x \in \mathbb{X}$ and $a \in A(x)$. When the constant K^* in Assumption HT is fixed, an average-cost optimal policy can be computed in strongly polynomial time by transforming the original problem into a discounted one with discount factor $(K^*-1)/K^*$ using the transformation in [5, Sec. 4.1]. It follows from [5, Prop. 8] that an average-cost optimal policy for the original MDP can be computed by solving the following LP.

$$\text{minimize} \quad \sum_{x\in\mathbb{X}}\sum_{a\in A(x)} \mu^*(x)^{-1}c(x,a)z_{x,a}$$

$$\text{such that}$$

$$\sum_{a\in A(x)} z_{x,a} - \sum_{y\in\mathbb{X}}\sum_{a\in A(y)} \frac{q(x|y,a)\mu^*(x)}{\mu^*(y)} z_{y,a} = 1,\ x\neq\ell,$$

$$\sum_{a\in A(\ell)} z_{\ell,a} - \sum_{y\in\mathbb{X}}\sum_{a\in A(y)} \frac{\mu^*(y)-1-\sum_{z\neq\ell} q(z|y,a)\mu^*(z)}{\mu^*(y)} z_{y,a} = 1,$$

$$z_{x,a}\geq 0,\ x\in\mathbb{X},\ a\in A(x).$$

The estimates for discounted MDPs in [15] imply the following complexity estimates for this LP.

PROPOSITION 4. *The block-pivoting simplex method that corresponds to Howard's policy iteration algorithm for discounted MDPs needs $O((m-n)K^*\log K^*)$ iterations to solve the above LP. In addition, the simplex method with Dantzig's rule needs at most $O(n(m-n)K^*\log K^*)$ iterations to solve the above LP.*

The following theorem can then be proven in a way analogous to the case of transient total-cost MDPs.

THEOREM 5. *Suppose the constant K^* in Assumption HT is fixed. Then both the block-pivoting simplex method corresponding to Howard's policy iteration algorithm for discounted MDPs, as well as the simplex method with Dantzig's rule, can be used to compute an average-cost optimal policy in strongly polynomial time.*

It follows from the definition of the transformation in [5, Sec. 4.1] that applying Howard's policy iteration algorithm to the constructed discounted MDP, which is equivalent to a block-pivoting simplex method for the above LP, corresponds to a block-pivoting simplex method for the LP that is typically used to solve unichain average-cost MDPs [11, Sec. 4.6]. On the other hand, this latter LP, even when Assumption HT holds with a constant K^*, may not satisfy the sufficient conditions given in [12] under which the simplex method with Dantzig's rule is strongly polynomial.

Acknowledgements. This research was partially supported by NSF grants CMMI-1335296 and CMMI-1636193.

5. REFERENCES

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

[2] R. J. Boucherie and N. M. van Dijk. *Markov Decision Processes in Practice*. Springer International Publishing, 2017.

[3] E. V. Denardo. Nearly strongly polynomial algorithms for transient dynamic programs. Preprint, February 1, 2016.

[4] E. A. Feinberg and J. Huang. Strong polynomiality of policy iterations for average-cost MDPs modeling replacement and maintenance problems. *Operations Research Letters*, 41:249–251, 2013.

[5] E. A. Feinberg and J. Huang. On the reduction of total-cost and average-cost MDPs to discounted MDPs. Preprint, `http://arxiv.org/abs/1507.00664`, 2015.

[6] E. A. Feinberg, J. Huang, and B. Scherrer. Modified policy iteration algorithms are not strongly polynomial for discounted dynamic programming. *Operations Research Letters*, 42:429–431, 2014.

[7] E. A. Feinberg and F. Yang. On polynomial cases of the unichain classification problem for Markov decision processes. *Operations Research Letters*, 36:527–530, 2008.

[8] T. D. Hansen, P. B. Miltersen, and U. Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM*, 60:1–16, 2013.

[9] R. Hollanders, J. Delvenne, and R. M. Jungers. The complexity of policy iteration is exponential for discounted Markov decision processes. In *Proceedings of the 51st IEEE Conference on Decision and Control*, 2012.

[10] R. A. Howard. *Dynamic Programming and Markov Processes*. The MIT Press, 1960.

[11] L. C. M. Kallenberg. *Linear Programming and Finite Markovian Control Problems*. Mathematisch Centrum, 1983.

[12] T. Kitahara and S. Mizuno. A bound for the number of different basic solutions generated by the simplex method. *Mathematical Programming Series A*, 137:579–586, 2013.

[13] I. Post and Y. Ye. The simplex method is strongly polynomial for deterministic Markov decision processes. *Mathematics of Operations Research*, 40:859–868, 2015.

[14] U. G. Rothblum and A. F. Veinott. Markov branching decision chains: Immigration-induced optimality. Technical Report 45, Department of Operations Research, Stanford University, 1992.

[15] B. Scherrer. Improved and generalized upper bounds on the complexity of policy iteration. *Mathematics of Operations Research*, 41:758–774, 2016.

[16] J. N. Tsitsiklis. NP-hardness of checking the unichain condition in average cost MDPs. *Operations Research Letters*, 35:319–323, 2007.

[17] A. F. Veinott. Discrete dynamic programming with sensitive discount optimality criteria. *The Annals of Mathematical Statistics*, 40:1635–1660, 1969.

[18] A. F. Veinott. Markov decision chains. In G. B. Dantzig and B. C. Eaves, editors, *Studies in Optimization, MAA Studies in Mathematics Vol. 10*, pages 124–159. Mathematical Association of America, 1974.

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

[20] 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*, 36:593–603, 2011.

[21] A. Zadorojniy, G. Even, and A. Shwartz. A strongly polynomial algorithm for controlled queues. *Mathematics of Operations Research*, 34:992–1007, 2009.