The Simplex Method is Strongly Polynomial for the Markov Decision Problem with a Fixed Discount Rate

Yinyu Ye *

April 20, 2010; revised May 14, 2010

Abstract

In this short paper we prove that the classic simplex method with the most-negative-reduced-cost pivoting rule (Dantzig 1947) for solving the Markov decision problem (MDP) with a fixed discount rate is a *strongly* polynomial-time algorithm. The result seems surprising since this very pivoting rule was shown to be exponential for solving a general linear programming (LP) problem, and the simplex (or simple policy iteration) method with the smallest-index pivoting rule was shown to be exponential for solving an MDP problem regardless of discount rates. As a corollary, the policy-iteration method (Howard 1960) is also a *strongly* polynomial-time algorithm for solving the MDP with a fixed discount rate.

1 Introduction

The infinite-horizon discounted Markov Decision Process (MDP) is one of the most fundamental decision models in mathematical, physical, management, and social sciences. Its applications include dynamic planning, reinforcement learning, social networking, and almost all other sequential decision makings. MDP is a special class of linear programming (LP), where general LP has a standard form

$$\begin{array}{rll} \text{Primal: minimize} & \mathbf{c}^T\mathbf{x} & \\ \text{subject to} & A\mathbf{x} = \mathbf{b}, & \\ & \mathbf{x} \geq \mathbf{0}, & \end{array} \tag{1}$$

*Department of Management Science and Engineering, Stanford University, Stanford, CA 94305; E-mail: yinyu-ye@stanford.edu. This researcher is supported in part by NSF Grant GOALI 0800151 and AFOSR Grant FA9550-09-1-0306.

and its dual

$$\begin{array}{rl} \text{Dual: maximize} & \mathbf{b}^T\mathbf{y} \\ \text{subject to} & \mathbf{s} = \mathbf{c} - A^T\mathbf{y} \geq \mathbf{0}, \end{array} \tag{2}$$

where $A \in \mathbb{R}^{m\times n}$ is a given real matrix with rank m, $\mathbf{c} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^m$ are given real vectors, and $\mathbf{x} \in \mathbb{R}^n$ and $(\mathbf{y} \in \mathbb{R}^m, \mathbf{s} \in \mathbb{R}^n)$ are unknown real decision vectors. Vector $\mathbf{s}$ is often called dual slack vector.

Then, due to de Ghellinck [6], D'Epenoux [7] and Manne [12], the infinite-horizon discounted MDP can be written as:

$$\begin{array}{lccccc} \text{minimize} & \mathbf{c}_1^T\mathbf{x}_1 & \ldots + \mathbf{c}_i^T\mathbf{x}_i + \ldots & +\mathbf{c}_k^T\mathbf{x}_k & & \\ \text{subject to} & (I-\gamma P_1)\mathbf{x}_1 & \ldots + (I-\gamma P_i)\mathbf{x}_i + \ldots & +(I-\gamma P_k)\mathbf{x}_k & = & \mathbf{e}, \\ & \mathbf{x}_1, & \ldots \quad \mathbf{x}_i, \quad \ldots & \mathbf{x}_k, & \geq & \mathbf{0}. \end{array} \tag{3}$$

where I is the $m \times m$ identity matrix, P_i is an $m \times m$ Markov or column stochastic matrix such that

$$\mathbf{e}^T P_i = \mathbf{e}^T \quad \text{and} \quad P_i \geq \mathbf{0}, \quad i = 1, \ldots, k,$$

and $\mathbf{e}$ is the vector of all ones. Here, decision vector of $\mathbf{x}_i \in \mathbb{R}^m$ is the policy vector where each state takes the ith action from its action set, and $\mathbf{c}_i$ is the cost vector associated with the policy vector. In the problem, γ is the so-called discount factor or rate such that

$$\gamma = \frac{1}{1+r} \leq 1,$$

where r is the interest rate and it is assumed strictly positive so that $0 \leq \gamma < 1$. When $\gamma = 1$, problem (3) is infeasible.

Comparing to the LP standard form, we have

$$A = [I - \gamma P_1, \ldots, I - \gamma P_k] \in \mathbb{R}^{m\times mk}, \quad \mathbf{b} = \mathbf{e} \in \mathbb{R}^m, \quad \text{and} \quad \mathbf{c} = (\mathbf{c}_1; \ldots; \mathbf{c}_k) \in \mathbb{R}^{mk}.$$

The dual (by adding slack variables) of (3) is given by:

$$\begin{array}{lcccc} \text{maximize} & \mathbf{e}^T\mathbf{y} & & \\ \text{subject to} & (I-\gamma P_1)^T\mathbf{y} + \mathbf{s}_1 & = & \mathbf{c}_1, \\ & \ldots & \ldots & \ldots \\ & (I-\gamma P_i)^T\mathbf{y} + \mathbf{s}_i & = & \mathbf{c}_i, \\ & \ldots & \ldots & \ldots \\ & (I-\gamma P_k)^T\mathbf{y} + \mathbf{s}_k & = & \mathbf{c}_k, \\ & \mathbf{s}_1, \ldots, \mathbf{s}_i, \ldots, \mathbf{s}_k & \geq & \mathbf{0}. \end{array} \tag{4}$$

The optimal solution of the dual expresses the maximal value of each state when the primal optimal policy solution is adapted. Indeed, the MDP problem is to find a best action, among its k actions, for each state so that the total cost is minimized or its dual

value is maximized. An optimal solution or policy vector to the MDP problem contains a specific action for each state, which forms an optimal basic feasible solution (BFS) to linear program (3).

There are several major events on developing methods for solving MDPs. Bellman (1957) [1] developed an approximate method, called the value-iteration method, to approximate the maximal state values. The other best known method is due to Howard (1960) [9] and is known as the policy-iteration method, which generate an optimal policy vector in a finite number of iterations. Since it was discovered in 1960 that the MDP has an LP representation, the simplex method of Dantzig (1947) [5] should be applicable to solving MDPs as well.

As the notion of computational complexity emerged, there were tremendous efforts in analyzing the complexity of the MDP, and the policy-iteration or the simplex method. On the positive side, Papadimitriou and Tsitsiklis [15] showed in 1987 that an MDP with deterministic transitions (i.e., the entries of P_i are all 0's and 1's) can be solved in *strongly* polynomial-time (i.e., the number of arithmetic operations is bounded by a polynomial of the numbers of states and actions as a Minimum-Mean-Cost-Cycle problem. Tseng [17] in 1990 showed that the value-iteration method generates an optimal policy in polynomial-time (i.e., the number of arithmetic operations is bounded by a polynomial of the numbers of states and actions, and the bit-size of the input data) for a fixed discount rate γ, built upon Bertsekas' 1987 work [2] that the value-iteration method converges to the optimal policy in finite number of iterations. Puterman [16] in 1994 showed that the policy-iteration method converges no more slowly than the value iteration method, so that it is also a polynomial-time algorithm for MDPs for a fixed discount rate. Then, Ye [21] in 2005 developed a *strongly* polynomial-time (combinatorial) interior-point algorithm for the MDP for a fixed discount rate γ, that is, the number of arithmetic operations is bounded by a polynomial of the numbers of states and actions when γ is fixed.

In terms of the worst-case complexity bound on the number of arithmetic operations, they (without a constant factor) are summarized in the following table for $k = 2$ (see Littman et al. [11], Mansour and Singh [13], Ye [21], and references therein).

Value-Iteration	Policy-Iteration	LP-Algorithms	Combinatorial IP
$\frac{m^2 L(P_i,\mathbf{c}_i,\gamma)}{1-\gamma}$	$\min\left\{m^3 \cdot \frac{2^m}{m}, \frac{m^2 L(P_i,\mathbf{c}_i,\gamma)}{1-\gamma}\right\}$	$m^3 L(P_i, \mathbf{c}_i, \gamma)$	$m^4 \cdot \log \frac{1}{1-\gamma}$

where $L(P_i, \mathbf{c}_i, \gamma)$ is the toal bit-size of the MDP input data $(P_i, \mathbf{c}_i, \gamma)$, $i = 1, \ldots, k$. One can see from the table, both the value-iteration and policy-iteration methods are *polynomial-time* algorithms if the discount rate $0 \leq \gamma < 1$ is fixed, but they are not *strongly* polynomial where the running time should be a polynomial only in m (or mk).

However, the policy-iteration method has been remarkably successful and shown to be a most effective and widely used method in practice where the number of iterations is

typically bounded by $O(mk)$. It turns out that the policy-iteration method is actually the simplex method with multiple pivots at each iteration; and the simplex method also remains one of the very few extremely effective methods for general LP solvers; see Bixby [3]. In the past 50 years, many efforts have been put to resolve the worst-case complexity issue of the policy-iteration method or the simplex method: Are the policy-iteration and the simplex methods *strongly* polynomial-time algorithms? So far, most of results were negative. The result of Klee and Minty [10] emerged in 1972 that the simplex method, with Dantzig's original most-negative-reduced-cost pivoting rule, necessarily takes an exponential number of iterations to solve a carefully designed LP problem. Later, a similar negative result of Melekopoglou and Condon [14] showed that one simple policy-iteration method, where in each iteration only the action for the state with the smallest index is updated, needs an exponential number of iterations to compute an optimal policy for a specific MDP problem regardless of discount rates (i.e., even $\gamma < 1$ is fixed). Most recently, Fearnley (2010) [8] showed that the policy-iteration method needs an exponential number of iterations for a undiscounted (i.e., $\gamma = 1$) but finite-horizon MDP. The only worst-case iteration upper bound for the policy-iteration method is $\frac{k^m}{m}$ given in 1999 by Mansour and Singh [13], and many researchers believe that the policy-iteration method is *strongly* polynomial.

In this short paper, we prove that the classic simplex method, or the simple policy-iteration method, with the most-negative-reduced-cost pivoting rule, is indeed a *strongly* polynomial-time algorithm for MDP with fixed discount rate $0 \leq \gamma < 1$. The number of its iterations is bounded by

$$\frac{m^2(k-1)}{1-\gamma} \cdot \log\left(\frac{m^2}{1-\gamma}\right).$$

Since the policy-iteration method with the all-negative-reduced-cost pivoting rule is at least as good as the the simple policy-iteration method, it is also a *strongly* polynomial-time algorithm with the same iteration complexity bound. Therefore, there is no complexity difference between the simplex method and interior-point algorithms for MDP with a fixed discount rate. Our proof is based on a combinatorial cross-over event similar to the one in Vavasis and Ye [18, 21]. We remark that, if the discount rate is an input, it remains open whether or not the policy-iteration method is polynomial for MDP, or whether or not there exists a *strongly* polynomial-time algorithm for MDP or LP in general.

2 MDP Properties and The Simplex Method

We first describe few general LP and MDP theorems and the classic simplex method. Without loss of generality and for simplicity, we fix k, the number of possible actions taken

by each state, to 2 in the rest of the paper:

$$\begin{array}{rccccc} \text{minimize} & \mathbf{c}_1^T\mathbf{x}_1 & & \mathbf{c}_2^T\mathbf{x}_2 & & \\ \text{subject to} & (I-\gamma P_1)\mathbf{x}_1 & +(I-\gamma P_2)\mathbf{x}_2 & = & \mathbf{e}, \\ & \mathbf{x}_1, & & \mathbf{x}_2 & \geq & \mathbf{0}; \end{array} \tag{5}$$

with its dual

$$\begin{array}{rccc} \text{maximize} & \mathbf{e}^T\mathbf{y} & & \\ \text{subject to} & (I-\gamma P_1)^T\mathbf{y}+\mathbf{s}_1 & = & \mathbf{c}_1, \\ & (I-\gamma P_2)^T\mathbf{y}+\mathbf{s}_2 & = & \mathbf{c}_2, \\ & \mathbf{s}=(\mathbf{s}_1;\mathbf{s}_2) & \geq & \mathbf{0}. \end{array} \tag{6}$$

Comparing to the LP standard form, we have

$$A=[I-\gamma P_1,\ I-\gamma P_2]\in \mathbb{R}^{m\times 2m}, \quad \mathbf{b}=\mathbf{e}\in\mathbb{R}^m, \quad \text{and} \quad \mathbf{c}=(\mathbf{c}_1;\mathbf{c}_2)\in\mathbb{R}^{2m}.$$

2.1 MDP Properties

The *optimality conditions* for all optimal solution of general LP may be written as follows:

$$\begin{array}{rcl} A\mathbf{x} & = & \mathbf{b}, \\ A^T\mathbf{y}+\mathbf{s} & = & \mathbf{c}, \\ SX\mathbf{e} & = & \mathbf{0}, \\ \mathbf{x}\geq\mathbf{0}, & & \mathbf{s}\geq\mathbf{0} \end{array}$$

where X denotes diag($\mathbf{x}$) and S denotes diag($\mathbf{s}$), and $\mathbf{0}$ denotes the vector of all 0's, and the third equation is often referred as the complementarity condition.

If LP has an optimal solution pair, then there exists a BFS pair $(\mathbf{x}_B,\mathbf{y})$, where B is a column index set such that $A_B\mathbf{x}_B=\mathbf{b}$ and $A_B^T\mathbf{y}=\mathbf{c}_B$, where A_B consists of m independent columns of A, and $\mathbf{x}_B\geq\mathbf{0}$ and $\mathbf{c}-A^T\mathbf{y}\geq\mathbf{0}$. Here, sub-vector $\mathbf{x}_B$ contains all x_j for $j\in B\subset\{1,\ldots,n\}$, and the variables of $\mathbf{x}_B$ are called basic variables. Note that any feasible basis A_B of the MDP has the Leontief substitution form

$$A_B=I-\gamma P$$

where P is an $m\times m$ Markov matrix chosen from columns of $[P_1,\ P_2]$, and the reverse is also true. (This can be seen from that, otherwise, the basis has at least one row consisting of all non-positive elements so that its inner product with $\mathbf{x}_B(\geq\mathbf{0})$ is non-positive, but the right-hand side is strictly positive.) The following lemma is given in [21] whose proof was based on Dantzig [4, 5], and Veinott [19].

Lemma 1 *The MDP has the following properties:*

1. *The feasible set of the primal MDP is bounded. More precisely,*

$$\mathbf{e}^T\mathbf{x} = \frac{m}{1-\gamma},$$

where $\mathbf{x} = (\mathbf{x}^1; \mathbf{x}^2)$ *of all feasible solutions.*

2. *Let* $\hat{\mathbf{x}}$ *be a BFS of the MDP. Then, any basic variable, say* $\hat{x}_i$*, has its value*

$$1 \leq \hat{x}_i \leq \frac{m}{1-\gamma}.$$

2.2 The Simplex and policy-iteration Method

Let us start with $\mathbf{x}_1$ being the initial basic feasible solution of (5) where the initial basic index set is denoted by B^0. Then, the MDP can be rewritten as an equivalent problem

$$\begin{array}{lcccc} \text{minimize} & & \bar{\mathbf{c}}_2^T\mathbf{x}_2 & & \\ \text{subject to} & (I-\gamma P_1)\mathbf{x}_1 & +(I-\gamma P_2)\mathbf{x}_2 & = & \mathbf{e}, \\ & \mathbf{x}_1, & \mathbf{x}_2 & \geq & \mathbf{0}. \end{array} \tag{7}$$

$\bar{\mathbf{c}}_2$ is called the reduced cost vector for the non-basic variables $\mathbf{x}_2$:

$$\bar{\mathbf{c}}_2 = \mathbf{c}_2 - (I-\gamma P_2)^T\mathbf{y}^0$$

and

$$\mathbf{y}^0 = (I-\gamma P_1)^{-T}\mathbf{c}_1.$$

The initial primal basic feasible solution is given by

$$\mathbf{x}^0 = (\mathbf{x}_1^0 = (I-\gamma P_1)^{-1}\mathbf{e};\ \mathbf{x}_2^0 = \mathbf{0}).$$

If $\bar{\mathbf{c}}_2 \geq \mathbf{0}$, then the current BFS is optimal. Otherwise, let $\Delta^0 = -\min(\bar{\mathbf{c}}_2)$ with $(\bar{\mathbf{c}}_2)_{\bar{i}} = -\Delta^0$. Then, the classic simplex method takes $(\mathbf{x}_2)_{\bar{i}}$ as the in-coming basic variable to replace the old one $(\mathbf{x}_1)_{\bar{i}}$, and the method repeats with the new BFS denoted by $\mathbf{x}^1$. The method will break a tie arbitrarily, and it updates exact one state action in one iteration, that is, it only updates the state with the most negative reduced cost. This is called the simplex or simple policy-iteration method with the most-negative-reduced-cost update or pivoting rule.

The policy-iteration method with the all-negative-reduced-cost update or pivoting rule is to update every state who has a negative reduced cost (for $k > 2$ each state will update one of its most negative reduced cost). Such a parallel update or pivot is possible due to the special structure of MDP, and it may not work for general LP. Again, the method repeats with the new BFS. Thus, both methods generate a sequence of BFSs denoted by $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^t, ...$

3 Proof of Strong Polynomiality

First, we have

Lemma 2 *Let z^* be the minimal objective value of (5). Then,*

$$z^* \geq \mathbf{c}^T\mathbf{x}^0 - \frac{m}{1-\gamma} \cdot \Delta^0.$$

Moreover,

$$\mathbf{c}^T\mathbf{x}^1 - z^* \leq \left(1 - \frac{1-\gamma}{m}\right)\left(\mathbf{c}^T\mathbf{x}^0 - z^*\right).$$

Proof. From Lemma 1, for problem (7), its minimal objective value is bounded from below by $-\frac{m}{1-\gamma} \cdot \Delta^0$, that is, with all the solution mass put to state $\bar{i}$. On the other hand, the minimal objective value of (7) differs from the one of (5) exactly by $\mathbf{c}^T\mathbf{x}^0$, the initial BFS objective value of (5). Thus, we have

$$z^* \geq \mathbf{c}^T\mathbf{x}^0 - \frac{m}{1-\gamma} \cdot \Delta^0.$$

Since at the new BFS $\mathbf{x}^1$, the new basic variable value for state $\bar{i}$ is greater than or equal to 1 from Lemma 1, the objective value of the new BFS of problem (7) is decreased by at least Δ^0. Thus, for problem (5),

$$\mathbf{c}^T\mathbf{x}^0 - \mathbf{c}^T\mathbf{x}^1 \geq \Delta^0 \geq \frac{1-\gamma}{m}\left(\mathbf{c}^T\mathbf{x}^0 - z^*\right),$$

which leads to the desired inequality. ∎

Lemma 3 *If the initial BFS $\mathbf{x}^0$ is not optimal, then there is $i^0 \in B^0$ such that*

$$(\mathbf{s}_1^*)_{i^0} \geq \frac{1-\gamma}{m^2}\left(\mathbf{c}^T\mathbf{x}^0 - z^*\right),$$

where $\mathbf{s}^*$ is an optimal dual slack vector of (6). And for any basic feasible solution $\mathbf{x}^t$ of (5), $t \geq 1$,

$$(\mathbf{x}_1^t)_{i^0} \leq \frac{m^2}{1-\gamma} \cdot \frac{\mathbf{c}^T\mathbf{x}^t - z^*}{\mathbf{c}^T\mathbf{x}^0 - z^*}.$$

Proof. Since

$$\mathbf{c}^T\mathbf{x}^0 - z^* = (\mathbf{s}^*)^T\mathbf{x}^0 = (\mathbf{s}_1^*)^T\mathbf{x}_1^0 = \sum_{i=1}^{m}(\mathbf{s}_1^*)_i(\mathbf{x}_1^0)_i,$$

there must be an $i^0 \in B^0$ such that

$$(\mathbf{s}_1^*)_{i^0}(\mathbf{x}_1^0)_{i^0} \geq \frac{1}{m}\left(\mathbf{c}^T\mathbf{x}^0 - z^*\right).$$

Then, from Lemma 1 we have $(\mathbf{x}_1^0)_{i^0} \leq \frac{m}{1-\gamma}$ so that

$$(\mathbf{s}_1^*)_{i^0} \geq \frac{1-\gamma}{m^2}\left(\mathbf{c}^T\mathbf{x}^0 - z^*\right).$$

Furthermore, for any basic feasible solution $\mathbf{x}^t$,

$$\mathbf{c}^T\mathbf{x}^t - z^* = (\mathbf{s}^*)^T\mathbf{x}^t \geq (\mathbf{s}^*)_{i^0}(\mathbf{x}^t)_{i^0},$$

so that

$$(\mathbf{x}_1^t)_{i^0} \leq \frac{\mathbf{c}^T\mathbf{x}^t - z^*}{(\mathbf{s}_1^*)_{i^0}} \leq \frac{m^2}{1-\gamma} \cdot \frac{\mathbf{c}^T\mathbf{x}^t - z^*}{\mathbf{c}^T\mathbf{x}^0 - z^*}.$$

∎

From Lemma 2, after t iterations of the simplex method, we have

$$\frac{\mathbf{c}^T\mathbf{x}^t - z^*}{\mathbf{c}^T\mathbf{x}^0 - z^*} \leq \left(1 - \frac{1-\gamma}{m}\right)^t.$$

Therefore, after $\frac{m}{1-\gamma} \cdot \log\left(\frac{m^2}{1-\gamma}\right)$ iterations from the initial BFS $\mathbf{x}^0$, we must have, from Lemma 3,

$$(\mathbf{x}_1^t)_{i^0} \leq \frac{m^2}{1-\gamma} \cdot \frac{\mathbf{c}^T\mathbf{x}^t - z^*}{\mathbf{c}^T\mathbf{x}^0 - z^*} < 1,$$

for all $t \geq \frac{m}{1-\gamma} \cdot \log\left(\frac{m^2}{1-\gamma}\right)$. But for any basic variable, its value should be greater than or equal to 1 from Lemma 1, hence it must be true $(\mathbf{x}_1^t)_{i^0} = 0$. This leads to our key result:

Theorem 1 *There is a basic variable in the initial basic feasible solution $\mathbf{x}^0$ that would never be in the basis again after $\frac{m}{1-\gamma} \cdot \log\left(\frac{m^2}{1-\gamma}\right)$ iterations of the simplex method with the most-negative-reduced-cost pivoting rule.*

The event described in Theorem 1 can be viewed as the cross-event of Vavasis and Ye [18, 21]: a variable, although we don't know which one, was a basic variable initially but it will never be a basic variable after a certain number of iterations during the iterative process of the simplex or simple policy-iteration method with the most-negative-reduced-cost pivoting rule. Clearly, these cross-over events can happen only $(mk - m)$ times for any k (we can subtract it by m because the cross event can only happen to a variable that is not in any optimal basis and there should be at least m basic variables in an optimal BFS), thus, we reach our final conclusion:

Theorem 2 *The simplex or simple policy-iteration method with the most-negative-reduced-cost pivoting rule of Dantzig for solving the Markov decision problem with a fixed discount rate* $0 \leq \gamma < 1$ *is a strongly polynomial-time algorithm. It terminates at most* $\frac{m^2(k-1)}{1-\gamma} \cdot \log\left(\frac{m^2}{1-\gamma}\right)$ *iterations, where each iteration uses* $O(m^2k)$ *arithmetic operations.*

As a corollary, we have

Corollary 1 *The policy-iteration method of Howard for solving the Markov decision problem with a fixed discount rate* $0 \leq \gamma < 1$ *is a strongly polynomial-time algorithm. It terminates at most* $\frac{m^2(k-1)}{1-\gamma} \cdot \log\left(\frac{m^2}{1-\gamma}\right)$ *iterations.*

4 Extensions and Remarks

The result can be extended to other MDPs where every basic feasible matrix of (3) exhibits the Leontief substitution form

$$A_B = I - P$$

for some nonnegative square matrix P with $P \geq \mathbf{0}$ and its spectral radius $\rho(P) \leq \gamma$ for a fixed $\gamma < 1$. This includes MDPs with sub-stochastic matrices and transient cases; see Veinott [20]. Note that the inverse of $(I - P)$ has the expansion form

$$(I - P)^{-1} = I + P + P^2 + \ldots$$

and

$$\|(I - P)^{-1}\mathbf{e}\|_2 \leq \|\mathbf{e}\|_2(1 + \gamma + \gamma^2 + \ldots) = \frac{\sqrt{m}}{1 - \gamma}$$

so that

$$\|(I - P)^{-1}\mathbf{e}\|_1 \leq \frac{m}{1 - \gamma}.$$

Thus, Lemma 1 still holds, and together with other Lemmas, they imply

Corollary 2 *Let every basic feasible matrix in the form* $(I - P)$ *of an MDP have* $P \geq \mathbf{0}$ *and its spectral radius less than or equal to a fixed* $\gamma < 1$*. Then, the simplex method with the most-negative-reduced-cost pivoting rule and the policy-iteration method are strongly polynomial-time algorithms. Each of them terminates at most* $\frac{m^2(k-1)}{1-\gamma} \cdot \log\left(\frac{m^2}{1-\gamma}\right)$ *iterations.*

One observation from our worst-case analyses is that there seems no difference between the simple policy iteration method, as long as the most-negative-reduced-cost pivoting rule is adapted, and the general policy-iteration method that makes multiple pivots in each iteration.

Therefore, we remark that the pivoting rule seems making a big difference. As we mentioned earlier, for the MDP with a fixed discount rate, the simplex or simple policy iteration method with the smallest-index pivoting rule was shown to be exponential, but the method with the most-negative-reduced-cost pivoting rule is strongly polynomial. On the other hand, the most-negative-reduced-cost pivoting rule is exponential for solving some LP problems. Thus, searching for suitable pivoting rules for solving different problems is the key, and one can not rule out the simplex method simply because the behavior of one rule on one problem is shown to be exponential.

A further research direction would be: can the simplex method or the policy-iteration method be strongly polynomial regardless of discount rates? Or, is there a strongly polynomial-time algorithm at all for solving the MDP regardless of discount rates?

References

[1] R. Bellman. *Dynamic Programming.* Princeton University Press, Princeton, New Jersey, 1957.

[2] D. P. Bertsekas. *Dynamic Programming, Deterministic and Stochastic Models.* Prentice-Hall, Englewood Cliffs, New Jersey, 1987.

[3] R. E. Bixby, *Progress in linear programming*, ORSA J. on Comput. **6**:1 (994) 15–22.

[4] G. B. Dantzig, *Optimal solutions of a dynamic Leontief model with substitution*, Econometrica **23** (1955), 295–302.

[5] G. B. Dantzig. *Linear Programming and Extensions.* Princeton University Press, Princeton, New Jersey, 1963.

[6] G. de Ghellinck, *Les Probléms de Décisions Séquentielles*, Cahiers du Centre d'Etudes de Recherche Opérationnelle **2** (1960), 161–179.

[7] F. D'Epenoux, *A Probabilistic Production and Inventory Problem*, Management Science **10** (1963), 98–108; Translation of an article published in Revue Fran**c**caise de Recherche Opérationnelle **14** (1960).

[8] J. Fearnley, *Exponential lower bounds for policy iteration*, arXiv:1003.3418v1, (March 2010).

[9] R. A. Howard, *Dynamic Programming and Markov Processes.* MIT, Cambridge, Massachusetts, 1960.

[10] V. Klee and G. J. Minty. How good is the simplex method. In O. Shisha, editor, *Inequalities III*, Academic Press, New York, NY, 1972.

[11] M. L. Littman, T. L. Dean and L. P. Kaelbling, *On the complexity of solving Markov decision problems*, Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI–95), 1995, pp. 394–402.

[12] A. S. Manne, *Linear programming and sequential decisions*, Management Science **6** (1960), 259–267.

[13] Y. Mansour and S. Singh, *On the complexity of policy iteration*, Proceedings of the 15th International Conference on Uncertainty in AI, 1999, pp. 401–408.

[14] M. Melekopoglou and A. Condon, *On the complexity of the policy improvement algorithm for Markov Decision Processes*, INFORMS Journal on Computing **6**:2 (1994), 188-192.

[15] C. H. Papadimitriou and J. N. Tsitsiklis, *The complexity of Markov decision processes*, Mathematics of Operations Research, **12**:3 (1987), 441-450.

[16] M. L. Puterman. *Markov Decision Processes.* John & Wiley and Sons, New York, 1994.

[17] P. Tseng, *Solving H-horizon, stationary Markov decision problems in time propotional to* $\log(H)$, Operations Research Letters, **9**:5 (1990), 287-297.

[18] S. Vavasis and Y. Ye, *A primal-dual interior-point method whose running time depends only on the constraint matrix*, Mathematical Programming **74** (1996) 79–120.

[19] A. Veinott, *Extreme points of Leontief substitution systems*, Linear Algebra and its Applications **1** (1968) 181-194.

[20] A. Veinott, *Discrete dynamic programming with sensitive discount optimality criteria*, The Annals of Mathematical Statistics **40**:5 (1969) 1635-1660.

[21] Y. Ye, *A new complexity result on solving the Markov decision problem*, Mathematics of Operations Research, **30**:3 (2005), 733-749.