The Simplex and Policy-Iteration Methods are Strongly Polynomial for the Markov Decision Problem with a Fixed Discount Rate

Yinyu Ye *

April 20, 2010; revised November 30, 2010

Abstract

We prove that the classic policy-iteration method (Howard 1960), including the Simplex method (Dantzig 1947) with the most-negative-reduced-cost pivoting rule, is a *strongly* polynomial-time algorithm for solving the Markov decision problem (MDP) with a fixed discount rate. Furthermore, the computational complexity of the policy-iteration method (including the Simplex method) is *superior* to that of the only known strongly polynomial-time interior-point algorithm ([28] 2005) for solving this problem. The result is surprising since the Simplex method with the same pivoting rule was shown to be exponential for solving a general linear programming (LP) problem, the Simplex (or simple policy-iteration) method with the smallest-index pivoting rule was shown to be exponential for solving an MDP regardless of discount rates, and the policy-iteration method was recently shown to be exponential for solving a undiscounted MDP. We also extend the result to solving MDPs with sub-stochastic and transient state transition probability matrices.

1 Introduction of the Markov decision problem and its linear programming formulation

Markov decision problems (MDPs), named after Andrey Markov, provide a mathematical framework for modeling decision-making in situations where outcomes are partly random

*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 partly under the control of a decision maker. The MDP is one of the most fundamental models for studying a wide range of optimization problems solved via dynamic programming and reinforcement learning. Today, it has been used in a variety of areas, including management, economics, bioinformatics, electronic commerce, social networking, and supply chains.

More precisely, an MDP is a discrete-time stochastic control process. At each time step, the process is in some state i, and the decision maker may choose any action, say action j, that is available in state i. The process responds at the next time step by randomly moving into a new state i', and giving the decision maker a corresponding immediate cost $c^j(i, i')$.

Let m denote the total number of states. The probability that the process enters i' as its new state is influenced by the chosen state-action j. Specifically, it is given by a state transition probability distribution $p^j(i, i') \geq 0$, $i' = 1, \cdots, m$, and

$$\sum_{i'=1}^{m} p^j(i, i') = 1, \ \forall i = 1, \cdots, m.$$

Thus, the next state i' depends on the current state i and the decision maker's chosen state-action j, but is conditionally independent of all previous states and actions; in other words, the state transitions of an MDP possess the Markov property.

The key decision of MDPs is to find a (stationary) policy for the decision maker: a set function $\pi = \{\pi_1, \pi_2, \cdots, \pi_m\}$ that specifies the action π_i that the decision maker will choose when in state i, for $i = 1, \cdots, m$. The goal of the problem is to find a (stationary) policy π that will minimize some cumulative function of the random costs, typically the expected discounted sum over an infinite horizon:

$$\sum_{t=0}^{\infty} \gamma^t c^{\pi_{i^t}}(i^t, i^{t+1}),$$

where $c^{\pi_{i^t}}(i^t, i^{t+1})$ represents the cost, at time t, incurred to an individual who is in state i^t and takes action π_{i^t}.

Here γ is the discount rate, where $\gamma \geq 0$ and is assumed to be strictly less than 1 in this paper. This MDP problem is called the infinite-horizon discounted Markov decision problem (DMDP), which serves as the core model for MDPs. Because of the Markov property, there is an optimal *stationary* policy, or policy for short, for the DMDP so that it can indeed be written as a function of i only; that is, π is independent of time t as described above.

Let k_i be the number of state-actions available in state i, $i = 1, \cdots, m$, and let

$$\mathcal{A}_1 = \{1, 2, \cdots, k_1\}, \quad \mathcal{A}_2 = \{k_1 + 1, k_1 + 2, \cdots, k_1 + k_2\}, ...$$

or, for $i = 1, 2, \cdots, m$ in general,

$$\mathcal{A}_i := \left\{ \left(\sum_{s=1}^{i-1} k_s \right) + 1, \left(\sum_{s=1}^{i-1} k_s \right) + 2, \cdots, \sum_{s=1}^{i} k_s \right\}.$$

Moreover, let $n = \sum_{i=1}^{m} k_i$, and the total n state-actions be ordered such that if $j \in \mathcal{A}_i$, then state-action j is controlled by state i. Note that the cardinality $|\mathcal{A}_i| = k_i$.

Suppose we know the state transition probability P and the cost function c, and we wish to calculate the policy that minimizes the expected discounted cost. Then a policy π would be associated with another array indexed by state, value vector $\mathbf{v} \in \mathbb{R}^m$, which contains cost-to-go values for all states. Furthermore, an optimal policy, $(\mathbf{v}^*, \pi^*)$, is then a fixed point of the following minimum cost operator,

$$\begin{array}{rcl} \pi_i^* := & \arg\min_{j \in \mathcal{A}_i} \left\{ \sum_{i'} p^j(i,i') \left(c^j(i,i') + \gamma v_{i'}^* \right) \right\}; \\ v_i^* := & \sum_{i'} p^{\pi_i^*}(i,i') \left(c^{\pi_i^*}(i,i') + \gamma v_{i'}^* \right), \ \forall i = 1, \cdots, m. \end{array} \tag{1}$$

Let $P_\pi \in \mathbb{R}^{m \times m}$ be the column stochastic matrix corresponding to a policy π, that is, the ith column of P_π be the probability distribution $p^{\pi_i}(i,i')$, $i' = 1, \cdots, m$. Then the equilibrium condition of (1) can be represented by a matrix form

$$\begin{array}{rcl} (I - \gamma P_{\pi^*}^T)\mathbf{v}^* = & \mathbf{c}_{\pi^*}, \\ (I - \gamma P_{\pi}^T)\mathbf{v}^* \leq & \mathbf{c}_{\pi}, \ \forall \pi, \end{array} \tag{2}$$

where the ith entry of column vector $\mathbf{c}_\pi \in \mathbb{R}^m$ equals $\sum_{i'} p^{\pi_i}(i,i') c^{\pi_i}(i,i')$.

Due to D'Epenoux [9] (also see Manne [17] and de Ghellinck [8]), the infinite-horizon discounted MDP can be formulated as a primal linear programming (LP) problem in the standard form

$$\begin{array}{rl} \text{minimize} & \mathbf{c}^T \mathbf{x} \\ \text{subject to} & A\mathbf{x} = \mathbf{b}, \\ & \mathbf{x} \geq \mathbf{0}, \end{array} \tag{3}$$

with the dual

$$\begin{array}{rl} \text{maximize} & \mathbf{b}^T \mathbf{y} \\ \text{subject to} & \mathbf{s} = \mathbf{c} - A^T \mathbf{y} \geq \mathbf{0}, \end{array} \tag{4}$$

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, $\mathbf{0}$ denotes the vector of all 0's, and $\mathbf{x} \in \mathbb{R}^n$ and $(\mathbf{y} \in \mathbb{R}^m, \mathbf{s} \in \mathbb{R}^n)$ are unknown primal and dual variables, respectively. Vector $\mathbf{s}$ is often called the dual slack vector. In what follows, "LP" stands for any of the following: "linear program", "linear programs", or "linear programming", depending on the context.

More precisely, the DMDP can be represented by the LP problems (3) and (4) with the following assignments of $(A, \mathbf{b}, \mathbf{c})$. First, the ith entry of the column vector $\mathbf{b} \in \mathbb{R}^n$ is 1 for all i, representing an initial population of individuals in state i. Secondly, the jth entry of

the column vector of $\mathbf{c} \in \mathbb{R}^n$ is the (expected) one-time unit cost of action j taken by a state. In particular, if $j \in \mathcal{A}_i$, then action j is controlled by state i and

$$c_j = \sum_{i'} p^j(i,i')c^j(i,i'). \tag{5}$$

The LP constraint matrix has the form

$$A = E - \gamma P \in \mathbb{R}^{m \times n}. \tag{6}$$

The jth column of P is the state transition probability distribution when the jth action is taken by a state. More precisely, for each state-action $j \in \mathcal{A}_i$, that is, each action j controlled by state i,

$$P_{i'j} = p^j(i,i'),\ \forall i' = 1, \cdots, m. \tag{7}$$

Finally, the ith element of the jth column of E is 1 if action j is controlled by state i and zero everywhere else:

$$E_{ij} = \begin{cases} 1, & \text{if } j \in \mathcal{A}_i, \\ 0, & \text{otherwise} \end{cases},\ \ \forall i = 1, \cdots, m,\ j = 1, \cdots, n. \tag{8}$$

Let $\mathbf{e}$ be the vector of all ones, where its dimension depends on the context. Then we have $\mathbf{b} = \mathbf{e}$, $\mathbf{e}^T P = \mathbf{e}$ (that is, P is a column stochastic matrix), $\mathbf{e}^T E = \mathbf{e}$, and $\mathbf{e}^T A = (1-\gamma)\mathbf{e}$.

The interpretations of the quantities defining the DMDP primal (3) and the DMDP dual (4) are as follows: $\mathbf{b} = \mathbf{e}$ means that there is one unit of the initial number of individuals in each state i. The jth entry, if $j \in \mathcal{A}_i$, of primal variables $\mathbf{x} \in \mathbb{R}^n$ is the state-action frequency for action j, or the expected present value of the number of times in which an individual is in state i and takes state-action j when $j \in \mathcal{A}_i$. Thus, solving the DMDP primal entails choosing state-action frequencies that minimize the expected present value sum, $\mathbf{c}^T\mathbf{x}$, of total costs subject to the conservation law $A\mathbf{x} = \mathbf{e}$. The conservation law ensures that for each state i, the expected present value of the number of individuals entering state i equals the expected present value of the number of individuals leaving i.

The DMDP dual variables $\mathbf{y} \in \mathbb{R}^m$ exactly represent the expected present cost-to-go values of the m states. Solving the dual entails choosing dual variables $\mathbf{y}$, one for each state i, together with $\mathbf{s} \in \mathbb{R}^n$ of slack variables, one for each state-action j, that maximizes $\mathbf{e}^T\mathbf{y}$ subject to $A^T\mathbf{y} + \mathbf{s} = \mathbf{c},\ \mathbf{s} \geq \mathbf{0}$ or simply $A^T\mathbf{y} \leq \mathbf{c}$. It is well known that there exist unique optimal $\mathbf{y}^*$ and $\mathbf{s}^*$ where, for each state i, y_i^* is the minimum expected present cost that an individual in state i and its progeny can incur.

A policy π of the original DMDP, containing exactly one action in $\mathcal{A}_i$ for each state i, actually corresponds to m basic variable indexes of a basic feasible solution (BFS) of the DMDP primal LP formulation. Obviously, we have a total of $\prod_{i=1}^m k_i$ different policies. Let matrix $A_\pi \in \mathbb{R}^{m \times m}$ (resp., P_π, E_π) be the columns of A (resp., P, E) with indexes in π. Then for a policy π, $E_\pi = I$ (where I is the identity matrix), so that A_π has the

Leontief substitution form $A_\pi = I - \gamma P_\pi$. It is also well known that A_π is nonsingular, has a nonnegative inverse and is a feasible basis for the DMDP primal. Let $\mathbf{x}^\pi$ be the BFS for a policy π in the DMDP primal form and let ν contain the rest indexes not in π.

Let $\mathbf{x}_\pi$ and $\mathbf{x}_\nu$ be the sub-vectors of $\mathbf{x}$ whose indexes are respectively in policy π and ν. Then the nonbasic variables $\mathbf{x}_\nu^\pi = \mathbf{0}$ and the basic variables $\mathbf{x}_\pi^\pi$ are the unique solution to $A_\pi \mathbf{x}_\pi = \mathbf{e}$. The corresponding basic solution of the dual is the vector $\mathbf{y}^\pi$ that is the unique solution to $A_\pi^T \mathbf{y} = \mathbf{c}_\pi^T$. The basic solution $\mathbf{y}^\pi$ of the dual is feasible if also $A_\nu^T \mathbf{y}^\pi \leq \mathbf{c}_\nu^T$ or $\mathbf{s}_\nu \geq \mathbf{0}$. The basic solution pair $\mathbf{x}^\pi$ and $\mathbf{y}^\pi$ of the DMDP primal and dual are optimal if and only if both are feasible. If policy π produces optimal $\mathbf{x}^\pi$ and $\mathbf{y}^\pi$, then π is an optimal policy π^* and $\mathbf{y}^{\pi^*}$ is exactly $\mathbf{v}^*$. Note that the constraints $A_{\pi^*}^T \mathbf{y}^{\pi^*} = \mathbf{c}_{\pi^*}^T$ and $A^T \mathbf{y}^{\pi^*} \leq \mathbf{c}^T$ describe the same condition for $\mathbf{v}^*$ in (2) for each policy π or for each state-action j.

2 The Markov decision problem methods and their complexities

There are several major events in developing methods for solving DMDPs. Bellman (1957) [1] developed a successive approximate method, called value-iteration, which computes the optimal total cost function assuming first a one stage finite horizon, then a two-stage finite horizon, and so on. The total cost functions so computed are guaranteed to converge in the limit to the optimal total cost function. It should be noted that, even prior to Bellman, Shapley (1953) [23] used value-iteration to solve DMDPs in the context of zero-sum two-person stochastic games.

The other best known method is due to Howard (1960) [12] and is known as policy-iteration, which generates an optimal policy in a finite number of iterations. Policy-iteration alternates between a value determination phase, in which the current policy is evaluated, and a policy improvement phase, in which an attempt is made to improve the current policy. In the policy improvement phase, the policy-iteration method updates possibly improved actions for every state in one iteration. If the the current policy is improved for at most one state in one iteration, then it is called simple policy-iteration. We will come back to the policy-iteration and simple policy-iteration methods later in terms of the LP formulation.

Since it was discovered in 1960 that the DMDP has an LP formulation, the Simplex method of Dantzig (1947) [5] can be used to solving DMDPs. It turns out that the Simplex method, when applied to solving DMDPs, is the simple policy-iteration method. Other general LP methods, such as the Ellipsoid method and interior-point algorithms are also capable to solve DMDPs.

As the notion of computational complexity emerged, there were tremendous efforts in analyzing the complexity of the MDP and its solution methods. On the positive side, since

it (with or without discount) can be formulated as an linear program, the MDP can be solved in polynomial time by either the Ellipsoid method (e.g., Khachiyan (1979) [14]) or the interior-point algorithm (e.g., Karmarkar (1984) [13]). Here, polynomial time means that the number of arithmetic operations needed to compute an optimal policy is bounded by a polynomial in the numbers of states, actions, and the bit-size of the input data, which are assumed to be rational numbers. Papadimitriou and Tsitsiklis [20] then showed in 1987 that an MDP with deterministic transitions (i.e., each entry of state transition probability matrices is either 0's or 1's) can be solved in *strongly* polynomial-time (i.e., the number of arithmetic operations is bounded by a polynomial in the numbers of states and actions only) as a Minimum-Mean-Cost-Cycle problem. Erickson [7] in 1988 showed that successive approximations suffice to produce: (1) an optimal stationary halting policy, or (2) show that no such policy exists in strongly polynomial time algorithm, based on the work of Eaves and Veinott [6] and Rothblum [22].

There were also great research interests and progresses in the value-iteration and policy-iteration methods for solving the DMDP. Bertsekas [2] in 1987 showed that the value-iteration method converges to the optimal policy in a finite number of iterations. Tseng [24] in 1990 showed that the value-iteration method generates an optimal policy in polynomial-time for the DMDP when the discount rate is fixed. Puterman [21] 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 the DMDP with a fixed discount rate. This fact was actually known to Veinott (and perhaps others) three decades earlier and used in dynamic programming courses he taught for a number of years well before 1994. Mansour and Singh [18] in 1994 also gave an upper bound on the number of iterations, $\frac{k^m}{m}$, for the policy-iteration method in solving the DMDP when each state has k actions. (Note that the total number of possible policies is k^m, so that the result is not much better than that of complete enumeration.) In 2005, Ye [28] developed a *strongly* polynomial-time combinatorial interior-point algorithm (CIPA) for the DMDP with a fixed discount rate, that is, the number of arithmetic operations is bounded by a polynomial in only the numbers of states and actions.

In terms of the worst-case complexity bound on the number of arithmetic operations, the current best results (within a constant factor) are summarized in the following table, when there are exact k actions in each of the m states; see Littman et al. [16], Mansour and Singh [18], Ye [28], and references therein.

Value-Iteration	Policy-Iteration	LP-Algorithms	CIPA
$\frac{m^2kL(P,\mathbf{c},\gamma)\log(1/(1-\gamma))}{1-\gamma}$	$\min\left\{\frac{m^3k\cdot k^m}{m}, \frac{m^3kL(P,\mathbf{c},\gamma)\log(1/(1-\gamma))}{1-\gamma}\right\}$	$m^3k^2L(P,\mathbf{c},\gamma)$	$m^4k^4\log\frac{m}{1-\gamma}$

Here, $L(P,\mathbf{c},\gamma)$ is the total bit-size of the DMDP input data in the linear programming form, given that $(P,\mathbf{c},\gamma)$ have only rational entries. As 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 needs to be a polynomial only in m and k. The proof of polynomial-time for the value and policy-iteration methods is essentially due to the argument that, when the gap between the objective value of the current policy (or BFS) and the optimal one is small than $2^{-L(P,\mathbf{c},\gamma)}$, the current policy must be optimal; e.g., see [11]. However, the proof of a *strongly* polynomial-time algorithm cannot rely on this argument, since $(P, \mathbf{c}, \gamma)$ may have irrational entries so that the bit-size of the data can be ∞.

In practice, the policy-iteration method, including the simple policy-iteration or Simplex method, has been remarkably successful and shown to be a most effective and widely used. The number of iterations is typically bounded by $O(mk)$. It turns out that the policy-iteration method is actually the Simplex method with block pivots at each iteration; and the Simplex method also remains one of the very few extremely effective methods for solving general LPs; see Bixby [3]. In the past 50 years, many efforts have been made to resolve the worst-case complexity issue of the policy-iteration method or the Simplex method, and to answer the question: are the policy-iteration and the Simplex methods *strongly* polynomial-time algorithms?

Unfortunately, so far most results have been negative. Klee and Minty [15] showed in 1972 that the classic 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 [19] showed that one simple policy-iteration method, where 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 DMDP problem regardless of discount rates (i.e., even when $\gamma < 1$ is fixed). Most recently, Fearnley (2010) [10] showed that the policy-iteration method needs an exponential number of iterations for an undiscounted but finite-horizon MDP. Thus, it seems impossible for the policy-iteration method to be a *strongly* polynomial-time algorithm for solving general MDPs.

What about DMDP with a fixed discount rate? Is there a pivoting rule to make the simplex and policy-iteration methods *strongly* polynomial for the DMDP?

In this 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 the DMDP with a fixed discount rate $0 \leq \gamma < 1$. The number of its iterations is bounded by

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

and each iteration uses at most $O(m^2k)$ arithmetic operations. The result seems surprising, given the earlier negative results mentioned above.

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, we prove that it is also a *strongly* polynomial-time algorithm with the same iteration complexity bound. Therefore, the worst-case operation complexity, $O(m^4k^2 \log m)$, of the Simplex method is actually *superior* to that, $O(m^4k^4 \log m)$, of the combinatorial interior-point algorithm [28] for solving DMDPs when the discount rate is a fixed constant.

If the number of actions varies among the states, our worst-case iteration bound would be $\frac{m(n-m)}{1-\gamma} \cdot \log\left(\frac{m^2}{1-\gamma}\right)$, and each iteration uses at most $O(mn)$ arithmetic operations, where n is again the total number of actions. One can see that the worst-case iteration complexity bound is linear in the total number of actions, as it is observed in practice.

We remark that, if the discount rate is an input, it remains open whether or not the policy-iteration method is polynomial for the MDP, or whether or not there exists a *strongly* polynomial-time algorithm for MDP or LP in general.

3 DMDP Properties and the Simplex and policy-iteration methods

We first describe a few general LP and DMDP theorems and the classic Simplex and policy-iteration methods. We will use the LP formulation (3) and (4) for DMDP and the terminology presented in the Introduction section. Recall that, for DMDP,

$$\mathbf{b} = \mathbf{e} \in \mathbb{R}^m, \ A = E - \gamma P \in \mathbb{R}^{m \times n},$$

and $\mathbf{c}$, P and E are defined in (5), (7) and (8), respectively.

3.1 DMDP Properties

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

$$\begin{aligned} A\mathbf{x} &= \mathbf{b}, \\ A^T\mathbf{y} + \mathbf{s} &= \mathbf{c}, \\ s_j x_j &= 0, \ \forall j = 1, \cdots, n, \\ \mathbf{x} \geq \mathbf{0}, &\quad \mathbf{s} \geq \mathbf{0} \end{aligned}$$

where the third condition is often referred as the complementarity condition.

Let π be the index set of state-actions corresponding to a policy. Then, as we briefly mentioned earlier, $\mathbf{x}^\pi$ is a BFS of the DMDP primal and basis A_π has the form $A_\pi = (I - \gamma P_\pi)$, and P_π is a column stochastic matrix , that is, $P_\pi \geq 0$ and $\mathbf{e}^T P_\pi = \mathbf{e}$. In fact, the converse is also true, that is, the index set π of basic variables of every BSF of the DMDP primal is a policy for the original DMDP. In other words, π must have exactly one variable or action index in $\mathcal{A}_i$, for each state i. Thus, we have the following lemma.

Lemma 1 *The DMDP primal linear programming formulation has the following properties:*

1. *There is a one-to-one correspondence between a (stationary) policy of the original DMDP and a basic feasible solution of the DMDP primal.*

2. *Let* $\mathbf{x}^\pi$ *be a basic feasible solution of the DMDP primal. Then any basic variable, say* $\mathbf{x}_i^\pi$*, has its value*
$$1 \le \mathbf{x}_i^\pi \le \frac{m}{1-\gamma}.$$

3. *The feasible set of the DMDP primal is bounded. More precisely,*
$$\mathbf{e}^T\mathbf{x} = \frac{m}{1-\gamma},$$
for every feasible $\mathbf{x} \ge \mathbf{0}$.

Proof. Let π be the basis set of any basic feasible solution for the DMDP primal. Then, the first statement can be seen as follows. Consider the coefficients of the ith row of A. From the structure of (6), (7) and (8), we must have $a_{ij} \le 0$ for all $j \notin \mathcal{A}_i$. Thus, if no basic variable is chosen from $\mathcal{A}_i$ or $\pi \cap \mathcal{A}_i = \emptyset$, then
$$1 = \sum_{j\in\pi} a_{ij}x_j^\pi = \sum_{j\in\pi, j\notin\mathcal{A}_i} a_{ij}x_j^\pi \le 0,$$
which is a contradiction. Thus, each state must have a state-action in π. On the other hand, $|\pi| = m$. Therefore, π must contain exactly one action index in $\mathcal{A}_i$ from each state $i = 1, \cdots, m$, that is, π is a policy.

The last two statements of the lemma were given in [28] whose proofs were based on Dantzig [4, 5] and Veinott [26]. ∎

From the first statement of Lemma 1, in what follows we simply call the basis index set π of any BFS of the DMDP primal a policy. For the basis $A_\pi = (I - \gamma P_\pi)$ of any policy π, the BFS $\mathbf{x}^\pi$ and the dual can be computed as
$$\mathbf{x}_\pi^\pi = (A_\pi)^{-1}\mathbf{e} \ge \mathbf{e},\ \mathbf{x}_\nu^\pi = \mathbf{0},\ \mathbf{y}^\pi = (A_\pi^T)^{-1}\mathbf{c}_\pi,\ \mathbf{s}_\pi^\pi = \mathbf{0},\ \mathbf{s}_\nu^\pi = \mathbf{c}_\nu - A_\nu^T(A_\pi^T)^{-1}\mathbf{c}_\pi,$$
where again ν contains the rest action indexes not in π. Since $\mathbf{x}^\pi$ and $\mathbf{s}^\pi$ are already complementary, if $\mathbf{s}_\nu^\pi \ge \mathbf{0}$, then π would be an optimal policy.

We now present the following strict complementarity result for the DMDP.

Lemma 2 *Let both linear programs (3) and (4) be feasible. Then there is a unique partition* $\mathcal{P} \subseteq \{1, 2, \cdots, n\}$ *and* $\mathcal{O} \subseteq \{1, 2, \cdots, n\}$, $\mathcal{P} \cap \mathcal{O} = \emptyset$ *and* $\mathcal{P} \cup \mathcal{Q} = \{1, 2, \cdots, n\}$, *such that for all optimal solution pair* $(\mathbf{x}^*, \mathbf{s}^*)$,
$$x_j^* = 0,\ \forall j \in \mathcal{O}, \quad \textit{and} \quad s_j^* = 0,\ \forall j \in \mathcal{P},$$

and there is at least one optimal solution pair $(\mathbf{x}^*, \mathbf{s}^*)$ *that is strictly complementary,*

$$x_j^* > 0, \forall j \in \mathcal{P}, \quad \textit{and} \quad s_j^* > 0, \ \forall j \in \mathcal{O},$$

for the DMDP linear program. In particular, every optimal policy $\pi^* \subseteq \mathcal{P}$ *so that* $|\mathcal{P}| \geq m$ *and* $|\mathcal{O}| \leq n - m$.

PROOF. The strict complementarity result for general LP is well known, where we call $\mathcal{P}$ the optimal (super) *basic variable* set and $\mathcal{O}$ the optimal *non-basic variable* set. The cardinality result is from the fact that there is always an optimal basic feasible solution or optimal policy where the basic variables (optimal state-action frequencies) are all strictly positive from Lemma 1, so that their indexes must all belong to $\mathcal{P}$. ∎

The interpretation of Lemma 2 is as folows: since there may exist multiple optimal policies π^* for a DMDP, $\mathcal{P}$ contains those state-actions each of which appears in at least one optimal policy, and $\mathcal{O}$ contains the rest state-actions neither of which appears in any optimal policy. Let's call each state-action in $\mathcal{O}$ a non-optimal state-action or simply non-optimal action. Then, any DMDP should have no more than $n - m$ non-optimal actions.

Note that, although there may be multiple optimal policies for a DMDP, the optimal dual basic feasible solution $(\mathbf{y}^*, \mathbf{s}^*)$ is *unique and invariant* among the multiple optimal policies. Thus, if j is a non-optimal action, its optimal dual slack value, s_j^*, must be strictly greater than 0, and the converse is also true by the lemma.

3.2 The Simplex and policy-iteration Methods

Let π be a policy and ν contain the remaining indexes of the non-basic variables. Then we can rewrite (3) as

$$\begin{array}{rlll} \text{minimize} & \mathbf{c}_\pi^T \mathbf{x}_\pi + \mathbf{c}_\nu^T \mathbf{x}_\nu & & \\ \text{subject to} & A_\pi \mathbf{x}_\pi + A_\nu \mathbf{x}_\nu & = & \mathbf{e}, \\ & \mathbf{x} = (\mathbf{x}_\pi; \mathbf{x}_\nu) & \geq & \mathbf{0}, \end{array} \tag{9}$$

with its dual

$$\begin{array}{rlll} \text{maximize} & \mathbf{e}^T \mathbf{y} & & \\ \text{subject to} & A_\pi^T \mathbf{y} + \mathbf{s}_\pi & = & \mathbf{c}_\pi, \\ & A_\nu^T \mathbf{y} + \mathbf{s}_\nu & = & \mathbf{c}_\nu, \\ & \mathbf{s} = (\mathbf{s}_\pi; \mathbf{s}_\nu) & \geq & \mathbf{0}. \end{array} \tag{10}$$

The (primal) Simplex method rewrites (9) into an equivalent problem

$$\begin{array}{rlll} \text{minimize} & (\bar{\mathbf{c}}_\nu)^T \mathbf{x}_\nu + \mathbf{c}_\pi^T (A_\pi)^{-1} \mathbf{e} & & \\ \text{subject to} & A_\pi \mathbf{x}_\pi + A_\nu \mathbf{x}_\nu & = & \mathbf{e}, \\ & \mathbf{x} = (\mathbf{x}_\pi; \mathbf{x}_\nu) & \geq & \mathbf{0}; \end{array} \tag{11}$$

where $\bar{\mathbf{c}}$ is called the *reduced cost* vector:

$$\bar{\mathbf{c}}_\pi = \mathbf{0} \quad \text{and} \quad \bar{\mathbf{c}}_\nu = \mathbf{c}_\nu - A_\nu^T \mathbf{y}^\pi,$$

and

$$\mathbf{y}^\pi = (A_\pi^T)^{-1} \mathbf{c}_\pi.$$

Note that the fixed quantity $\mathbf{c}_\pi^T (A_\pi)^{-1} \mathbf{e} = \mathbf{c}^T \mathbf{x}^\pi$ in the objective function of (11) is the objective value of the current policy π for (9). In fact, problem (9) and its equivalent form (11) share exactly the same objective value for every feasible solution $\mathbf{x}$.

The Simplex method

If $\bar{\mathbf{c}} \geq \mathbf{0}$, the current policy is optimal. Otherwise, let $0 < \Delta = -\min(\bar{\mathbf{c}})$ with $j^+ = \arg\min(\mathbf{c})$, that is, $\bar{c}_{j^+} = -\Delta > 0$. Then we must have $j^+ \notin \pi$, since $\bar{c}_j = 0$ for all $j \in \pi$. Let $j^+ \in \mathcal{A}_i$, that is, let j^+ be a state-action controlled by state i. Then, the classic Simplex method (Dantzig 1947) takes x_{j^+} as the incoming basic variable to replace the old one x_{π_i}, and the method repeats with the new policy denoted by π^+ where $\pi_i \in \mathcal{A}_i$ is replaced by $j^+ \in \mathcal{A}_i$. The method will break a tie arbitrarily, and it updates exactly one state-action in one iteration, that is, it only updates the state with the most negative reduced cost. This is the classic Simplex, or the simple policy-iteration, method that uses the most-negative-reduced-cost updating or pivoting rule.

The policy-iteration method

The original policy-iteration method (Howard 1960 [12]) is to update every state that has a negative reduced cost. For each state i, let $\Delta_i = -\min_{j \in \mathcal{A}_i}(\bar{\mathbf{c}})$ with $j_i^+ = \arg\min_{j \in \mathcal{A}_i}(\bar{\mathbf{c}})$. Then for every state i such that $\Delta_i > 0$, let $j_i^+ \in \mathcal{A}_i$ replace $\pi_i \in \mathcal{A}_i$ already in the current policy π. The method repeats with the new policy denoted by π^+, where possibly multiple $\pi_i \in \mathcal{A}_i$ are replaced by $j_i^+ \in \mathcal{A}_i$. The method will also break a tie in each state arbitrarily.

Therefore, both methods would generate a sequence of polices denoted by $\pi^0, \pi^1, \ldots, \pi^t, \ldots$, starting from any initial policy π^0. We comment that the Simplex and policy-iteration methods with the greedy or the most-negative-reduced-cost updating rule are special versions of generic policy improvement. In what follows, the most-negative-reduced-cost pivoting rule is used as a default for the Simplex and policy-iteration methods, unless otherwise stated.

4 Proof of strong polynomiality

We first prove our strongly polynomial-time result for the Simplex method. For the improvement of new policy π^+ over any policy π of the Simplex method, we have

Lemma 3 *Let z^* be the optimal objective value of (9) . Then, in any iteration of the Simplex method from current policy π to new policy π^+*

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

Moreover,

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

Therefore, the Simplex method generates a sequence of polices $\pi^0, \pi^1, \ldots, \pi^t, \ldots$ such that

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

Proof. From problem (11), we see that the objective function value for any feasible $\mathbf{x}$ is

$$\mathbf{c}^T\mathbf{x} = \mathbf{c}^T\mathbf{x}^\pi + \bar{\mathbf{c}}^T\mathbf{x} \geq \mathbf{c}^T\mathbf{x}^\pi - \Delta \cdot \mathbf{e}^T\mathbf{x} = \mathbf{c}^T\mathbf{x}^\pi - \Delta \cdot \frac{m}{1-\gamma},$$

where the first inequality follows from $\bar{\mathbf{c}} \geq \Delta \cdot \mathbf{e}$, by the most-negative-reduced-cost pivoting rule adapted in the method, and the last equality is based on the third statement of Lemma 1. In particular, the optimal objective value is

$$z^* = \mathbf{c}^T\mathbf{x}^* \geq \mathbf{c}^T\mathbf{x}^\pi - \frac{m}{1-\gamma} \cdot \Delta,$$

which proves the first inequality of the lemma.

Since at the new policy π^+, the value of new basic variable $x_{j^+}^{\pi^+}$ is greater than or equal to 1, from the second statement of Lemma 1, the objective value of the new policy for problem (11) is decreased by at least Δ. Thus, for problem (9),

$$\mathbf{c}^T\mathbf{x}^\pi - \mathbf{c}^T\mathbf{x}^{\pi^+} = \Delta \cdot x_{j^+}^{\pi^+} \geq \Delta \geq \frac{1-\gamma}{m}\left(\mathbf{c}^T\mathbf{x}^\pi - z^*\right),$$

or

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

which proves the second inequality.

Replacing π by π^t and using the above inequality, for all $t = 0, 1, \ldots,$ we have

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

which leads to the third desired inequality by induction. ∎

We now present the following key technical lemma.

Lemma 4 *1. If a policy π is not optimal, then there is a state-action $j \in \pi \cap \mathcal{O}$ (i.e., a non-optimal state-action j in the current policy) such that*

$$s_j^* \geq \frac{1-\gamma}{m^2}\left(\mathbf{c}^T\mathbf{x}^\pi - z^*\right),$$

where $\mathcal{O}$, together with $\mathcal{P}$, is the strict complementarity partition stated in Lemma 2, and $\mathbf{s}^*$ is the optimal dual slack vector of (10).

2. For any sequence of polices $\pi^0, \pi^1, \ldots, \pi^t, \ldots$ generated by the Simplex method where π^0 is not optimal, let $j^0 \in \pi^0 \cap \mathcal{O}$ be the state-action index identified above in the initial policy π^0. Then, if $j^0 \in \pi^t$, we must have

$$x_{j^0}^{\pi^t} \leq \frac{m^2}{1-\gamma} \cdot \frac{\mathbf{c}^T\mathbf{x}^{\pi^t} - z^*}{\mathbf{c}^T\mathbf{x}^{\pi^0} - z^*}, \ \forall t \geq 1.$$

Proof. Since all non-basic variable of $\mathbf{x}^\pi$ have zero values,

$$\mathbf{c}^T\mathbf{x}^\pi - z^* = \mathbf{c}^T\mathbf{x}^\pi - \mathbf{e}^T\mathbf{y}^* = (\mathbf{s}^*)^T\mathbf{x}^\pi = \sum_{j\in\pi} s_j^* x_j^\pi.$$

Since the number of non-negative terms in the sum is m, there must be a state-action $j \in \pi$ such that

$$s_j^* x_j^\pi \geq \frac{1}{m}\left(\mathbf{c}^T\mathbf{x}^\pi - z^*\right).$$

Then, from Lemma 1, $x_j^\pi \leq \frac{m}{1-\gamma}$, so that

$$s_j^* \geq \frac{1-\gamma}{m^2}\left(\mathbf{c}^T\mathbf{x}^\pi - z^*\right) > 0,$$

which also implies $j \in \mathcal{O}$ from Lemma 2.

Now, suppose the initial policy π^0 is not optimal and let $j^0 \in \pi^0 \cap \mathcal{O}$ be the index identified at policy π^0 such that the above inequality holds, that is,

$$s_{j^0}^* \geq \frac{1-\gamma}{m^2}\left(\mathbf{c}^T\mathbf{x}^{\pi^0} - z^*\right).$$

Then, for any policy π^t generated by the Simplex method, if $j^0 \in \pi^t$, we must have

$$\mathbf{c}^T\mathbf{x}^{\pi^t} - z^* = (\mathbf{s}^*)^T\mathbf{x}^{\pi^t} \geq s_{j^0}^* x_{j^0}^{\pi^t},$$

so that

$$x_{j^0}^{\pi^t} \leq \frac{\mathbf{c}^T\mathbf{x}^{\pi^t} - z^*}{s_{j^0}^*} \leq \frac{m^2}{1-\gamma} \cdot \frac{\mathbf{c}^T\mathbf{x}^{\pi^t} - z^*}{\mathbf{c}^T\mathbf{x}^{\pi^0} - z^*}.$$

∎

These lemmas lead to our key result:

Theorem 1 *Let π^0 be any given non-optimal policy. Then there is a state-action $j^0 \in \pi^0 \cap \mathcal{O}$, i.e., a non-optimal action j^0 in policy π^0, that would never appear in any of the policies generated by the Simplex method after $T := \lceil \frac{m}{1-\gamma} \cdot \log\left(\frac{m^2}{1-\gamma}\right) \rceil$ iterations starting from π^0.*

Proof. From Lemma 3, after t iterations of the Simplex method, we have

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

Therefore, after $t \geq T+1$ iterations from the initial policy π^0, $j^0 \in \pi^t$ implies, by Lemma 4,

$$x_{j^0}^{\pi^t} \leq \frac{m^2}{1-\gamma} \cdot \frac{\mathbf{c}^T\mathbf{x}^{\pi^t} - z^*}{\mathbf{c}^T\mathbf{x}^{\pi^0} - z^*} \leq \frac{m^2}{1-\gamma} \cdot \left(1 - \frac{1-\gamma}{m}\right)^t < 1.$$

The last inequality above comes from the fact $\log(1-x) \leq -x$ for all $x < 1$ so that

$$\log \frac{m^2}{1-\gamma} + t \cdot \log\left(1 - \frac{1-\gamma}{m}\right) \leq \log \frac{m^2}{1-\gamma} + t \cdot \left(-\frac{1-\gamma}{m}\right) < 0$$

if $t \geq 1 + T \geq 1 + \frac{m}{1-\gamma} \cdot \log\left(\frac{m^2}{1-\gamma}\right)$. But $x_{j^0}^{\pi^t} < 1$ is a contradiction to Lemma 1, which states that every basic variable value must be greater or equal to 1. Thus, $j^0 \notin \pi^t$ for all $t \geq T+1$. ∎

The event described in Theorem 1 can be viewed as a *crossover event* of Vavasis and Ye [25, 28]: a state-action, although we don't know which one it is, was in the initial policy but it will *never* stay in or return to the policies after a certain number of iterations, during the iterative process of the Simplex or simple policy-iteration method.

We now repeat the same proof for policy π^{T+1}, if it is not optimal yet, in the policy sequence generated by the Simplex method. Since policy π^{T+1} is not optimal, there must be a non-optimal state-action, $j^1 \in \pi^{T+1} \cap \mathcal{O}$ and $j^1 \neq j^0$ (because of Theorem 1), that would never stay in or return to the policies generated by the Simplex method after $2T$ iterations starting from π^0. Again, we can repeat this process for policy π^{2T+1} if it is not optimal yet, and so on.

In each of these cycles of T Simplex iterations, at least one *new non-optimal state-action* is eliminated from appearance in any of the future policy cycles generated by the Simplex method. However, we have at most $|\mathcal{O}|$ many such non-optimal state-actions to eliminate, where $|\mathcal{O}| \leq n - m$ from Lemma 2. Hence, the Simplex method can cycle at most $n - m$ times, and we reach our main conclusion:

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

The arithmetic operations count is well known for the Simplex method: it uses $O(m^2)$ arithmetic operations to update the inverse of the basis $(A_{\pi^t})^{-1}$ of the current policy π^t and the dual basic solution $\mathbf{y}^{\pi^t}$, as well as $O(mn)$ arithmetic operations to calculate the reduced cost, and then chooses the incoming basic variable.

We now turn our attention to the policy-iteration method, and we have the following corollary:

Corollary 1 *The original policy-iteration method of Howard for solving the discounted Markov decision problem with a fixed discount rate is a strongly polynomial-time algorithm. Starting from any policy, it terminates in at most* $\frac{m(n-m)}{1-\gamma} \cdot \log\left(\frac{m^2}{1-\gamma}\right)$ *iterations.*

Proof. First, Lemmas 1 and 2 hold since they are independent of which method is being used. Secondly, Lemma 3 still holds for the policy-iteration method, since at any policy π the incoming basic variable $j^+ = \arg\min(\mathbf{c})$ (that is, $\bar{\mathbf{c}}_{j^+} = -\Delta = -\min(\bar{\mathbf{c}})$) for the Simplex method is always one of the incoming basic variables for the policy-iteration method. Thirdly, the facts established by Lemma 4 are also independent of how the policy sequence is generated as long as the state-action with the most-negative-reduced-cost is included in the next policy, so that they hold for the policy-iteration method as well. Thus, we can conclude that there is a state-action $j^0 \in \pi^0 \cap \mathcal{O}$, i.e., a non-optimal state-action j^0 in the initial non-optimal policy π^0, that would *never* stay in or return to the policies generated by the policy-iteration method after T iterations. Thus, Theorem 1 also holds for the policy-iteration method, which proves the corollary. ∎

Note that, for the policy-iteration method, each iteration could use up to $O(m^2 n)$ arithmetic operations.

5 Extensions and Remarks

Our result can be extended to other undiscounted MDPs where every basic feasible matrix of (9) exhibits the Leontief substitution form:

$$A_\pi = I - P,$$

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 [27]. 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, each basic variable value is still between 1 and $\frac{m}{1-\gamma}$, so that Lemma 1 is true with an inequality (actually stronger for our proof):

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

for every feasible solution $\mathbf{x}$. Consequently, Lemmas 2, 3, and 4 all hold, which leads to the following corollary.

Corollary 2 *Let every feasible basis of an MDP have the form $I - P$ where $P \geq \mathbf{0}$, with a spectral radius less than or equal to a fixed $\gamma < 1$. Then, the Simplex and policy-iteration methods are strongly polynomial-time algorithms. Starting from any policy, each of them terminates in at most $\frac{m(n-m)}{1-\gamma} \cdot \log\left(\frac{m^2}{1-\gamma}\right)$ iterations.*

One observation from our worst-case analyses is that there is no iteration-count difference between the Simplex method and the policy-iteration method that makes block pivots in each iteration, as long as the most-negative-reduced-cost pivoting rule is adapted. However, each iteration of the Simplex method is more efficient than that the policy-iteration method.

Finally, we remark that the pivoting rule seems to make the difference. As we mentioned earlier, for the DMDP with a fixed discount rate, the simplex or simple policy-iteration method with the smallest-index pivoting rule (a rule popularly used against cycling in the presence of degeneracy) was shown to be exponential. This is in contrast to the method that uses the most-negative-reduced-cost pivoting rule, which is proven to be strongly polynomial in this paper. On the other hand, the most-negative-reduced-cost pivoting rule is exponential for solving some other LP problems. Thus, searching for suitable pivoting rules for solving different LP problems is essential, and one cannot rule out the Simplex method simply because the behavior of one pivoting rule on one problem is shown to be exponential.

Further possible research directions may answer the questions: can the Simplex method or the policy-iteration method be strongly polynomial for solving the MDP regardless of discount rates? Or, is there any strongly polynomial-time algorithm for solving the MDP regardless of discount rates?

Acknowledgments. I thank Pete Veinott and four anonymous Referees for many insightful discussions and suggestions on this subject, which have greatly improved the presentation of the paper.

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 (1994) 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] B. E. Eaves and A. F. Veinott, Maximum-Stopping-Value Policies in Finite Markov Population Decision Chains, manuscript, Stanford University, 2007.

[7] R. E. Erickson, Optimality of Stationary Halting Policies and Finite Termination of Successive Approximations, *Mathematics of Operations Research* **13**:1 (1988), 90-98.

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

[9] 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).

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

[11] M. Grötschel, L. Lovász and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer, Berlin, 1988.

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

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

[14] L. G. Khachiyan, A polynomial algorithm in linear programming, *Dokl. Akad. Nauk SSSR* **244** (1979), 1093–1086; Translated in *Soviet Math. Dokl.* **20** 191–194.

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

[16] 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.

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

[18] 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.

[19] 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.

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

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

[22] U. Rothblum, *Multiplicative Markov Decision Chains*, Doctoral Dissertation, Department of Operations Research, Stanford University, Stanford, 1974.

[23] L. S. Shapley, Stochastic Games, *Proc Natl Acad Sci U S A* **39**:10 (1953), 10951100.

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

[25] 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.

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

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

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