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

Yinyu Ye *

April 20, 2010

Abstract

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

1 Introduction

General linear programming (LP) has a standard form

\begin{align}
\text{Primal: minimize } \quad & c^T x \\
\text{subject to } \quad & Ax = b, \\
\text{} & x \geq 0,
\end{align}

and its dual

\begin{align}
\text{Dual: maximize } \quad & b^T y \\
\text{subject to } \quad & s = c - A^T y \geq 0,
\end{align}

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

The Markov Decision Problem or Process (MDP) is a special class of linear programming due to de Ghellinck [3], D’Epenoux [4] and Manne [8]:

$$\begin{align*}
\text{minimize} & \quad c^T_1 x_1 + \ldots + c^T_i x_i + \ldots + c^T_k x_k \\
\text{subject to} & \quad (I - \gamma P_1)x_1 + \ldots + (I - \gamma P_i)x_i + \ldots + (I - \gamma P_k)x_k = e,
\end{align*}$$

(3)

where $I$ is the $m \times m$ identity matrix, $P_i$ is an $m \times m$ Markov or column stochastic matrix such that $e^T P_i = e^T$ and $P_i \geq 0$, $i = 1, \ldots, k$,

and $e$ is the vector of all ones. Here, decision vector of $x_i \in \mathbb{R}^m$ is the policy vector where each state takes the $i$th action from its action set, and $c_i$ is the cost vector associated with the policy vector. Comparing to the LP standard form, we have

$$A = [I - \gamma P_1, \ldots, I - \gamma P_k] \in \mathbb{R}^{m \times mk}, \quad b = e \in \mathbb{R}^m, \quad \text{and} \quad c = (c_1; \ldots; c_k) \in \mathbb{R}^{mk}.$$

Note that the dual (by adding slack variables) of (3) is given by:

$$\begin{align*}
\text{maximize} & \quad e^T y \\
\text{subject to} & \quad (I - \gamma P_1)^T y + s_1 = c_1, \\
& \quad \ldots \ldots \ldots \ldots \\
& \quad (I - \gamma P_i)^T y + s_i = c_i, \\
& \quad \ldots \ldots \ldots \ldots \\
& \quad (I - \gamma P_k)^T y + s_k = c_k, \\
& \quad s_1, \ldots, s_i, \ldots, s_k \geq 0.
\end{align*}$$

(4)

In MDP, $\gamma$ is the so-called discount factor or rate such that $\gamma = \frac{1}{1 + r} \leq 1$, where $r$ is the interest rate and it is assumed strictly positive so that $0 \leq \gamma < 1$. When $\gamma = 1$, problem (3) is infeasible and $y = e$ is a certificate for the infeasibility.

The MDP problem is to find a best action, among its $k$ actions, for each state so that the total cost is minimized. Thus, an optimal solution or policy vector to the MDP problem contains a specific action for each state, which forms an optimal basic feasible solution (BFS) to linear program (3). There are four current general methods for solving the MDP: the value iteration method, the policy-iteration method, regular LP interior-point algorithm, and a combinatorial interior-point algorithm. In terms of the worst-case complexity bound on the number of arithmetic operations, they (without a constant factor) are summarized in the following table for $k = 2$ (see Littman et al. [7], Mansour and Singh [9], Ye [14], and references therein).
where \( L(P_i, c_i, \gamma) \) is a size/complexity measure of data \((P_i, c_i, \gamma), i = 1, \ldots, k\). When data \((P_i, c_i, \gamma)\) are rational numbers, this measure is generally bounded above by the total bit size of all input data.

One can see from the table, both the value-iteration and policy-iteration methods are polynomial-time algorithms if the discount rate \(0 \leq \gamma < 1\) is fixed, but they are not strongly polynomial where the running time should be a polynomial only in \(m\) (or \(mk\)). The only known strongly polynomial-time algorithm is the combinatorial interior-point algorithm of Ye [14] when the rate is fixed. However, the policy-iteration method, due to Howard in 1960 [5], has been remarkably successful in practice where the number of iterations is typically bounded by \(O(mk)\). It turns out that the policy-iteration method is actually the simplex method, invented by Dantzig in late 1947 [2], for solving general LP with multiple pivots at each iteration; and the simplex method remains one of the very few extremely effective methods for general LP solvers.

In the past 50 years, many efforts have been put to resolve the worst-case complexity issue of the policy-iteration method or the simplex method. The negative result of Klee and Minty [6] emerged in 1972 that the simplex method, with the 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 [10] showed that one simple policy-iteration method, where in each iteration only the action for the state with the smallest index is updated, needs an exponential number of iterations to compute an optimal policy for a specific MDP problem regardless of discount rates. On the positive side, Papadimitriou and Tsitsiklis [11] showed in 1987 that an MDP with deterministic transitions (the entries of \(P_i\) are all 0’s and 1’s) can be solved in strongly polynomial time as a Minimum-Mean-Cost-Cycle problem. However, it is still unknown if the policy-iteration method can solve the deterministic MDP in polynomial time. So far, the best worst-case iteration upper bound for the policy-iteration method, independent of \(c(P_i, c_i, \gamma)\), is \(\frac{2m}{m} (with k = 2)\) given in 1999 by Mansour and Singh [9].

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 also a strongly polynomial-time algorithm for MDP with fixed discount rate \(0 \leq \gamma < 1\). The number of its iterations is bounded by

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

Since the policy-iteration method with the all-negative-reduced-cost pivoting rule is at least as good as the the simple policy-iteration method, it is also a strongly polynomial-time
algorithm with the same iteration complexity bound. Therefore, there is no complexity difference between the simplex method and interior-point algorithms for MDP with a fixed discount rate. Our proof is based on a combinatorial cross-over event similar to the one in Vavasis and Ye [12, 14]. We remark that, if the discount rate is an input, it remains open whether or not the policy-iteration method is polynomial for MDP, or whether or not there exists a strongly polynomial-time algorithm for MDP or LP in general.

2 MDP Properties and The Simplex Method

We first describe few general LP and MDP theorems and the classic simplex method. Without loss of generality and for simplicity, we fix $k$, the number of possible actions taken by each state, to 2 in the rest of the paper:

$$\begin{align*}
\text{minimize} & \quad c_1^T x_1 + c_2^T x_2 \\
\text{subject to} & \quad (I - \gamma P_1) x_1 + (I - \gamma P_2) x_2 = e, \\
& \quad x_1, x_2 \geq 0;
\end{align*}$$

with its dual

$$\begin{align*}
\text{maximize} & \quad e^T y \\
\text{subject to} & \quad (I - \gamma P_1)^T y + s_1 = c_1, \\
& \quad (I - \gamma P_2)^T y + s_2 = c_2, \\
& \quad s = (s_1; s_2) \geq 0.
\end{align*}$$

Comparing to the LP standard form, we have

$$A = [I - \gamma P_1, I - \gamma P_2] \in \mathbb{R}^{m \times 2m}, \quad b = e \in \mathbb{R}^m, \quad \text{and} \quad c = (c_1; c_2) \in \mathbb{R}^{2m}.$$ 

2.1 MDP Properties

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

$$\begin{align*}
Ax &= b, \\
A^T y + s &= c, \\
SXe &= 0, \\
x \geq 0, \quad s \geq 0
\end{align*}$$

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

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

\[
A_B = I - \gamma P
\]

where \( P \) is an \( m \times m \) Markov matrix chosen from columns of \([P_1, P_2]\), and the reverse is also true. (This can be seen from that, otherwise, the basis has at least one row consisting of all non-positive elements so that its inner product with \( x_B(\geq 0) \) is non-positive, but the right-hand side is strictly positive.) The following lemma is given in [14] whose proof was based on Dantzig [1, 2], and Veinott [13].

**Lemma 1** The MDP has the following properties:

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

\[
e^T x = \frac{m}{1 - \gamma},
\]

where \( x = (x^1; x^2) \) of all feasible solutions.

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

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

### 2.2 The Simplex and Policy Iteration Method

Let us start with \( x_1 \) being the initial basic feasible solution of (5) where the initial basic index set is denoted by \( B^0 \). Then, the MDP can be rewritten as an equivalent problem

\[
\begin{align*}
\text{minimize} & \quad \bar{c}_2^T x_2 \\
\text{subject to} & \quad (I - \gamma P_1)x_1 + (I - \gamma P_2)x_2 = e, \\
& \quad x_1, x_2 \geq 0.
\end{align*}
\]

(7)

\( \bar{c}_2 \) is called the reduced cost vector for the non-basic variables \( x_2 \):

\[
\bar{c}_2 = c_2 - (I - \gamma P_2)^T y^0
\]

and

\[
y^0 = (I - \gamma P_1)^{-T} c_1.
\]

The initial primal basic feasible solution is given by

\[
x^0 = (x^0_1 = (I - \gamma P_1)^{-1} e; \ x^0_2 = 0).
\]
If $\bar{c}_2 \geq 0$, then the current BFS is optimal. Otherwise, let $\Delta^0 = -\min(\bar{c}_2)$ with $\bar{c}_2 = -\Delta^0$. Then, the classic simplex method takes $(x_2)_{\bar{i}}$ as the in-coming basic variable to replace the old one $(x_1)_{\bar{i}}$, and the method repeats with the new BFS denoted by $x^1$. The method will break a tie arbitrarily, and it updates exact one state action in one iteration, that is, it only updates the state with the most negative reduced cost. This is called the simplex or simple policy-iteration method with the most-negative-reduced-cost update or pivoting rule.

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

3 Proof of Strong Polynomiality

First, we have

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

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

Moreover,

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

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

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

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

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

which leads to the desired inequality. \[\square\]
Lemma 3 If the initial BFS $x^0$ is not optimal, then there is $i^0 \in B^0$ such that
\[(s^*_i)_{i^0} \geq \frac{1 - \gamma}{m^2} \left( c^T x^0 - z^* \right),\]
where $s^*$ is an optimal dual slack vector of (6). And for any basic feasible solution $x^t$ of (5), $t \geq 1$,
\[(x^t_i)_{i^0} \leq \frac{m^2}{1 - \gamma} \cdot \frac{c^T x^t - z^*}{c^T x^0 - z^*}.

Proof. Since
\[c^T x^0 - z^* = (s^*)_T x^0 = (s^*_i)^T x^0 = \sum_{i=1}^m (s^*_i)_i (x^0_i),\]
there must be an $i^0 \in B^0$ such that
\[(s^*_i)_{i^0} (x^0_i) \geq \frac{1}{m} \left( c^T x^0 - z^* \right).

Then, from Lemma 1 we have
\[(x^0_i)_{i^0} \leq \frac{m}{1 - \gamma} \text{ so that }\]
\[(s^*_i)_{i^0} \geq \frac{1 - \gamma}{m^2} \left( c^T x^0 - z^* \right).

Furthermore, for any basic feasible solution $x^t$,
\[c^T x^t - z^* = (s^*)_T x^t \geq (s^*)_{i^0} (x^t)_{i^0},\]
so that
\[(x^t_i)_{i^0} \leq \frac{c^T x^t - z^*}{(s^*_i)_{i^0}} \leq \frac{m^2}{1 - \gamma} \cdot \frac{c^T x^t - z^*}{c^T x^0 - z^*}.

From Lemma 2, after $t$ iterations of the simplex method, we have
\[\frac{c^T x^t - z^*}{c^T x^0 - z^*} \leq \left( 1 - \frac{1 - \gamma}{m} \right)^t.

Therefore, after $\frac{m}{1 - \gamma} \cdot \log \left( \frac{m^2}{1 - \gamma} \right)$ iterations from the initial BFS $x^0$, we must have, from Lemma 3,
\[(x^t_i)_{i^0} \leq \frac{m^2}{1 - \gamma} \cdot \frac{c^T x^t - z^*}{c^T x^0 - z^*} < 1,

for all $t \geq \frac{m}{1 - \gamma} \cdot \log \left( \frac{m^2}{1 - \gamma} \right)$. But for any basic variable, its value should be greater than or equal to 1 from Lemma 1, hence it must be true $(x^t_i)_{i^0} = 0$. This leads to our key result:
Theorem 1 There is a basic variable in the initial basic feasible solution \( x^0 \) that would never be in the basis again after \( \frac{m}{1-\gamma} \cdot \log \left( \frac{m^2}{1-\gamma} \right) \) iterations of the simplex method with the most-negative-reduced-cost pivoting rule.

The event described in Theorem 1 can be viewed as the cross-event of Vavasis and Ye [12, 14]: a variable, although we don’t know which one, was a basic variable initially but it will never be a basic variable after a certain number of iterations during the iterative process of the simplex or simple policy-iteration method with the most-negative-reduced-cost pivoting rule. Clearly, these cross-over events can happen only \((mk - m)\) times for any \(k\) (we can subtract it by \(m\) because the cross event can only happen to a variable that is not in any optimal basis and there should be at least \(m\) basic variables in an optimal BFS), thus, we reach our final conclusion:

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

As a corollary, we have

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

References


