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

Yinyu Ye

Department of Management Science and Engineering and
Institute of Computational and Mathematical Engineering
Stanford University

October 26, 2010

Outline

- The MDP Problem and its History
- The Simplex and Policy-Iteration Methods
- The Main Result: Strong Polynomiality
- Proof Sketch of the Main Result
- Remarks and Open Questions

The Markov Decision Process

- Markov decision processes (MDPs), named after Andrey Markov, provide a mathematical framework for modeling sequential decision-making in situations where outcomes are partly random and partly under the control of a decision maker.
- MDPs are useful for studying a wide range of optimization problems solved via dynamic programming, where it was known at least as early as the 1950s (cf. Bellman 1957).
- At each time step, the process is in some state/agent s, and the decision maker choose any action that is available in state s. The process responds at the next time step by randomly moving into a new state s', and giving the decision maker a corresponding reward or cost $C_a(s, s')$.

The Markov Decision Process continued

- The probability that the process chooses s' as its new state is influenced by the chosen action. Specifically, it is given by the state transition function $P_a(s, s')$. Thus, the next state s' depends on the current state s and the decision maker's action a.
- But given s and a, it is conditionally independent of all previous states and actions; in other words, the state transitions of an MDP possess the Markov property.
- Each state (or agent) is myopic and can be selffish. But when every state chooses an optimal action among its available ones, the process reaches optimality, and they form an optimal policy for all states.

Applications of The Markov Decision Process

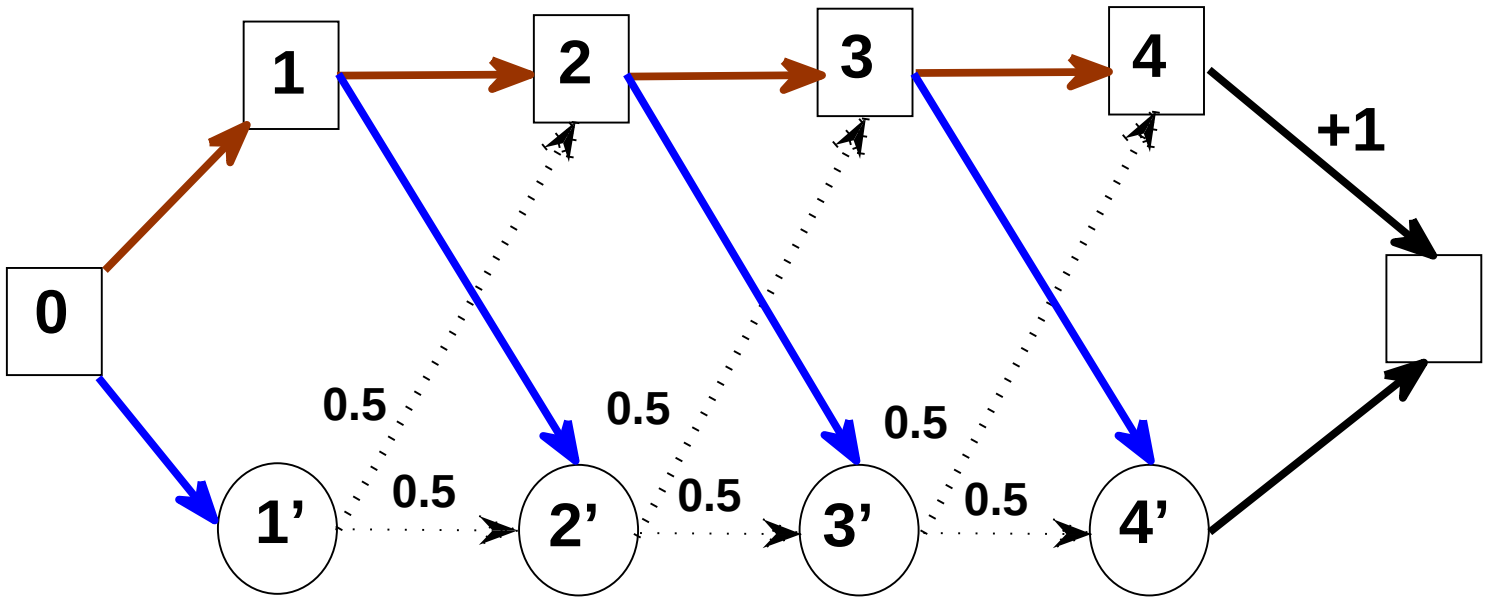
MDP is one of the most fundamental dynamic decision models in

- Mathematical science
- Physical science
- Management science
- Social Science

Modern applications include dynamic planning, reinforcement learning, social networking, and almost all other sequential decision makings.

A Markov Decision Process Example

by Melekopoglou and Condon 1990; red edges are actions currently taken:



The LP Form of The Markov Decision Process

$$\begin{array}{rccccc} \text{minimize} & \mathbf{c}_1^T\mathbf{x}_1 & \dots & +\mathbf{c}_m^T\mathbf{x}_m & & \\ \text{subject to} & (E_1-\gamma P_1)\mathbf{x}_1 & \dots & +(E_m-\gamma P_k)\mathbf{x}_k & = & \mathbf{e}, \\ & \mathbf{x}_1, & \dots & \mathbf{x}_m, & \geq & \mathbf{0}. \end{array}$$

where E_i is the $m \times k$ matrix where the ith row are all ones and everywhere else are zeros, P_i is an $m \times k$ Markov or column stochastic matrix such that

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

and $\mathbf{e}$ is the vector of all ones.

The LP Form of The Markov Decision Process continued

Decision $\mathbf{x}_i \in \mathbf{R}^k$ is the action vector of state i and only one of them is chosen as the winning action, ant its value represents visiting frequencies of state i, and $\mathbf{c}_i$ is the cost vector associated with the action vector.

In the MDP, γ is the so-called discount rate or factor such that

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

where r is an interest rate and it is assumed strictly positive so that $0 \leq \gamma < 1$.

The MDP Example in LP form

a:	(0_1)	(0_2)	(1_1)	(1_2)	(2_1)	(2_2)	(3_1)	(3_2)	(4_1)	$(4'_1)$
c:	0	0	0	0	0	0	0	0	1	0
(0)	1	1	0	0	0	0	0	0	0	0
(1)	$-\gamma$	0	1	1	0	0	0	0	0	0
(2)	0	$-\gamma/2$	$-\gamma$	0	1	1	0	0	0	0
(3)	0	$-\gamma/4$	0	$-\gamma/2$	$-\gamma$	0	1	1	0	0
(4)	0	$-\gamma/8$	0	$-\gamma/4$	0	$-\gamma/2$	$-\gamma$	0	$1-\gamma$	0
(4′)	0	$-\gamma/8$	0	$-\gamma/4$	0	$-\gamma/2$	0	$-\gamma$	0	$1-\gamma$

The Dual Problem

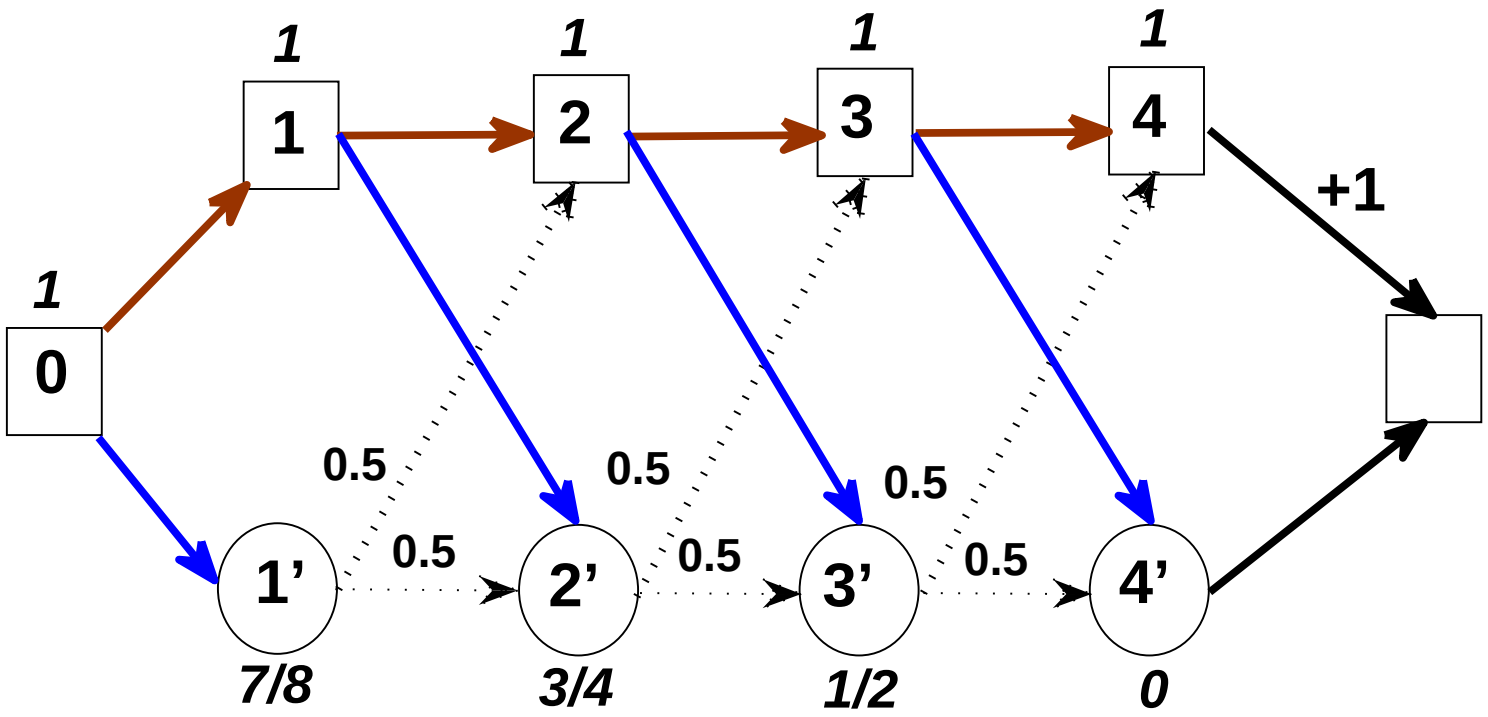
$$
\begin{array}{rrcl}
\text{maximize} & \mathbf{e}^T\mathbf{y} & & \\
\text{subject to} & (E_1 - \gamma P_1)^T\mathbf{y} & \leq & \mathbf{c}_1, \\
& \cdots & \cdots & \cdots \\
& (E_i - \gamma P_i)^T\mathbf{y} & \leq & \mathbf{c}_i, \\
& \cdots & \cdots & \cdots \\
& (E_m - \gamma P_m)^T\mathbf{y} & \leq & \mathbf{c}_m.
\end{array}
$$

The MDP-LP Formulation

- A policy consists of a chosen action for each state among its all possible actions that form a basic feasible solution $\mathbf{x}_B$ to the LP problem, which represent visiting frequencies the states, and entries of $\mathbf{y}$ of the dual represent the state values at the current policy.
- An optimal policy consists of a winning action for each state from its all possible actions that form an optimal basic feasible solution $\mathbf{x}^*$ to the LP problem, which represent visiting frequencies the states, and entries of $\mathbf{y}^*$ of the dual represent the optimal state values.
- The MDP-LP problem is to find an optimal policy for the net present value of the process over the infinite horizon with discount rate γ.

Pricing: the Values of the States

Red edges are actions currently taken, and with a value on each state:



Historical Events of the Markov Decision Process Methods I

- Bellman (1957) developed an approximation method called the value iteration method to approximate the optimal state values.
- Another best known method is due to Howard (1960) and is known as the policy iteration method, which generate an optimal policy in finite number of iterations in a distributed and decentralized way
- de Ghellinck (1960), D'Epenoux (1960) and Manne (1960) showed that the MDP has an LP representation, so that it can be solved by the simplex method of Dantzig (1947) in finite number of steps, and the Ellipsoid method of Kachiyan (1979) in polynomial time.

Historical Events of the Markov Decision Process Methods II

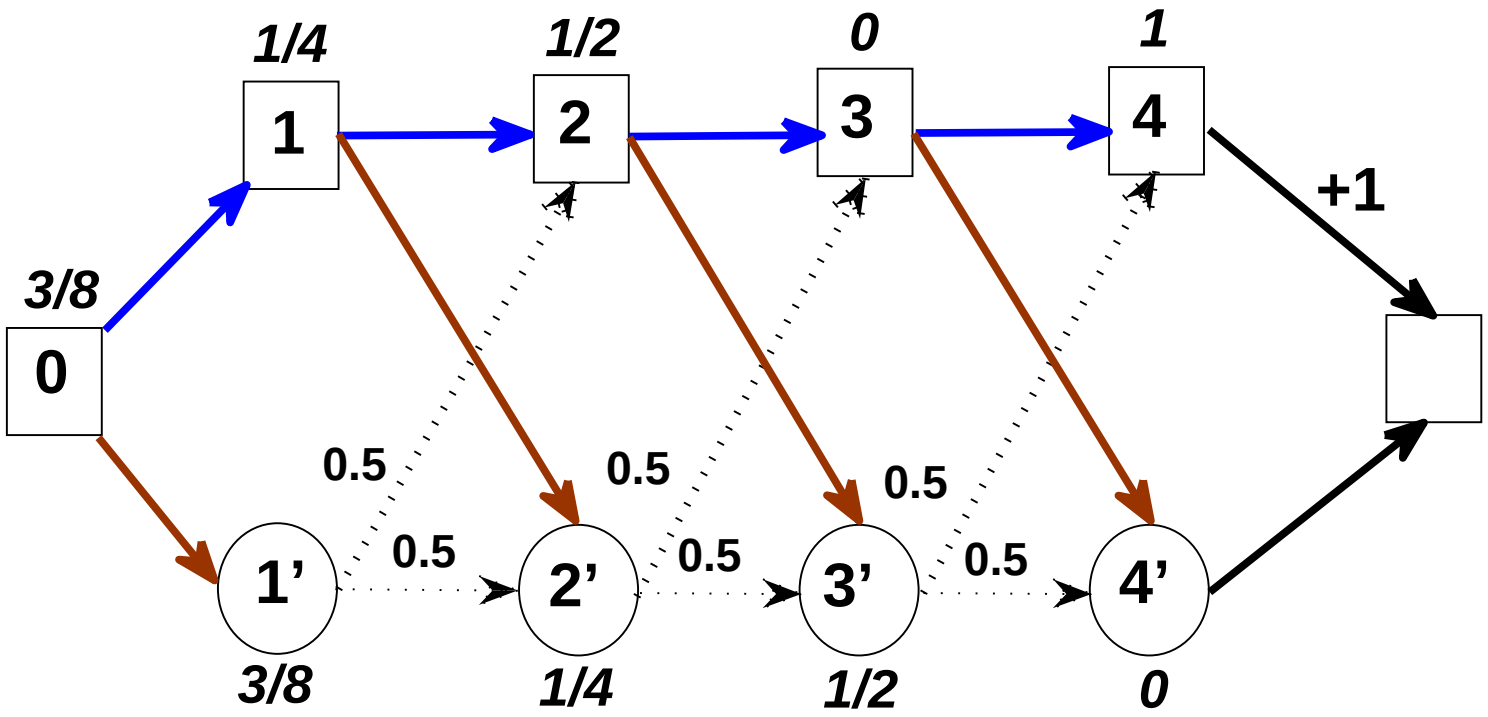
- Papadimitriou and Tsitsiclis (1987) gave a theoretical complexity analysis of the MDP and showed that if P_i is deterministic, then the MDP can be solved in *strongly* polynomial time.
- Tseng (1990) showed that the value iteration method generates an optimal policy in polynomial time for fixed discount γ, built upon Bertsekas' (1987) work that the value iteration method converges to the optimal policy in finite number of iterations.
- Y (2005) showed that the MDP, for fixed discount γ, can be solved in *strongly* polynomial time by a combinatorial interior-point method.

Polynomial vs Strongly Polynomial

- If the computation time of an algorithm, the total number of basic arithmetic operations needed, of solving the problem is bounded by a polynomial in m, k, and the total bits, L, to encode the problem data, then the algorithm is called polynomial-time algorithms.
- If the computation time of an algorithm, the total number of basic arithmetic operations needed, of solving the problem is bounded by a polynomial in m and k, then the algorithm is called *strongly* polynomial-time algorithms.

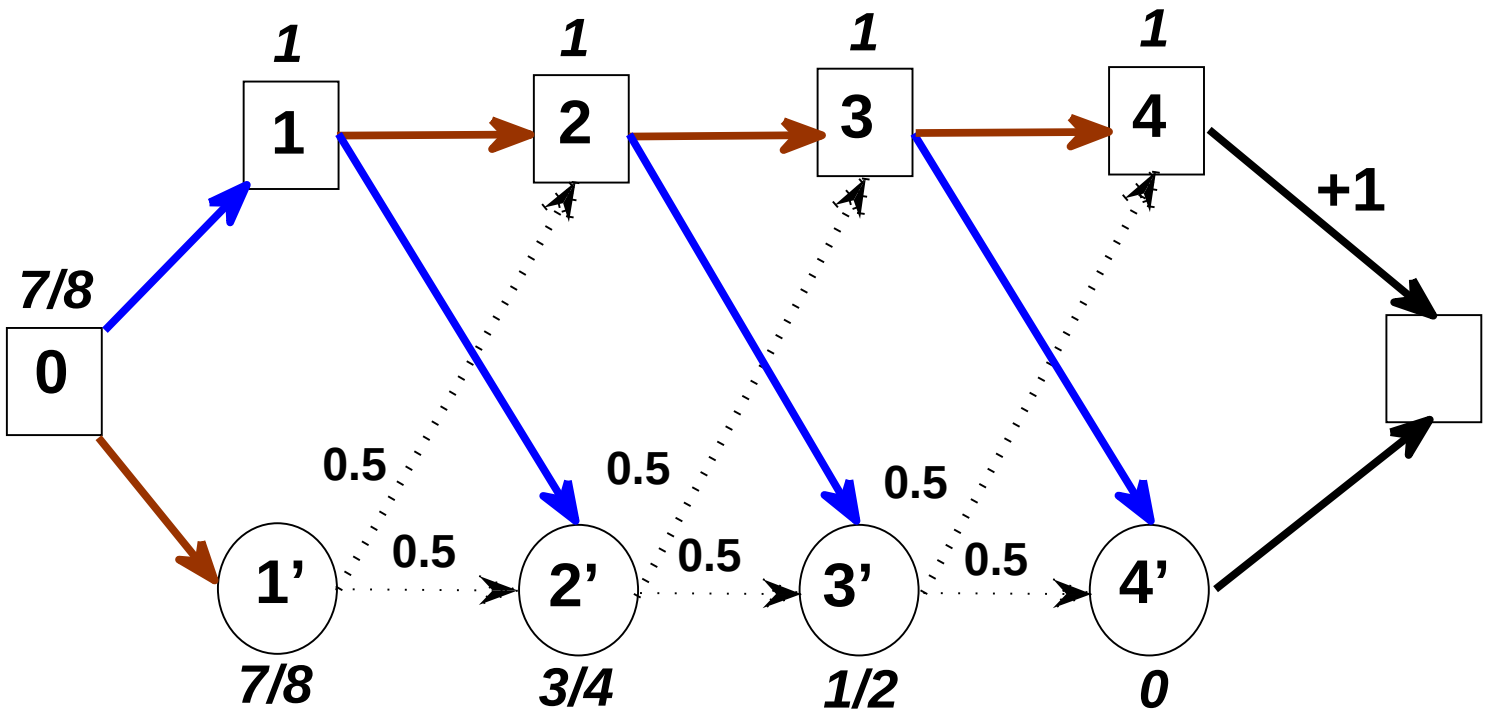
The Policy Iteration Method

Red edges are actions currently taken, with a new value on each state:



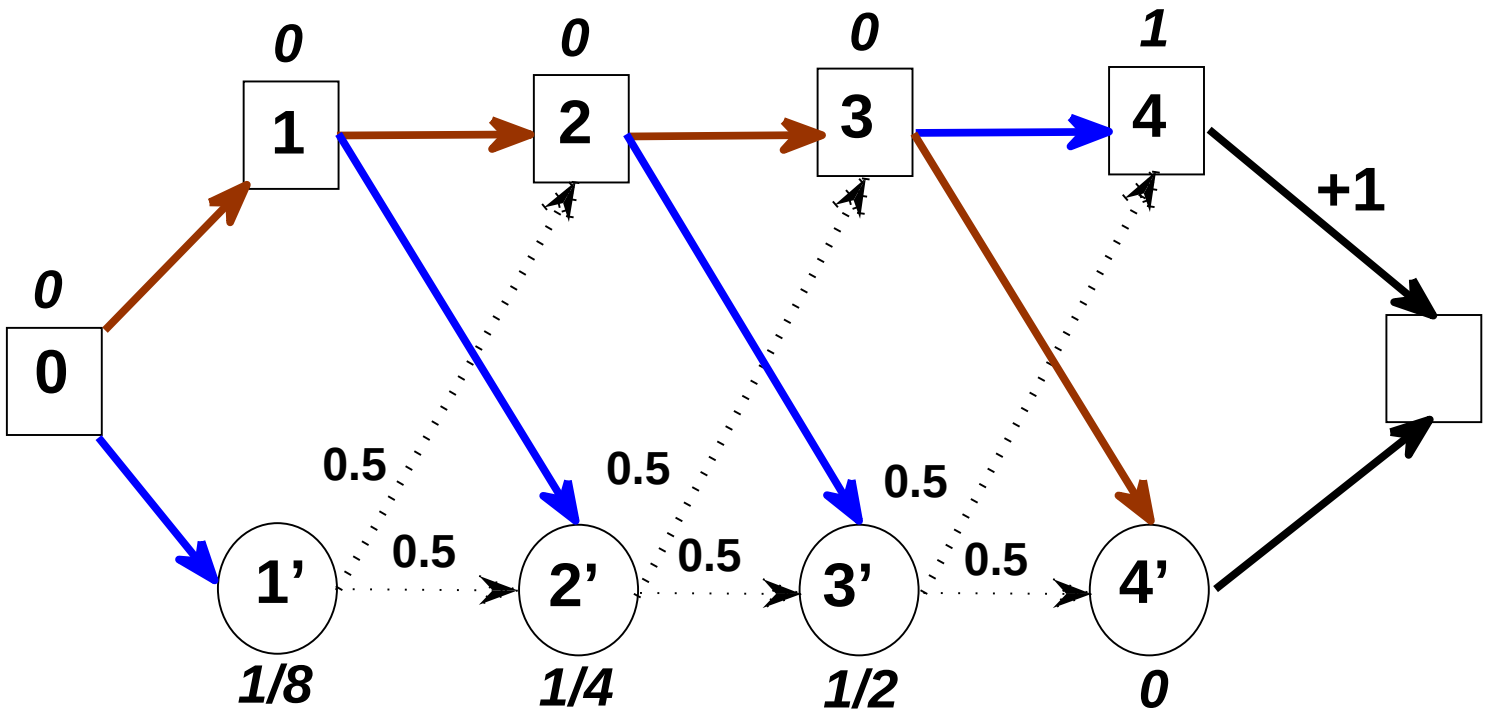
The Simple Policy Iteration or Simplex Method: index rule

Red edges are actions currently taken, with a new value on each state:



The Simple Policy Iteration or Simplex Method: greedy rule

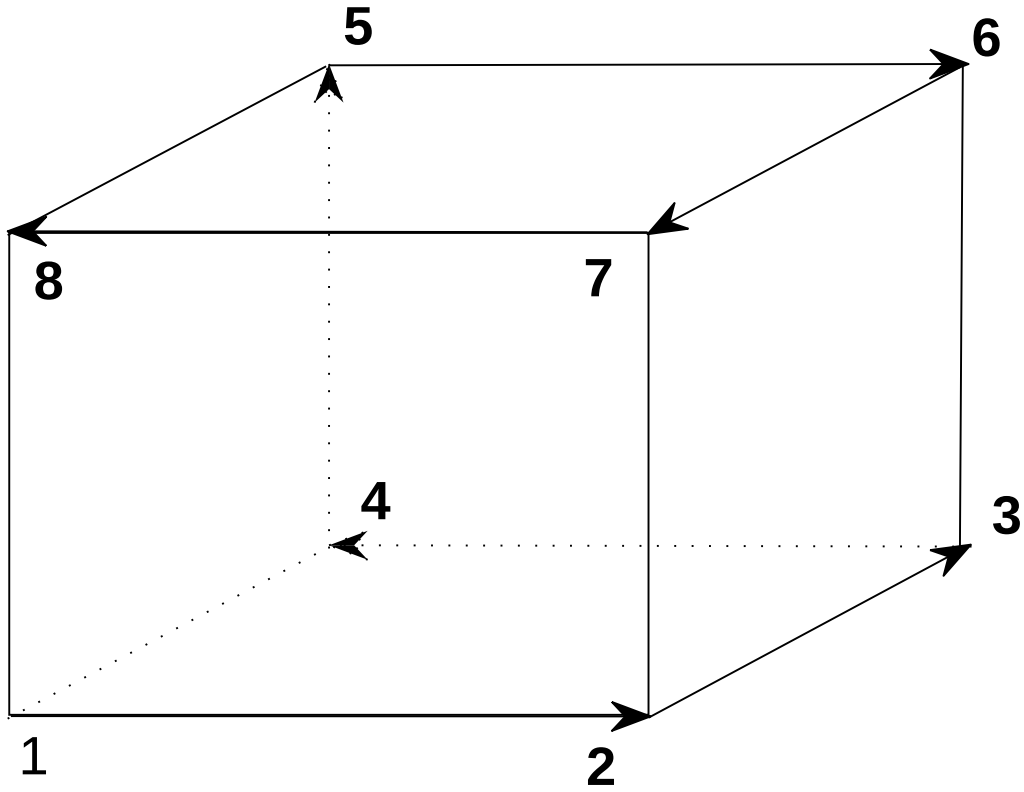
Red edges are actions currently taken, with a new value on each state:



Facts of the Policy Iteration and Simplex Methods

- The policy iteration method has been shown to be the one of the most effective and widely used methods in practice for solving MDPs.
- It turned out that the policy-iteration method is actually the simplex method for solving general LP with block pivots at each iteration; and the simplex method also remains one of the extremely effective general LP solvers.
- In theory, Klee and Minty (1972) have showed that the simplex method, with the greedy (most-negative-reduced-cost) pivoting rule, necessarily takes an exponential number of iterations to solve a carefully designed LP problem.

The Klee and Minty Example I



The Klee and Minty Example II

The figure above illustrates the fact that the sequence of vectors v^k corresponds to a path on the edges of the 3-cube. The path visits each vertex of the cube once and only once. Such a path is said to be Hamiltonian.

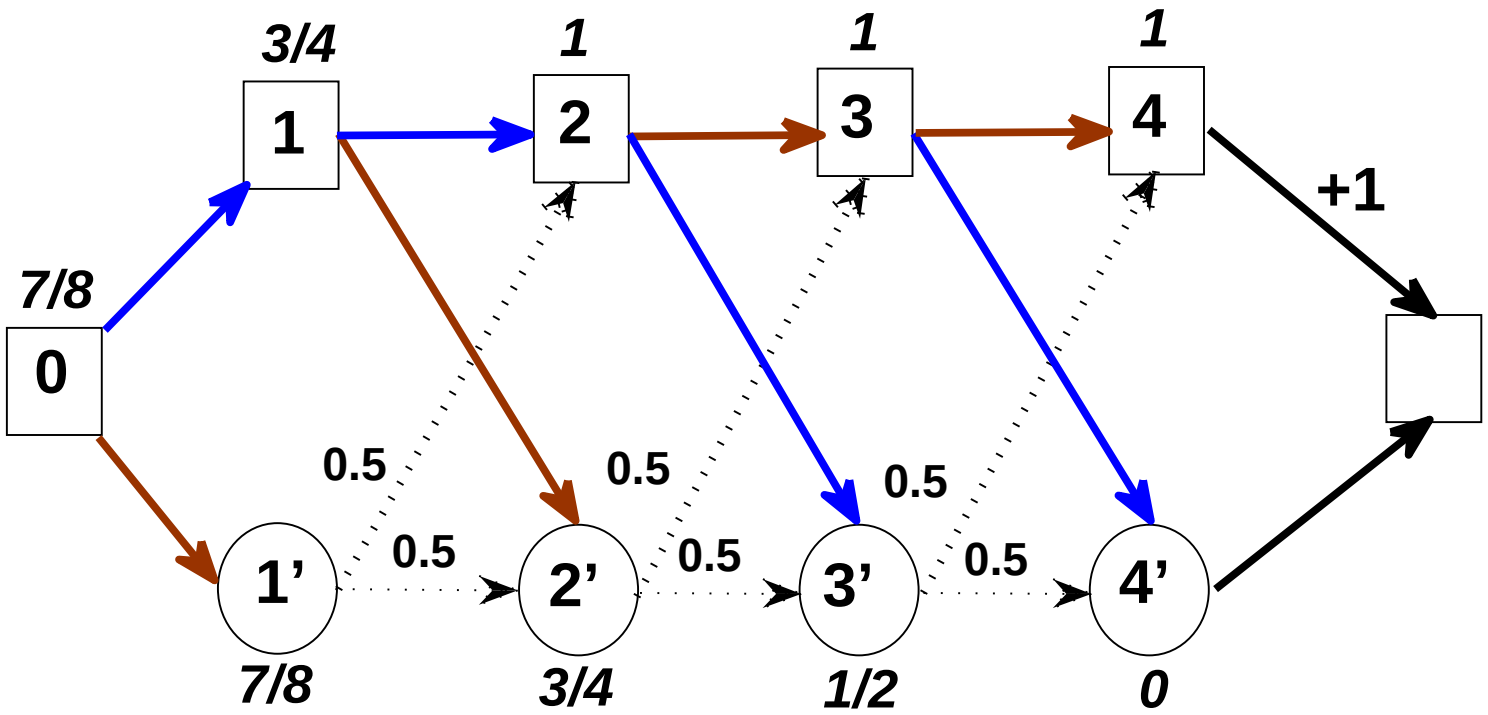
There is an amusing recreational literature that connects Hamiltonian path with certain puzzles. See Martin Gardner, "Mathematical games, the curious properties of the Gray code and how it can be used to solve puzzles," *Scientific American* 227 (August 1972) pp. 106-109. See also, S.N. Afriat, *The Ring of Linked Rings*, London: Duckworth, 1982.

More Negative Facts of the Policy Iteration and Simplex Methods

- A similar negative result of Melekopoglou and Condon (1990) showed that a 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.
- Most recently, Fearnley (2010) showed that the policy iteration method needs an exponential number of iterations for a undiscounted finite-horizon MDP.

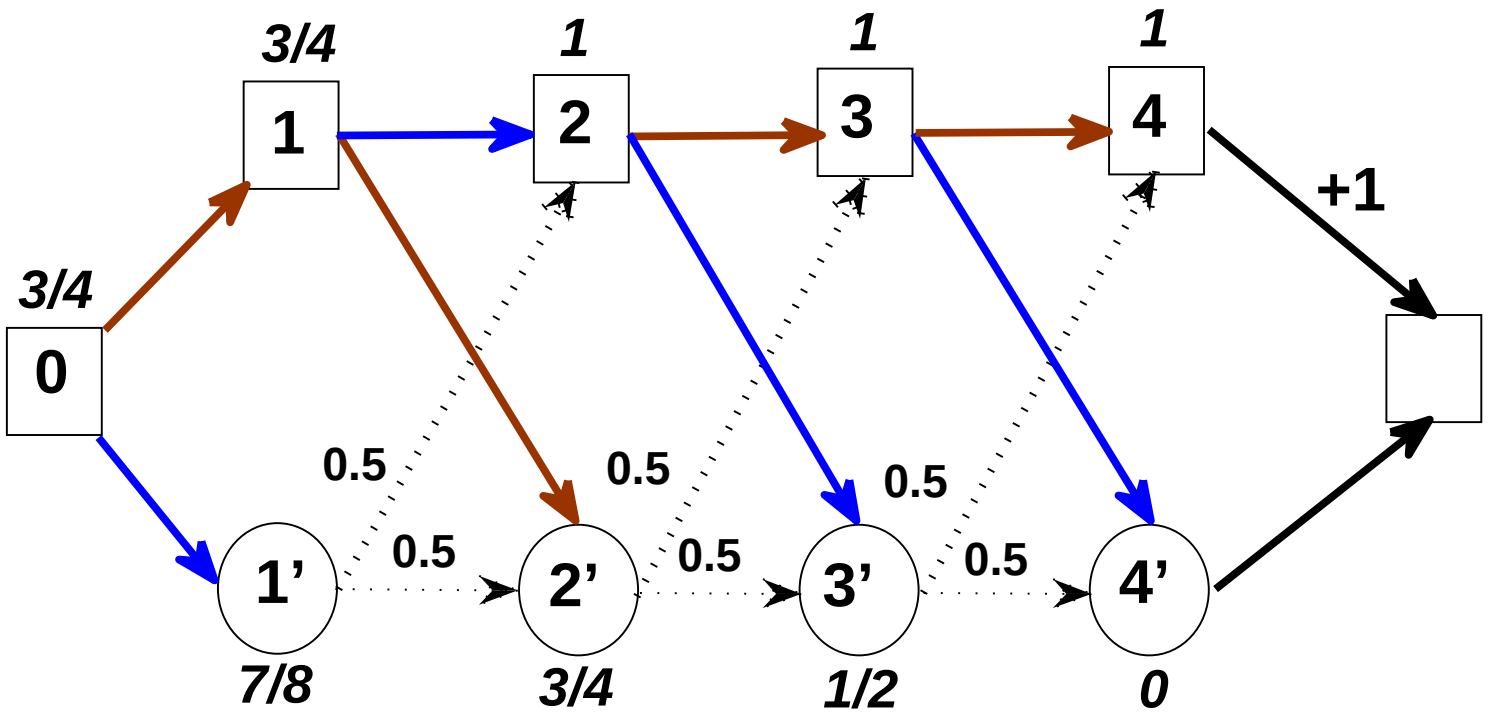
The Simple Policy Iteration or Simplex Method: index rule II

Red edges are actions currently taken, with a new value on each state:



The Simple Policy Iteration or Simplex Method: index rule III

Red edges are actions currently taken, with a new value on each state:



Positive Facts of the Policy Iteration and Simplex Methods

- Puterman (1994) showed that the policy iteration method converges no more slowly than the value iteration method, so that it is a polynomial time algorithm for the discounted MDP.
- Mansour and Singh (1999) gave an upper bound of the policy iteration method of $\frac{k^m}{m}$ iterations for the MDP.

Complexity Summaries of The MDP Methods

Value-Iter	Policy-Iter	LP-Alg	Comb IP
$\frac{m^2 L}{1-\gamma}$	$\min\left\{\frac{m^3 2^m}{m}, \frac{m^2 L}{1-\gamma}\right\}$	$m^3 L$	$m^4 \cdot \log \frac{1}{1-\gamma}$

where L is a total bits to encode the problem data $(P_i, \mathbf{c}_i, \gamma)$, $i = 1, \ldots, k$ (for simplicity we have assumed that $k = 2$ in this table).

The main objective of our work is to narrow the gap between the practical performance and theoretical complexity of the simplex and/or policy iteration methods.

Our Result

- The classic simplex method, or the simple policy-iteration method, with the greedy (steepest descent) or most-negative-reduced-cost pivoting rule, is a *strongly* polynomial-time algorithm for MDP with fixed discount rate $0 \leq \gamma < 1$. The number of its iterations is bounded by

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

- The policy-iteration method with the all-negative-reduced-cost pivoting rule is at least as good as the simple policy-iteration method, it is also a strongly polynomial-time algorithm with the same iteration complexity bound.

High Level Ideas of the Proof

- Create a combinatorial event similar to the one in Vavasis and Ye (1994) developed for general LP and Y (2005) for MDP.
- The event will happen in a strongly polynomial number of iterations.
- In particular, in a polynomial number of iterations, a non-winning action for at least one state will be found, and the state would never take this non-winning action again for the rest of iterations.
- The event then repeats for another non-winning action for a state

The MDP Solution Properties I

$$\begin{array}{rccccc} \text{minimize} & \mathbf{c}_1^T\mathbf{x}_1 & & \mathbf{c}_2^T\mathbf{x}_2 & & \\ \text{subject to} & (I-\gamma P_1)\mathbf{x}_1 & +(I-\gamma P_2)\mathbf{x}_2 & = & \mathbf{e}, & (1) \\ & \mathbf{x}_1, & & \mathbf{x}_2 & \geq & \mathbf{0}; \end{array}$$

where $\mathbf{x}_j$ is the decision vector of all states when each of them takes the jth action. Its dual

$$\begin{array}{rccc} \text{maximize} & \mathbf{e}^T\mathbf{y} & & \\ \text{subject to} & (I-\gamma P_1)^T\mathbf{y}+\mathbf{s}_1 & = & \mathbf{c}_1, \\ & (I-\gamma P_2)^T\mathbf{y}+\mathbf{s}_2 & = & \mathbf{c}_2, \\ & \mathbf{s}_1, \mathbf{s}_2 & \geq & \mathbf{0}. \end{array} \qquad (2)$$

The MDP Solution Properties II

Every basic feasible solution (BFS) basis has the Leontief substitution form:

$$A_B = I - \gamma P,$$

and the reverse is also true.

Lemma

The MDP has the following properties:

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

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

   for all feasible solutions $\mathbf{x}$.

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

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

The Simplex and Policy-Iteration Methods I

Let us start with $\mathbf{x}_1$ being the initial basic feasible solution of (1) where the initial basic index set is denoted by B^0. Then, the MDP can be rewritten as a reduced-cost problem

$$\begin{array}{rccccc} \text{minimize} & & \bar{\mathbf{c}}_2^T\mathbf{x}_2 & & & \\ \text{subject to} & (I-\gamma P_1)\mathbf{x}_1 & +(I-\gamma P_2)\mathbf{x}_2 & = & \mathbf{e}, & (3) \\ & \mathbf{x}_1, & \mathbf{x}_2 & \geq & \mathbf{0}. & \end{array}$$

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

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

and

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

The initial primal basic feasible solution is given by

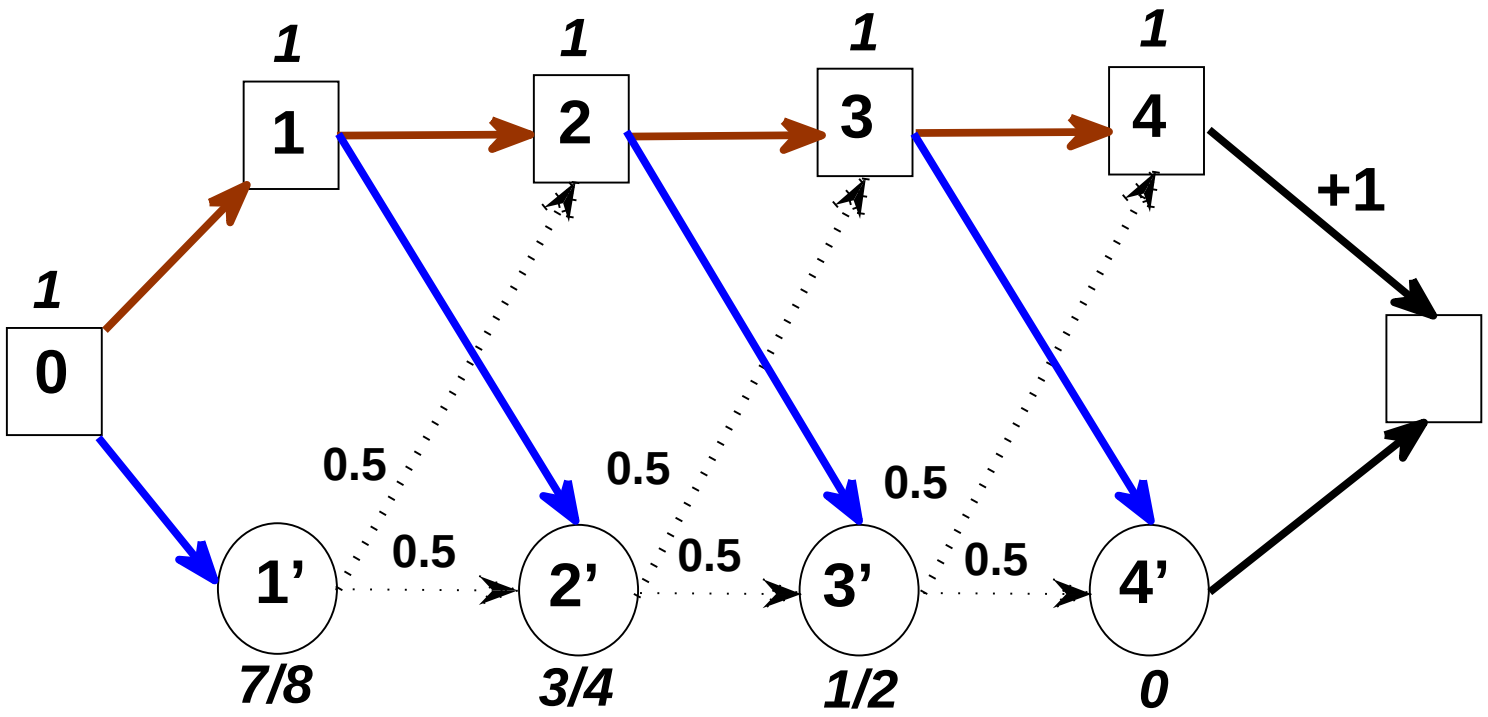
$$\mathbf{x}^0 = (\mathbf{x}_1^0 = (I-\gamma P_1)^{-1}\mathbf{e};\ \mathbf{x}_2^0 = \mathbf{0}).$$

$$\mathbf{y}^0 = (0;\ 0;\ 0;\ 0;\ -1).$$

a:	(0_1)	(0_2)	(1_1)	(1_2)	(2_1)	(2_2)	(3_1)	(3_2)
c:	0	$-1/8$	0	$-1/4$	0	$-1/2$	0	-1
(0)	1	1	0	0	0	0	0	0
(1)	-1	0	1	1	0	0	0	0
(2)	0	$-1/2$	-1	0	1	1	0	0
(3)	0	$-1/4$	0	$-1/2$	-1	0	1	1
(4)	0	$-1/8$	0	$-1/4$	0	$-1/2$	-1	-1

Reduced Cost of the States at the Current Policy

Red edges are actions currently taken, with a new value on each state:



The Simplex and Policy-Iteration Methods II

- Let $\Delta^0 = -\min(\bar{\mathbf{c}}_2) > 0$ with $(\bar{\mathbf{c}}_2)_{\bar{i}} = -\Delta^0$. Then, the classic simplex method takes $(\mathbf{x}_2)_{\bar{i}}$ as the in-coming basic variable to replace the old one $(\mathbf{x}_1)_{\bar{i}}$, and the method repeats with the new BFS denoted by $\mathbf{x}^1$.
- The method will break a tie arbitrarily, and it updates exact one state action in one iteration, that is, it only updates a state action with the most negative reduced cost.
- The method generates a sequence of BFSs or policies denoted by $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^t, \ldots$.

The Simplex and Policy-Iteration Methods III

- In contrast to the simplex method, the policy iteration method with the all-negative-reduced-cost update 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).
- This is possible due to the structure of the MDP, and the method repeats with the new BFS.
- Again, the method generate a sequence of BFSs denoted by $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^t, \ldots$.

Proof of Strong Polynomiality I

Lemma

Let z^* be the minimal objective value of (1). Then,

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

Moreover,

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

Proof Sketch of the Lemma

For problem (3), its minimal objective value is bounded from below by $-\frac{m}{1-\gamma} \cdot \Delta^0$.

The minimal objective value of (3) differs from the one of (1) exactly by $\mathbf{c}^T\mathbf{x}^0$.

Since at the new BFS $\mathbf{x}^1$, the new basic variable value for state $\bar{i}$ is greater than or equal to 1 from Lemma 1, the objective value of the new BFS of problem (3) is decreased by at least Δ^0. Thus,

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

Proof of Strong Polynomiality II

Lemma

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

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

where $\mathbf{s}^*$ *is an optimal dual slack vector of (2). And for any basic feasible solution* $\mathbf{x}^t$ *of (1),* $t \geq 1$,

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

Proof Sketch of the Lemma

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

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

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

But $(\mathbf{x}_1^0)_{i^0} \leq \frac{m}{1-\gamma}$ so that

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

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

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

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

Proof of Strong Polynomiality III

Theorem

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

There is a state who will stick to its winning action regardless what others states do, although we still don't know which state is.

This is called the "cross-over" event.

Proof Sketch of the Theorem

After t iterations of the simplex method, we have

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

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

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

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

Strong Polynomiality the Simplex Method

Clearly, these "cross-over events" can happen only $(mk - m)$ times for any k, thus we reach the final conclusion:

Theorem

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* $O(m^2 k)$ *arithmetic operations.*

Strong Polynomiality of the Policy Iteration Method

The objective reduction rate of the policy iteration method is also guaranteed by

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

and the winning action taken by the simplex method is always included in the winning actions taken by the policy iteration method.

Corollary

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.*

Hansen, Miltersen, and Zwick (2010): $\frac{m(k-1)}{1-\gamma} \cdot \log\left(\frac{m}{1-\gamma}\right)$ (only for the policy iteration method).

Extensions to other MDPs

Every BFS basis of the MDP has the Leontief substitution form:

$$A_B = I - P,$$

where $P \geq \mathbf{0}$ and the spectral radius of P is bounded by $\gamma < 1$. Then, $\mathbf{e}^T(I-P)^{-1}\mathbf{e} \leq \frac{m}{1-\gamma}$.

Corollary

The simplex or simple policy-iteration method with the most-negative-reduced-cost pivoting rule terminates at most $\frac{m^2(k-1)}{1-\gamma} \cdot \log\left(\frac{m^2}{1-\gamma}\right)$ *iterations, and so does the policy iteration method, for solving MDPs with bounded spectral radius for every BFS.*

Remarks and Open Questions I

- The performance of the simplex method is very sensitive to the pivoting rule.
- Tatonnement and decentralized process works under the Markov property.
- Greedy or Steepest Descent works when there is a discount!
- Multi-updates works better than a single-update does; policy iteration vs. simplex iteration.
- Don't believe the negative news, and be positive ...

Remarks and Open Questions II

- What about the value iteration method?
- Can the iteration bound for the simplex method be reduced to linear?
- Is the policy iteration method polynomial for the MDP regardless of discount rate γ or input data?
- Is there an MDP algorithm at all whose running time is *strongly* polynomial regardless of discount rate γ?
- Is there a strongly polynomial-time algorithm for LP?