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

Yinyu Ye
Department of Management Science and Engineering, Stanford University, Stanford, CA
email: yinyu-ye@tanford.edu

Initial submission date: May 15, 2010; first revision Nov. 30, 2010; second revision August 16, 2011.

We prove that the classic policy-iteration method (Howard 1960) and the original simplex method with the most-negative-reduced-cost pivoting rule (Dantzig 1947) are *strongly* polynomial-time algorithms for solving the Markov decision problem (MDP) with a fixed discount rate. Furthermore, the computational complexity of the policy-iteration and simplex methods is *superior* to that of the only known strongly polynomial-time interior-point algorithm (Ye 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 (Klee and Minty 1972), the simplex method with the smallest-index pivoting rule was shown to be exponential for solving an MDP regardless of discount rates (Melekopoglou and Condon 1994), and the policy-iteration method was recently shown to be exponential for solving undiscounted MDPs under the average cost criterion. We also extend the result to solving MDPs with transient substochastic transition matrices whose spectral radii uniformly minorize one.

Key words: Simplex Method, Policy-Iteration Method, Markov Decision Problem; Linear Programming; Dynamic Programming, Strongly Polynomial Time

MSC2000 Subject Classification: Primary: 90C40, 90C05; Secondary: 68Q25, 90C39

OR/MS subject classification: Primary: Linear Programming, Algorithm; Secondary: Dynamic Programming, Markov

---

1. Introduction Markov decision problems (MDPs) provide a mathematical framework for modeling decision-making in situations where outcomes are partly random 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 is being used in a variety of areas, including management, economics, bioinformatics, electronic commerce, social networking, and supply chains.

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

Let m denote the total number of states. The probability that the process enters state i' after a transition is determined by the current state i and the chosen action j, but is conditionally independent of all previous states and actions; in other words, the state transition of an MDP possesses the Markov property. Specifically, it is given by a transition probability distribution $p^j(i, i') \geq 0$, $i' = 1, \cdots, m$, and $\sum_{i'=1}^{m} p^j(i, i') = 1$, $\forall i = 1, \cdots, m$ (For transient substochastic systems, the sum is allowed to be less than or equal to one).

The key goal of MDPs is to identify an optimal (stationary) policy for the decision maker, which can be represented as 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 $i = 1, \cdots, m$. A (stationary) policy π is optimal if it minimizes some cumulative function of the random costs, typically the (expected) discounted sum over the infinite horizon

$$\sum_{t=0}^{\infty} \gamma^t C(\pi, t),$$

where $C(\pi, t)$ is the (expected) cost at time t when the decision maker follows policy π, and γ is the discount rate assumed to be strictly less than one in this paper.

This MDP problem is called the infinite-horizon discounted Markov decision problem (DMDP), which serves as a core model for MDPs. Presumably, the decision maker is allowed to take different policies at different time steps to minimize the above cost. However, it is known that there exists an optimal *stationary* policy (or policy for short in the rest of this paper) for the DMDP that it is *independent* of time t so that it can be written as a set function of state $i = 1, \cdots, m$ only as described above.

Let k_i be the number of actions available in state $i \in \{1, \cdots, m\}$ 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,

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

Clearly, $|\mathcal{A}_i| = k_i$. The total number of actions is $n = \sum_{i=1}^{m} k_i$ and these, without loss of generality, can be ordered such that if $j \in \mathcal{A}_i$, then action j is available in state i.

Suppose we know the state transition probabilities $p^j(i,i')$ and the immediate cost coefficients $c^j(i,i')$ ($j = 1, \cdots, n$ and $i, i' = 1, \cdots, m$) and we wish to find a policy π^* that minimizes the expected discounted cost. Then the policy π would be associated with another array indexed by state, value vector $\mathbf{v}^* \in \mathbf{R}^m$, which contains cost-to-go values for all states. Furthermore, an optimal pair $(\pi^*, \mathbf{v}^*)$ is then a fixed point of the following operator,

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

where again $c^j(i,i')$ represents the immediate cost incurred by an individual in state i who takes action j and enters state i'. Let $P_\pi \in \mathbf{R}^{m \times m}$ 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 (1) can be represented in matrix form as follows

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

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

D'Epenoux [8] and de Ghellinck [7] (also see Manne [20], Kallenberg [14] and Altman [1]) showed that 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 $A \in \mathbf{R}^{m \times n}$ is a given real matrix with rank m, $\mathbf{c} \in \mathbf{R}^n$ and $\mathbf{b} \in \mathbf{R}^m$ are given real vectors, $\mathbf{0}$ denotes the vector of all 0's, and $\mathbf{x} \in \mathbf{R}^n$ and ($\mathbf{y} \in \mathbf{R}^m, \mathbf{s} \in \mathbf{R}^n$) 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.

The DMDP can be formulated as LP problems (3) and (4) with the following assignments of $(A, \mathbf{b}, \mathbf{c})$: First, every entry of the column vector $\mathbf{b} \in \mathbf{R}^m$ is one. Secondly, the jth entry of vector $\mathbf{c} \in \mathbf{R}^n$ is the (expected) immediate cost of action j. In particular, if $j \in \mathcal{A}_i$, then action j is available in state i and

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

Thirdly, the LP constraint matrix has the form

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

Here, the jth column of P is the transition probability distribution when the jth action is chosen. Specifically, for each action $j \in \mathcal{A}_i$

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

and

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

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 initially, there is one individual in each state. For $j \in \mathcal{A}_i$, the primal variable x_j is the action frequency of action j, or the expected present value of the number of times in which an individual is in state i and takes action j. Thus, solving the DMDP primal entails choosing action frequencies that minimize the expected present cost, $\mathbf{c}^T\mathbf{x}$, 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 $\mathbf{y} \in \mathbf{R}^m$ 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 $\mathbf{s} \in \mathbf{R}^n$ of slack variables, one for each 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 solution pair $(\mathbf{y}^*, \mathbf{s}^*)$ 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 indices 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 $A_\pi \in \mathbf{R}^{m\times m}$ (resp., P_π, E_π) be the columns of A (resp., P, E) with indices 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 remaining indices not in π.

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

2. Literature and Complexity There are several major events in developing methods for solving DMDPs. Bellman (1957) [2] developed a successive approximations 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) [26] used value-iteration to solve DMDPs in the context of zero-sum two-person stochastic games.

The other best known method to solve DMDPs is due to Howard (1960) [13] 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 original and classic policy-iteration method updates the actions possibly in 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 by D'Epenoux [8] and de Ghellinck [7] that the DMDP has an LP formulation, the simplex method of Dantzig (1947) [6] 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 of solving DMDPs.

As the notion of computational complexity emerged, there were tremendous efforts in analyzing the complexity of solving MDPs and its solution methods. On the positive side, since it (with or without discounting [1]) can be formulated as an linear program, the MDP can be solved in polynomial time by either the Ellipsoid method (e.g., Khachiyan (1979) [16]) or the interior-point algorithm (e.g., Karmarkar (1984) [15]). 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 [23] then showed in 1987 that an MDP with deterministic transitions (i.e., each entry of state transition probability matrices

[1]When there is no discounting, the optimality criterion or objective function needs to be changed where a popular one is to minimize the average cost $\lim_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N-1} C(\pi, t)$.

is either 0's or 1's) and the average cost optimality criterion 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 the well-known Minimum-Mean-Cost-Cycle problem. Erickson [11] in 1988 showed that finitely many iterations of 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 [10] and Rothblum [25].

Bertsekas [3] in 1987 also showed that the value-iteration method converges to the optimal policy in a finite number of iterations. Tseng [27] 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 [24] 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.[2] Mansour and Singh [21] 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 [32] developed a *strongly* polynomial-time combinatorial interior-point algorithm (CIPA) for solving DMDPs 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.

The worst-case running-times (within a constant factor), in terms of the number of arithmetic operations, of these methods are summarized in the following table, when there are exact k actions in each of the m states; see Littman et al. [18], Mansour and Singh [21], Ye [32], and references therein.

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

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 $0 \le \gamma < 1$ 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 complexity 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 [19]. 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 has been remarkably successful and shown to be most effective and widely used; where the number of iterations is typically bounded by $O(mk)$ [18]. It turns out that the policy-iteration method is actually the simplex method with block pivots in each iteration; and the simplex method also remains one of the very few extremely effective methods for solving general LPs; see Bixby [4]. In the past 50 years, many efforts have been made to resolve the worst-case complexity issues of the policy-iteration and simplex methods for solving MDPs and general LPs, 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 [17] showed in 1972 that the original simplex method with Dantzig's 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 [22] showed that the simplex or simple policy-iteration method with the smallest-index pivoting rule (that is, only the action with the smallest index is pivoted or updated among the actions with negative reduced costs) needs an exponential number of iterations to compute an optimal policy for a specific DMDP problem regardless of discount rate (i.e., even when $\gamma < 1$ is fixed). [3] Is there a pivoting rule to make the simplex and policy-iteration methods *strongly* polynomial for solving DMDPs?

In this paper, we prove that the original, and now classic, simplex method of Dantzig (1947), or the

[2]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 before 1994 [31].

[3]Most recently, Fearnley (2010) [12] showed that the policy-iteration method also needs an exponential number of iterations for an undiscounted but finite-horizon MDP under the average cost optimality criterion.

simple policy-iteration method with the most-negative-reduced-cost pivoting rule, is indeed a *strongly* polynomial-time algorithm for solving DMDPs with a fixed discount rate $0 \le \gamma < 1$. The number of its iterations is actually 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 original policy-iteration method of Howard (1960) updates the actions possibly in every state in one iteration, the action updated in one iteration of the simplex method will be included in the actions updated in one iteration of the policy-iteration method. Consequently, we prove that it is also a *strongly* polynomial-time algorithm with the same iteration complexity bound. Note that the worst-case arithmetic 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 [32] for solving DMDPs when the discount rate is a fixed constant.

If the number of actions varies among the states, our iteration complexity bound of the simplex method 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 observed in practice [18], when the number of states is fixed.

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

3. Preliminaries and Formulations

We first describe a few general LP and DMDP theorems and the 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,

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

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:

$$\begin{array}{rcl} A\mathbf{x} & = & \mathbf{b}, \\ A^T\mathbf{y} + \mathbf{s} & = & \mathbf{c}, \\ s_j x_j & = & 0, \; \forall j = 1, \cdots, n, \\ \mathbf{x} \ge \mathbf{0}, & & \mathbf{s} \ge \mathbf{0} \end{array}$$

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

Let π be the index set of 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, $P_\pi \ge 0$ 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 3.1 *The DMDP primal linear programming formulation has the following properties:*

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

(*ii*) *Let* $\mathbf{x}^\pi$ *be a basic feasible solution of the DMDP primal. Then any basic variable, say* $\mathbf{x}^\pi_i$, $i = 1, \cdots, m$, *has its value*

$$1 \le \mathbf{x}^\pi_i \le \frac{m}{1-\gamma}.$$

(*iii*) *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 $a_{ij} \leq 0$ 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

$$1 = \sum_{j \in \pi} a_{ij} x_j^\pi = \sum_{j \in \pi, j \notin \mathcal{A}_i} a_{ij} x_j^\pi \leq 0,$$

which is a contradiction. Thus, each state must have an action in π. On the other hand, since $\mathbf{x}^\pi$ is a BFS, $|\pi| \leq m$, so, π must contain exactly one action index in $\mathcal{A}_i$ for each state $i = 1, \cdots, m$, that is, π is a policy.

The last two statements of the lemma were given in [32] whose proofs were based on Dantzig [5, 6] and Veinott [29]. □

From the first statement of Lemma 3.1, 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

$$\mathbf{x}_\pi^\pi = (A_\pi)^{-1}\mathbf{e} \geq \mathbf{e},\ \mathbf{x}_\nu^\pi = \mathbf{0},\ \mathbf{y}^\pi = (A_\pi^T)^{-1}\mathbf{c}_\pi,\ \mathbf{s}_\pi^\pi = \mathbf{0},\ \mathbf{s}_\nu^\pi = \mathbf{c}_\nu - A_\nu^T (A_\pi^T)^{-1}\mathbf{c}_\pi,$$

where ν contains the remaining action indices not in π. Since $\mathbf{x}^\pi$ and $\mathbf{s}^\pi$ are already complementary, if $\mathbf{s}_\nu^\pi \geq \mathbf{0}$, then π would be an optimal policy.

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

Lemma 3.2 *If both linear programs (3) and (4) are feasible; then there is a unique partition $\mathcal{O} \subseteq \{1, 2, \cdots, n\}$ and $\mathcal{N} \subseteq \{1, 2, \cdots, n\}$, $\mathcal{O} \cap \mathcal{N} = \emptyset$ and $\mathcal{O} \cup \mathcal{N} = \{1, 2, \cdots, n\}$, such that for each optimal solution pair $(\mathbf{x}^*, \mathbf{s}^*)$,*

$$x_j^* = 0,\ \forall j \in \mathcal{N}, \quad \text{and} \quad s_j^* = 0,\ \forall j \in \mathcal{O},$$

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

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

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

Proof. The strict complementarity result for general LP is well known, where we call $\mathcal{O}$ the optimal (super) *basic variable* set and $\mathcal{N}$ 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 action frequencies) are all strictly positive from Lemma 3.1, so that their indices must all belong to $\mathcal{O}$. □

The interpretation of Lemma 3.2 is as follows: since there may exist multiple optimal policies π^* for a DMDP, $\mathcal{O}$ contains those actions each of which appears in at least one optimal policy, and $\mathcal{N}$ contains the remaining actions none of which appears in any optimal policy. Let's call each action in $\mathcal{N}$ a *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 pair $(\mathbf{y}^*, \mathbf{s}^*)$ is *unique and invariant* among the multiple optimal policies. Thus, if j is a non-optimal action, then its optimal dual slack value, s_j^*, must be positive, and the converse is true by the lemma.

3.2 The Simplex and Policy-Iteration Methods Let π be a policy and ν contain the remaining indices of the non-basic variables. Then we can rewrite (3) as

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

with its dual

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

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

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

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 $\mathbf{x}$.

We now describe the simplex (or simple policy-iteration) method of Dantzig and the policy-iteration method of Howard in terms of the LP formulation.

The simplex method If $\bar{\mathbf{c}} \geq \mathbf{0}$, the current policy is optimal. Otherwise, let $0 < \Delta = -\min\{\bar{c}_j : j = 1, 2, \cdots, n\}$ with $j^+ = \arg\min\{\bar{c}_j : \ j = 1, 2, \cdots, n\}$, 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 an action available in state i. Then, the original 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 action in one iteration, that is, it only updates the state with the most negative reduced cost.

The policy-iteration method The original policy-iteration method (Howard 1960) updates the actions in every state that has a negative reduced cost. Specifically, for each state i, let $\Delta_i = -\min\{\bar{c}_j : j \in \mathcal{A}_i\}$ with $j_i^+ = \arg\min\{\bar{c}_j : \ j \in \mathcal{A}_i\}$. 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.

Thus, 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. Clearly, one can see that the policy-iteration method is actually the simplex method with block pivots or updates in each iteration; and the action, j^+, updated by the simplex method will be always included in the actions updated by the policy-iteration method in every iteration. In what follows, we analyze the complexity bounds of the simplex and policy-iteration methods with their action updating or pivoting rules as exactly described above.

4. Strong Polynomiality

We first prove the strongly polynomial-time result for the simplex method.

LEMMA 4.1 *Suppose z^* is the optimal objective value of (9), π is the current policy, π^+ is the improvement of π, and Δ is the quantity defined in an iteration of the simplex method. Then, in any iteration of the simplex method from a current policy π to a 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 3.1. 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 the new basic variable $x^{\pi^+}_{j^+}$ is greater than or equal to 1 from the second statement of Lemma 3.1, the objective value of the new policy for problem (11) is decreased by at least Δ. Thus, for problem (9),

$$\mathbf{c}^T\mathbf{x}^\pi - \mathbf{c}^T\mathbf{x}^{\pi^+} = \Delta \cdot x^{\pi^+}_{j^+} \geq \Delta \geq \frac{1-\gamma}{m}\left(\mathbf{c}^T\mathbf{x}^\pi - z^*\right), \tag{12}$$

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 4.2 *(i) If a policy π is not optimal, then there is an action $j \in \pi \cap \mathcal{N}$ (i.e., a non-optimal action j in π) such that*

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

where $\mathcal{N}$, together with $\mathcal{O}$, is the strict complementarity partition stated in Lemma 3.2, and $\mathbf{s}^*$ is the optimal dual slack vector of (10).

(ii) For any sequence of polices $\pi^0, \pi^1, \ldots, \pi^t, \ldots$ generated by the simplex method starting with a non-optimal policy π^0, let $j^0 \in \pi^0 \cap \mathcal{N}$ be the action index identified above in the initial policy π^0. Then, if $j^0 \in \pi^t$, we must have

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

Proof. Since all non-basic variables of $\mathbf{x}^\pi$ are zero,

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

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

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

Then, from Lemma 3.1, $x^\pi_j \leq \frac{m}{1-\gamma}$, so that

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

which also implies $j \in \mathcal{N}$ from Lemma 3.2.

Now, suppose the initial policy π^0 is not optimal and let $j^0 \in \pi^0 \cap \mathcal{N}$ be the index identified in 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

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

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 4.1 *Let π^0 be any given non-optimal policy. Then there is an action $j^0 \in \pi^0 \cap \mathcal{N}$, 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 $\mathcal{T} := \lceil \frac{m}{1-\gamma} \cdot \log\left(\frac{m^2}{1-\gamma}\right) \rceil$ iterations starting from π^0.*

PROOF. Let $j^0 \in \pi^0$ be a non-optimal action identified in Lemma 4.2. From Lemma 4.1, 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 $t \geq \mathcal{T}+1$ iterations from the initial policy π^0, $j^0 \in \pi^t$ implies, by Lemma 4.2,

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

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 $t \geq 1 + \mathcal{T} \geq 1 + \frac{m}{1-\gamma} \cdot \log\left(\frac{m^2}{1-\gamma}\right)$. But $x_{j^0}^{\pi^t} < 1$ contradicts the second statement of Lemma 3.1. Thus, $j^0 \notin \pi^t$ for all $t \geq \mathcal{T}+1$. □

The event described in Theorem 4.1 can be viewed as a *crossover event* of Vavasis and Ye [28, 32]: an 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 $\mathcal{T}$ iterations of the iterative process of the simplex method.

We now repeat the same proof for policy $\pi^{\mathcal{T}+1}$, if it is not optimal yet, in the policy sequence generated by the simplex method. Since policy $\pi^{\mathcal{T}+1}$ is not optimal, there must be a non-optimal action, $j^1 \in \pi^{\mathcal{T}+1} \cap \mathcal{N}$ and $j^1 \neq j^0$ (because of Theorem 4.1), that would never stay in or return to the policies generated by the simplex method after $2\mathcal{T}$ iterations starting from π^0. Again, we can repeat this process for policy $\pi^{2\mathcal{T}+1}$ if it is not optimal yet, and so on.

In each of these loops (each loop consists of $\mathcal{T}$ simplex method iterations), at least one new non-optimal action is *eliminated* from consideration in any of the future loops of the simplex method. However, we have at most $|\mathcal{N}|$ many such non-optimal actions to eliminate, where $|\mathcal{N}| \leq n - m$ from Lemma 3.2. Hence, the simplex method loops at most $n - m$ times, which gives our main result:

THEOREM 4.2 *The simplex, or simple policy-iteration, method of Dantzig for solving the discounted Markov decision problem with a fixed discount rate γ 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 4.1 *The policy-iteration method of Howard for solving the discounted Markov decision problem with a fixed discount rate γ 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 3.1 and 3.2 hold since they are independent of which method is being used. Secondly, Lemma 4.1 still holds for the policy-iteration method, since at any policy π the incoming basic variable $j^+ = \arg\min\{\bar{c}_j : j = 1, 2, \cdots, n\}$ (that is, $\bar{c}_{j^+} = -\Delta = -\min\{\bar{c}_j : j = 1, 2, \cdots, n\}$) for the simplex method is always one of the incoming basic variables for the policy-iteration method. Thus, inequality (12) in the proof of Lemma 4.1 can actually be strengthened to

$$\mathbf{c}^T\mathbf{x}^\pi - \mathbf{c}^T\mathbf{x}^{\pi^+} = \sum_{i:\Delta_i>0} \Delta_i x^{\pi^+}_{j_i^+} \geq \Delta \cdot x^{\pi^+}_{j^+} \geq \Delta \geq \frac{1-\gamma}{m}\left(\mathbf{c}^T\mathbf{x}^\pi - z^*\right),$$

where $\Delta_i = -\min\{\bar{c}_j : j \in \mathcal{A}_i\}$ and $j_i^+ = \arg\min\{\bar{c}_j : j \in \mathcal{A}_i\}$, which implies the final results of the lemma. Thirdly, the facts established by Lemma 4.2 are also independent of how the policy sequence is generated as long as the 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 an action $j^0 \in \pi^0 \cap \mathcal{N}$, i.e., a non-optimal 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 $\mathcal{T}$ iterations. Thus, Theorem 4.1 also holds for the policy-iteration method, which proves the corollary. □

Note that, for the policy-iteration method, each iteration could use up to $O(m^2n)$ arithmetic operations, depending on how many actions would be updated in one iteration.

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 transient substochastic matrices; see Veinott [30] and Eaves and Rothblum [9]. 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 3.1 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 3.2, 4.1, and 4.2 all hold, which leads to the following corollary.

Corollary 5.1 *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 complexity difference between the simplex method and the policy-iteration method. However, each iteration of the simplex method is more efficient than that the policy-iteration method. Thus, an open question would be whether or not the iteration complexity bound of the policy-iteration method can be reduced.

Finally, we remark that the pivoting rule seems to make the difference. As we mentioned earlier, for solving DMDPs with a fixed discount rate, the simplex 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 original most-negative-reduced-cost pivoting rule of Dantzig, which is proven to be strongly polynomial in the current paper. On the other hand, the most-negative-reduced-cost pivoting rule is exponential for solving some other general LP problems. Thus, searching for suitable pivoting rules for solving different LP problems seems 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: Are the classic simplex method and policy-iteration methods strongly polynomial-time algorithms for solving MDPs under other optimality criteria? 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. The author would like to thank Pete Veinott and five anonymous referees for many insightful discussions and suggestions on this subject, which have greatly improved the presentation of the paper. This research is supported in part by NSF Grant GOALI 0800151 and AFOSR Grant FA9550-09-1-0306.

References

[1] E. Altman. *Constrained Markov Decision Processes.* Chapman and Hall/CRC, 1999.

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

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

[4] R. E. Bixby. Progress in linear programming. *ORSA J. on Comput.*, 6(1):15–22, 1994.

[5] G. B. Dantzig. Optimal solutions of a dynamic leontief model with substitution. *Econometrica*, 23:295–302, 1955.

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

[7] G. de Ghellinck. Les problém s de décisions séquentielles. *Cahiers du Centre d'Etudes de Recherche Opérationnelle*, 2:161–179, 1960.

[8] F. D'Epenoux. A probabilistic production and inventory problem. *Management Science*, 10:98–108, 1963.

[9] B. E. Eaves and U. G. Rothblum. A theory for extending algorithms for parametric problems. *Mathematics of Operations Research*, 14(3):502–533, 1989.

[10] B. E. Eaves and A. F. Veinott. Maximum-stopping-value policies in finite markov population decision chains. Technical report, 2007. manuscript, Stanford University.

[11] R. E. Erickson. Optimality of stationary halting policies and finite termination of successive approximations. *Mathematics of Operations Research*, 13(1):90–98, 1988.

[12] J. Fearnley. Exponential lower bounds for policy iteration. Technical report, 2010. arXiv:1003.3418v1.

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

[14] L. C. M. Kallenberg. *Linear Programming and Finite Markovian Control Problems.* Mathematical Centre Tracts, 1983.

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

[16] L. G. Khachiyan. A polynomial algorithm in linear programming. *Dokl. Akad. Nauk SSSR*, 244:1086–1093, 1979.

[17] V. Klee and G. J. Minty. How good is the simplex method. Technical report, 1972. in O. Shisha, editor, *Inequalities III*, Academic Press, New York.

[18] M. L. Littman, T. L. Dean, and L. P. Kaelbling. On the complexity of solving markov decision problems. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI–95)*, pages 394–402, 1994.

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

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

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

[22] M. Melekopoglou and A. Condon. On the complexity of the policy improvement algorithm for markov decision processes. *INFORMS Journal on Computing*, 6(2):188–192, 1994.

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

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

[25] U. G. Rothblum. Multiplicative markov decision chains. Technical report, 1974. Doctoral Dissertation, Department of Operations Research, Stanford University, Stanford.

[26] L. S. Shapley. Stochastic games. *Proceedings of the National Acadamy of Science*, 39(10):1095–1100, 1953.

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

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

[29] A. F. Veinott. Extreme points of leontief substitution systems. *Linear Algebra and its Applications*, 1:181–194, 1968.

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

[31] A. F. Veinott. Lectures in dynamic programming and stochastic control. Technical report, 2008. Lecture Notes of MS&E351, Stanford University, Stanford.

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