Recent Progresses on the Simplex Method

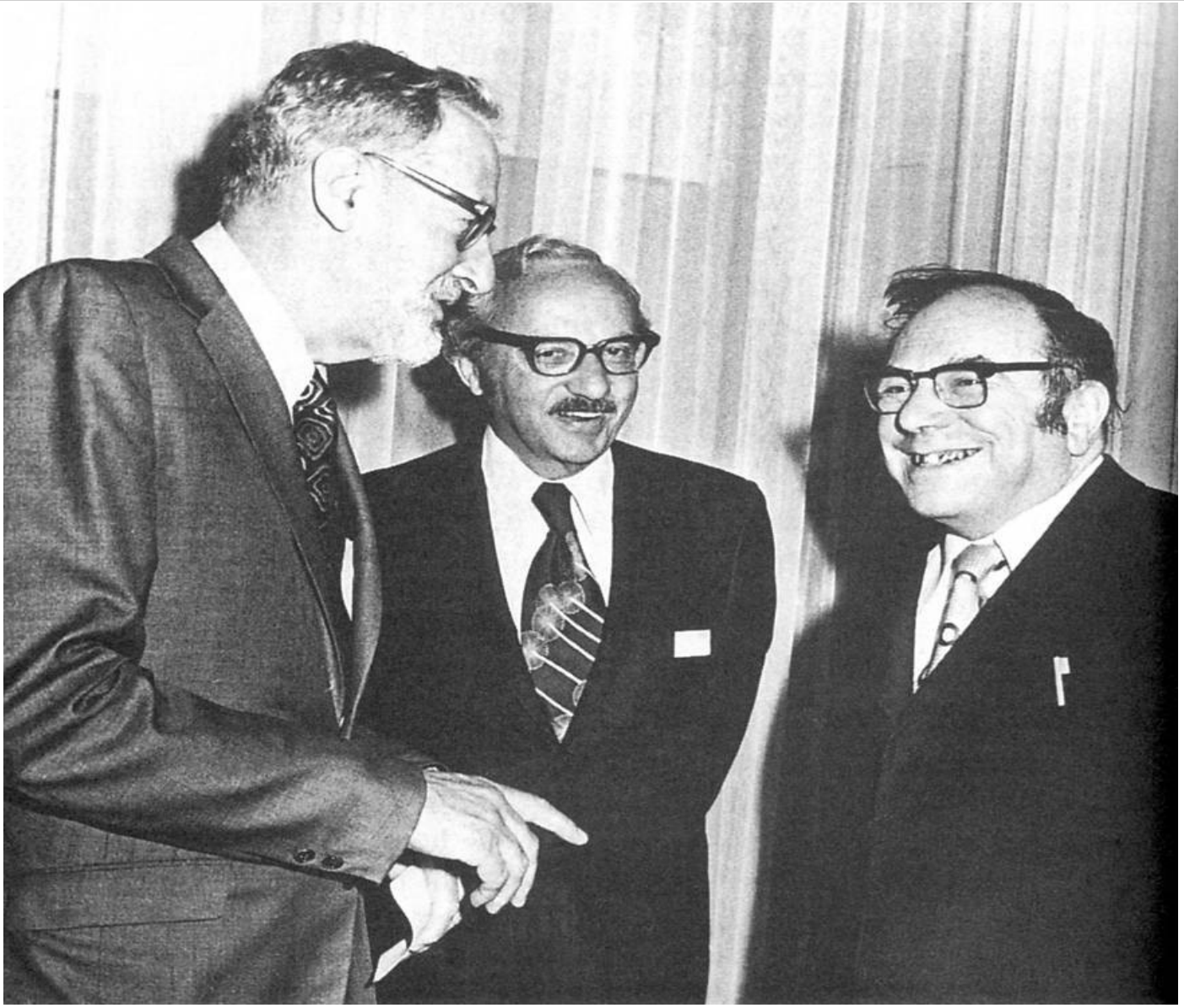
Yinyu Ye
www.stanford.edu/ yyye

K.T. Li Professor of Engineering
Stanford University
and
International Center of Management Science and Engineering
Nanjing University

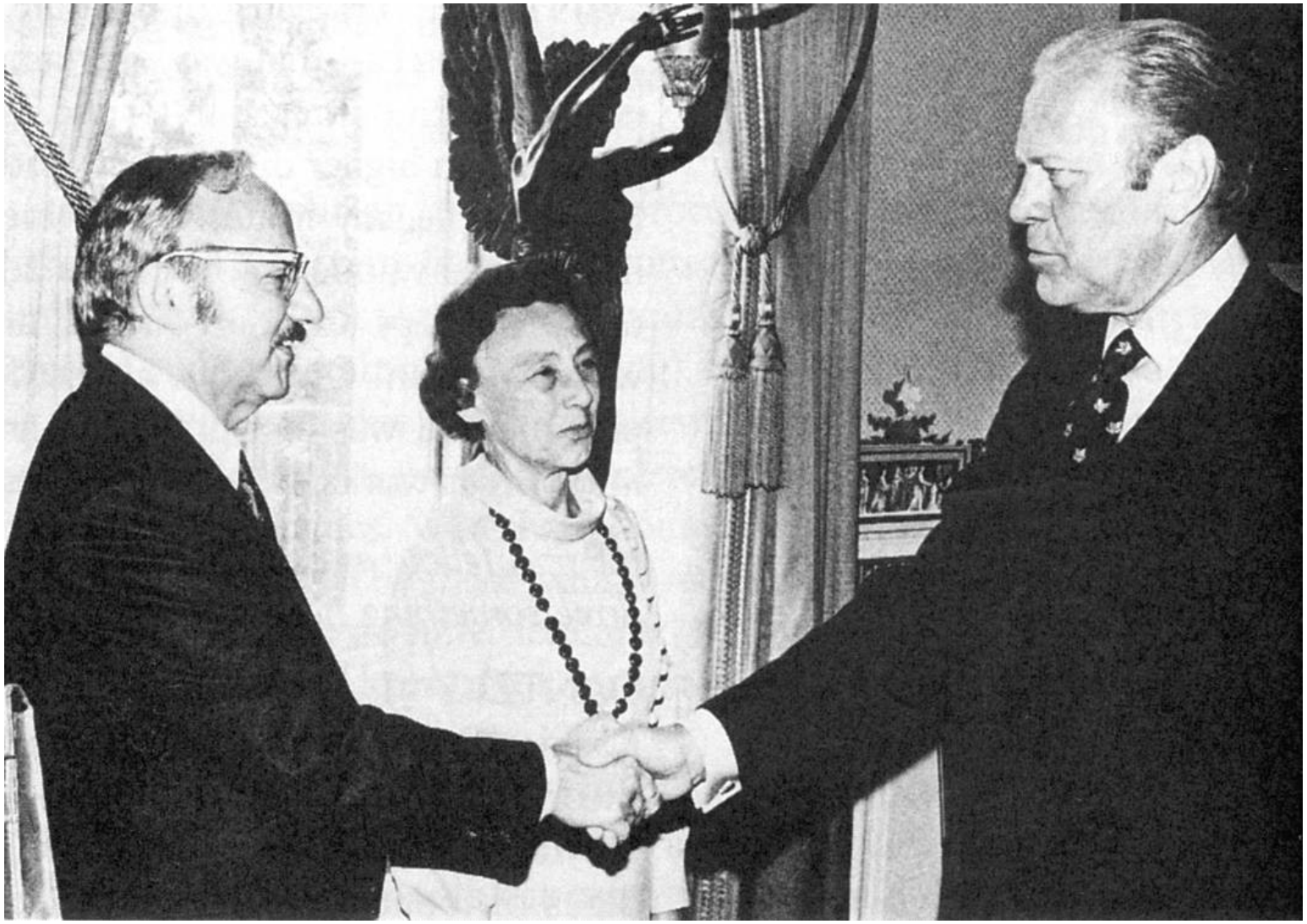
Outlines

- Linear Programming (LP) and the Simplex Method
- Markov Decision Process (MDP) and its LP Formulation
- Simplex and policy-iteration methods for MDP and Zero-Sum Game with fixed discounts
- Simplex method for general non-degenerate LP (including the unbounded case)
- Open Problems

Linear Programming started…



… with the simplex method



LP Model in Dimension *d*

$$\max \quad c_1 x_1 + c_2 x_2 + \ldots + c_d x_d$$

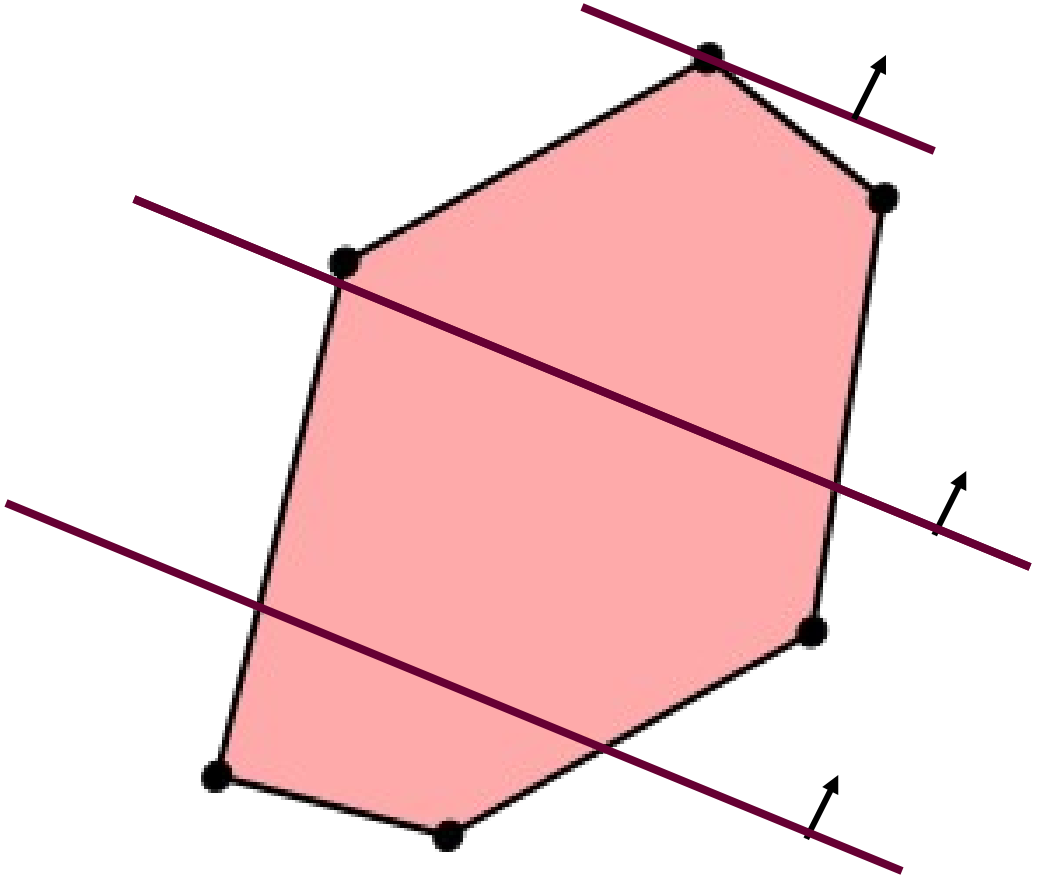
s.t.

$$\begin{aligned}
a_{11} x_1 + a_{12} x_2 + \ldots + a_{1d} x_d &\le b_1 \\
a_{21} x_1 + a_{22} x_2 + \ldots + a_{2d} x_d &\le b_2 \\
\ldots \qquad & \ldots \\
a_{n1} x_1 + a_{n2} x_2 + \ldots + a_{nd} x_d &\le b_n
\end{aligned}$$

The feasible region is a polyhedron defined by *n* inequalities in *d* dimensions.

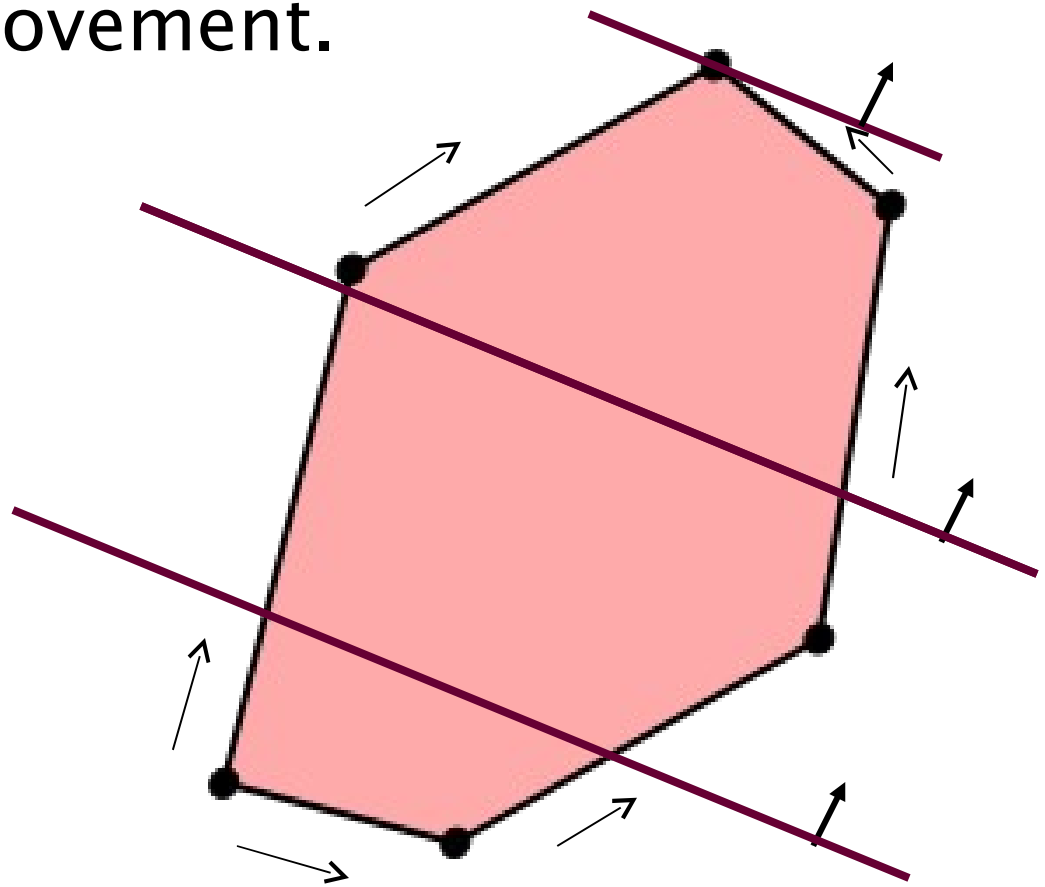
LP Geometry and Theorems

- Optimize a linear objective function over a convex polyhedron, and there is always a vertex optimal solution.



The Simplex Method

- Start with any vertex, and move to an adjacent vertex with an improved objective value. Continue this process till no improvement.



Pivoting rules …

- The simplex method is governed by a pivot rule, i.e. a method of choosing adjacent vertices with a better objective function value.
- Dantzig's original greedy pivot rule.
- The lowest index pivot rule.
- The random edge pivot rule chooses, from among all improving pivoting steps (or edges) from the current basic feasible solution (or vertex), one uniformly at random.

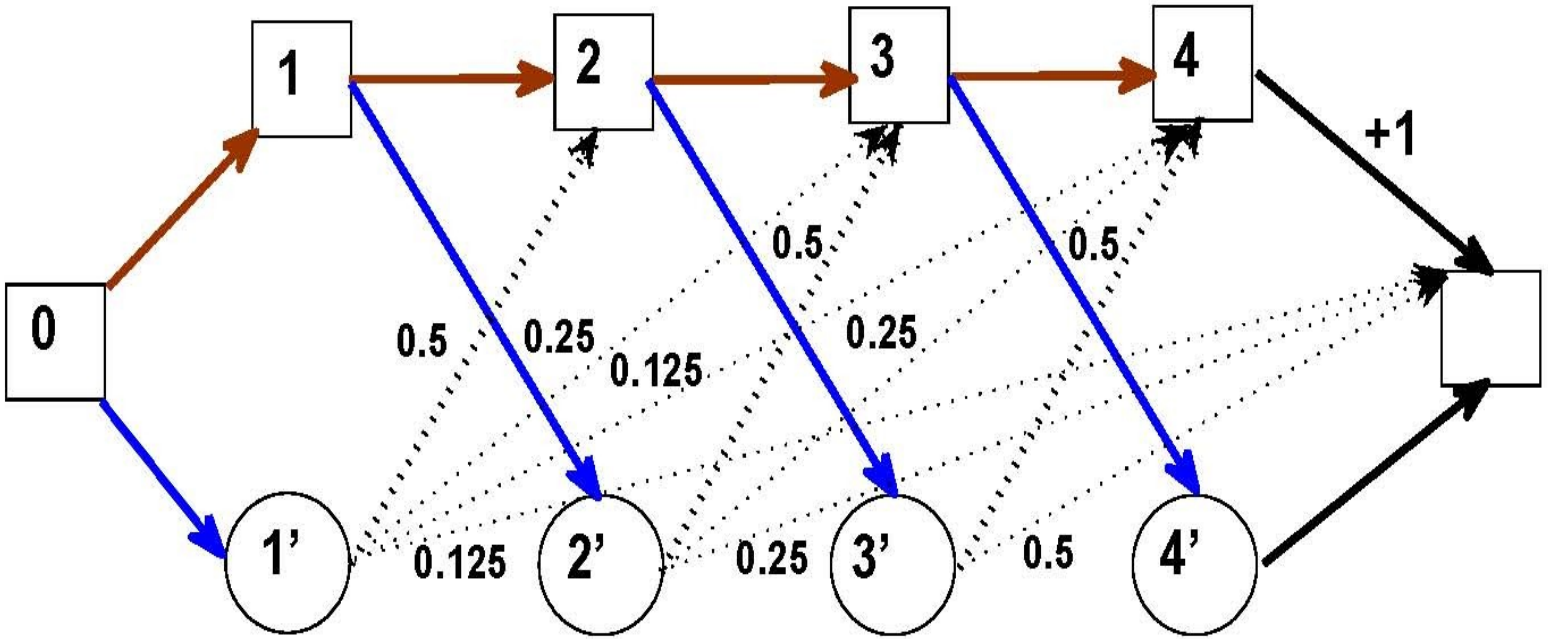
Markov Decision Process

- Markov decision process provides 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. Shapley 1953, Bellman 1957).
- Modern applications include dynamic planning, reinforcement learning, social networking, and almost all other dynamic/sequential decision making problems in Mathematical, Physical, Management, Economics, and Social Sciences.

States and Actions

- At each time step, the process is in some state $i = 1, \ldots, m$, and the decision maker chooses an action $j \in A_i$ that is available for state i, say of total n actions.
- The process responds at the next time step by randomly moving into a new state i', and giving the decision maker an immediate corresponding cost c_j.
- The probability that the process enters i' as its new state is influenced by the chosen action j. Specifically, it is given by the state transition probability distribution P_j.
- But given action j, the probability is conditionally independent of all previous states and actions; in other words, the state transitions of an MDP possess the Markov property.

A Simple MDP Problem I

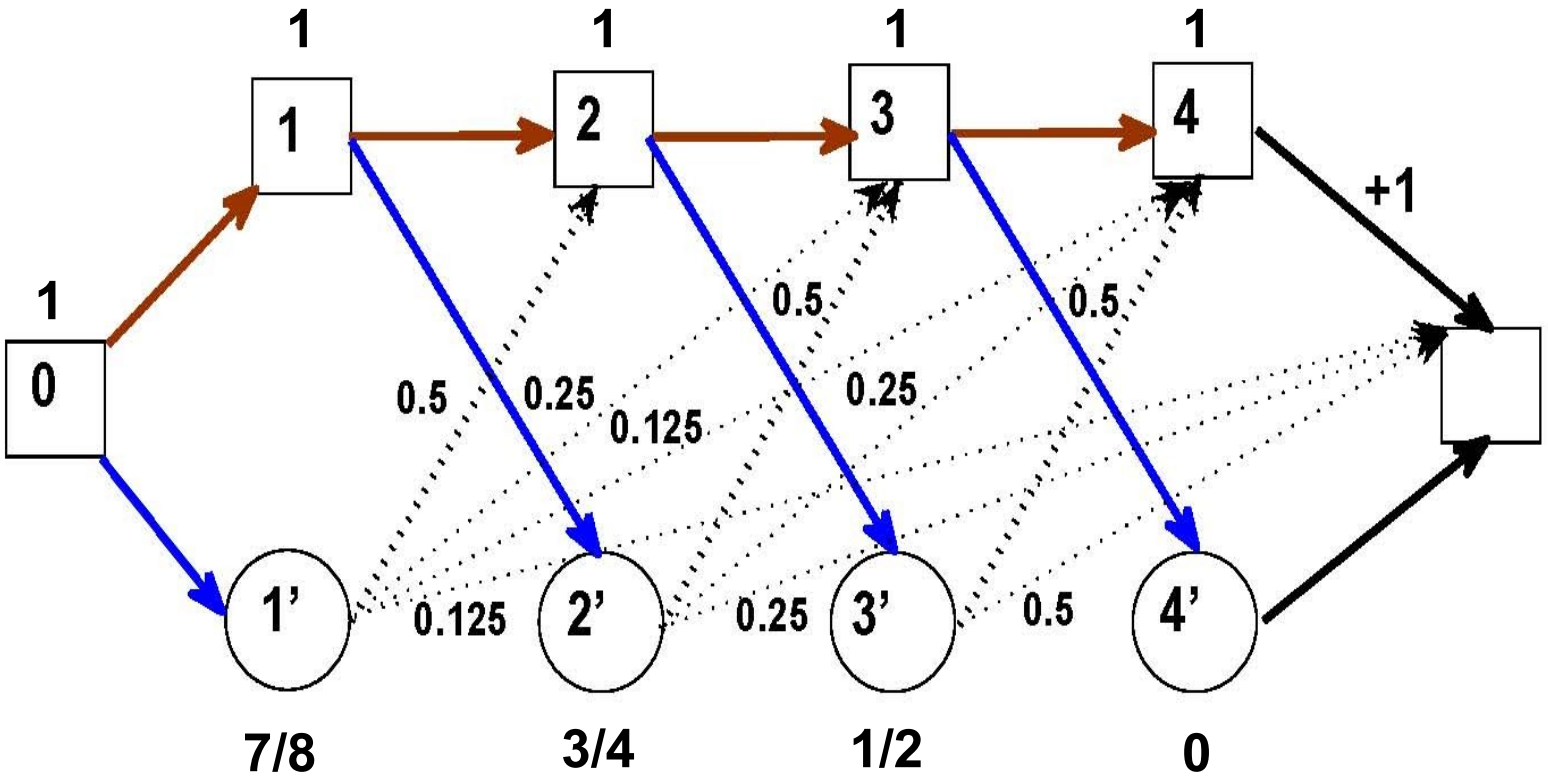


Policy and Discount Factor

- A policy of MDP is a set function $\pi = \{j_1, j_2, \cdots, j_m\}$ that specifies one action $j_i \in A_i$ that the decision maker will choose for each state i.
- The MDP is to find an optimal (stationary) policy to minimize the expected discounted sum over an infinite horizon with a discount factor $0 \leq \gamma < 1$.
- One can obtain an LP that models the MDP problem in such a way that there is a one-to-one correspondence between policies of the MDP and basic feasible solutions of the (dual) LP, and between improving switches and improving pivots.

  de Ghellinck (1960), D'Epenoux (1960) and Manne (1960)

Cost-to-Go-Values



Chosen actions in Red

Cost-to-Go values and LP formulation

- Let $y \in R^m$ represent the expected present cost-to-go values of the m states, respectively, for a given policy. Then, the cost-to-go vector of the optimal policy is a Fixed Point of

$$y_i = \min\{ c_j + \gamma\ p_j^T y, j \in\ A_i\}, \forall\ i,$$

$$j_i = \arg\min\{ c_j + \gamma\ p_j^T y, j \in\ A_i\}, \forall\ i.$$

- Such a fixed point computation can be formulated as an LP

$$\max \quad \sum_{i=1}^{m} y_i$$

$$\text{s.t.} \quad y_i \le\ c_j + \gamma\ p_j^T y, \forall\ j \in\ A_i; \forall\ i.$$

The dual of the MDP-LP

$$\min \quad \sum_{i=1}^{n} c_j x_j$$

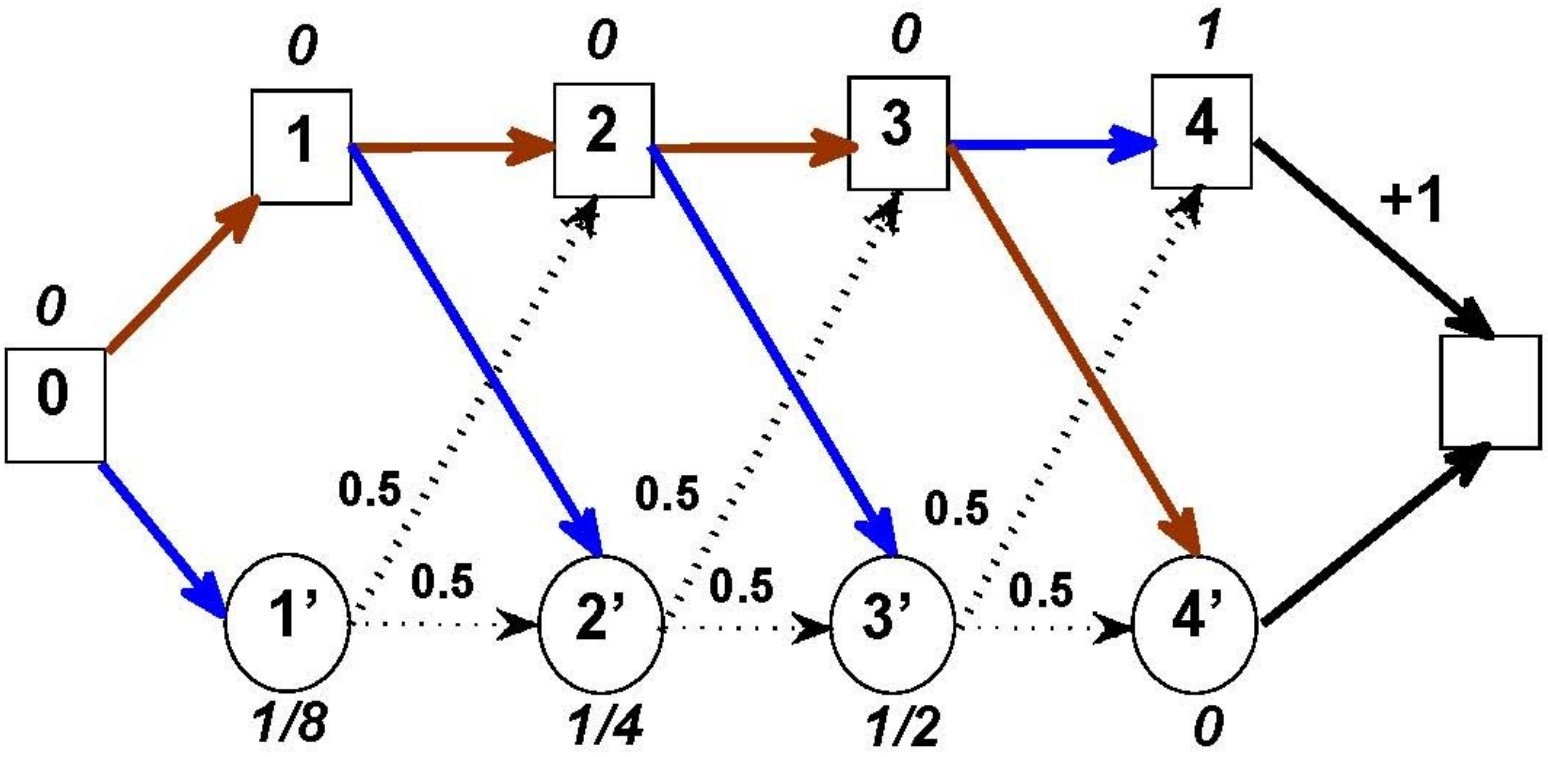
$$\text{s.t.} \quad \sum_{j=1}^{n} (e_{ij} - \gamma\ p_{ij}) x_j = 1, \forall\ i,$$

$$x_j \geq 0, \forall\ j.$$

where $e_{ij} = 1$ if $j \in A_i$ and 0 otherwise.

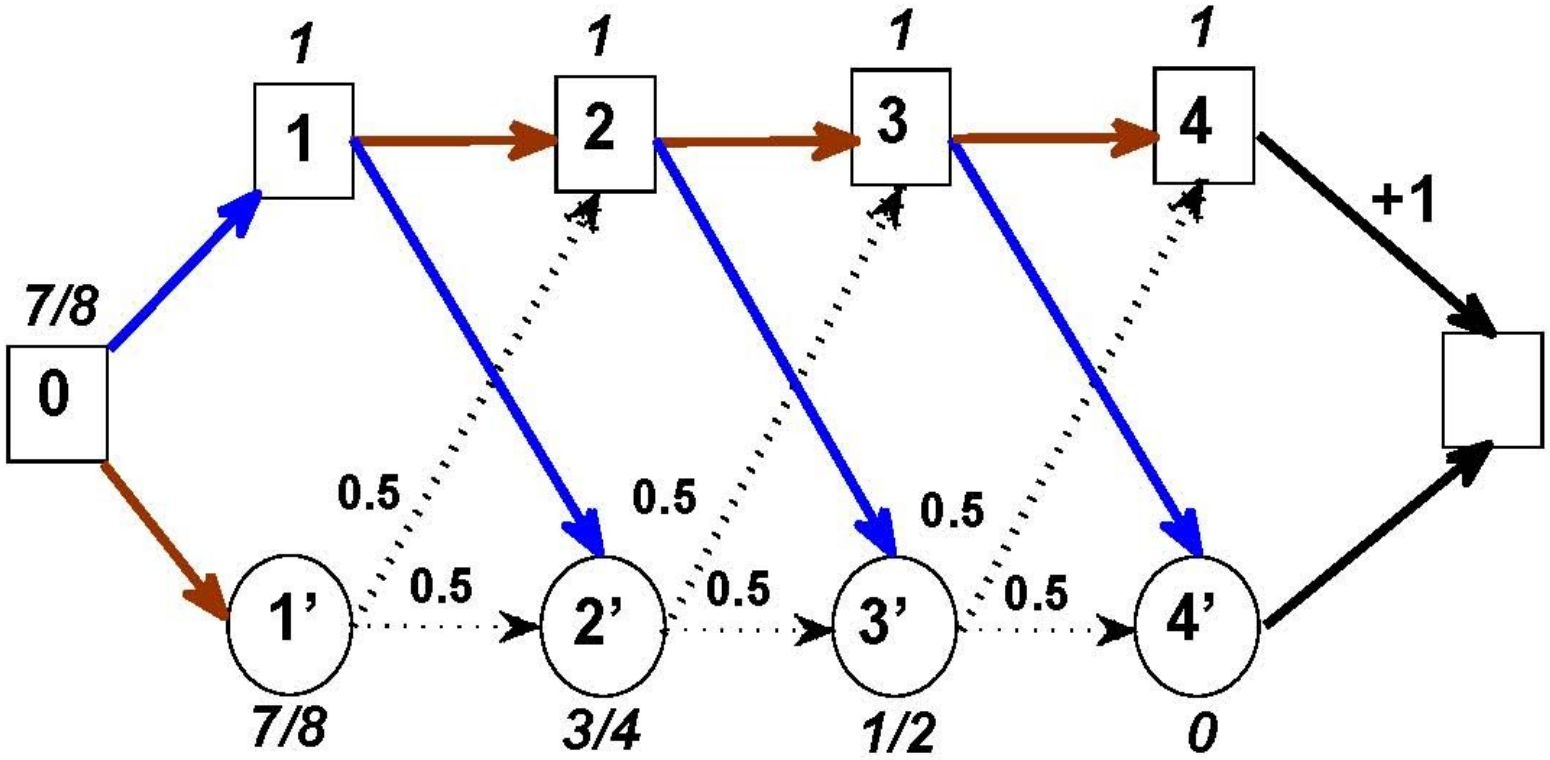
Dual variable x_j represents the expected action flow or visit-frequency, that is, the expected present value of the number of times action j is used.

Greedy Simplex Rule



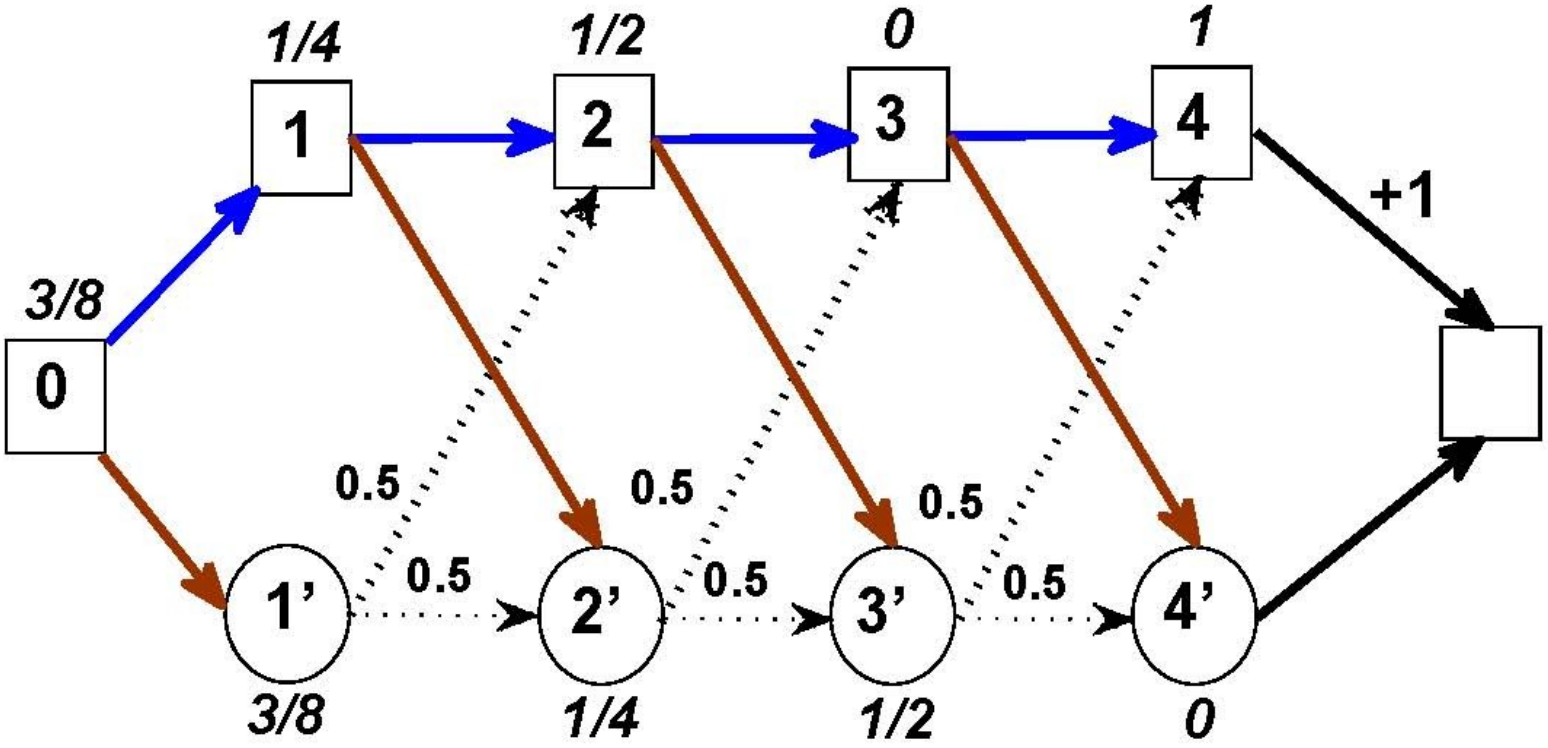
Chosen actions in Red

Lowest-Index Simplex Rule



Chosen actions in Red

Policy Iteration Rule (Howard 1960)



Chosen actions in Red

Exponentially bad examples

- Klee and Minty (1972) showed that Dantzig's original greedy pivot rule may require exponentially many steps for a LP example.
- Melekopoglou and Condon (1990) showed that the simplex method with the smallest index pivot rule needs an exponential number of iterations for a MDP example regardless of discount factors.
- Fearnley (2010) showed that the policy-iteration method needs an exponential number of iterations for a undiscounted finite-horizon MDP example.
- Friedmann, Hansen and Zwick (2011) gave an undiscounted MDP example that the random edge pivot rule needs sub-exponentially many steps.

Any Good News?

- In practice, the policy-iteration method, including the simplex method with greedy pivot rule, has been remarkably successful and shown to be most effective and widely used.

- Any good news in theory?

Bound on the simplex/policy methods

- Y (2011): The classic simplex and policy iteration methods, with the greedy pivoting rule, terminate in no more than

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

  pivot steps, where n is the total number of actions in an m-state MDP with discount factor γ.
- This is a strongly polynomial-time upper bound when γ is bounded above by a constant less than one.

Roadmap of proof

- Define a combinatorial event that cannot repeats more than *n* times. More precisely, at any step of the pivot process, there exists a non-optimal action *j* that will never re-enter future policies or bases after

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

  pivot steps
- There are at most *(n – m)* such non-optimal action to eliminate from appearance in any future policies generated by the simplex or policy-iteration method.
- The proof relies on the duality, the reduced-cost vector at the current policy and the optimal reduced-cost vector to provide a lower and upper bound for a non-optimal action when the greedy rule is used.

Improvement and extension

Hansen, Miltersen and Zwick (2011):

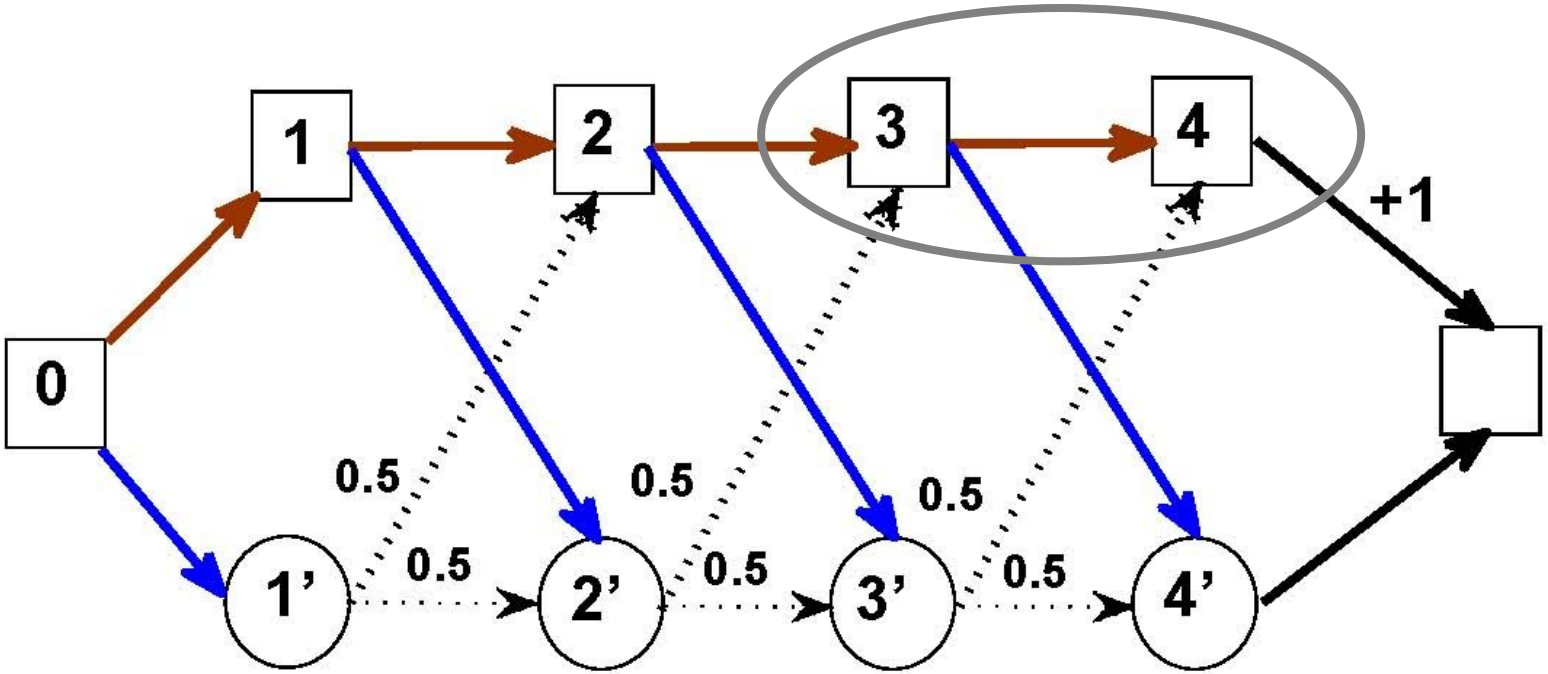
- For the policy iteration method terminates in no more

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

steps.

- The simplex and policy iteration methods, with the greedy pivoting rule, are strongly polynomial-time algorithms for Turn-Based Two-Person Zero-Sum Stochastic Game with any fixed discount factor, which problem cannot even be formulated as an LP.

A Turn-Based Zero-Sum Game



Deterministic MDP with discounts

Distribution vector $p_j \in R^m$ contains exactly one *1* and *0* everywhere else

$$y_i = \min\{ c_j + \gamma_j p_j^T y, j \in A_i \}, \forall\ i,$$

$$j_i = \arg\min\{ c_j + \gamma_j p_j^T y, j \in A_i \}, \forall\ i.$$

$$\max \quad \sum_{i=1}^{m} y_i$$

$$\text{s.t.} \quad y_i \le c_j + \gamma_j p_j^T y, \forall\ j \in A_i; \forall\ i.$$

It has uniform discounts if all γ_j are identical.

The dual resembles a generalized flow

$$\min \quad \sum_{i=1}^{n} c_j x_j$$

$$\text{s.t.} \quad \sum_{j=1}^{n} (e_{ij} - \gamma_j p_{ij}) x_j = 1, \forall\, i,$$

$$x_j \geq 0, \forall\, j.$$

where $e_{ij} = 1$ if $j \in A_i$ and 0 otherwise.

Dual variable x_j represents the expected action flow or frequency, that is, the expected present value of the number of times action j is chosen.

Efficiency of simplex/policy methods

- They are not known to be polynomial-time algorithms for deterministic MDP even with uniform discounts.
- There are quadratic lower bounds on these methods for solving MDP with uniform discounts.
- Ian Post and Y (2012): The Simplex method with the greedy pivot rule terminates in at most

$$0(m^3 n^2 \log^2 m)$$

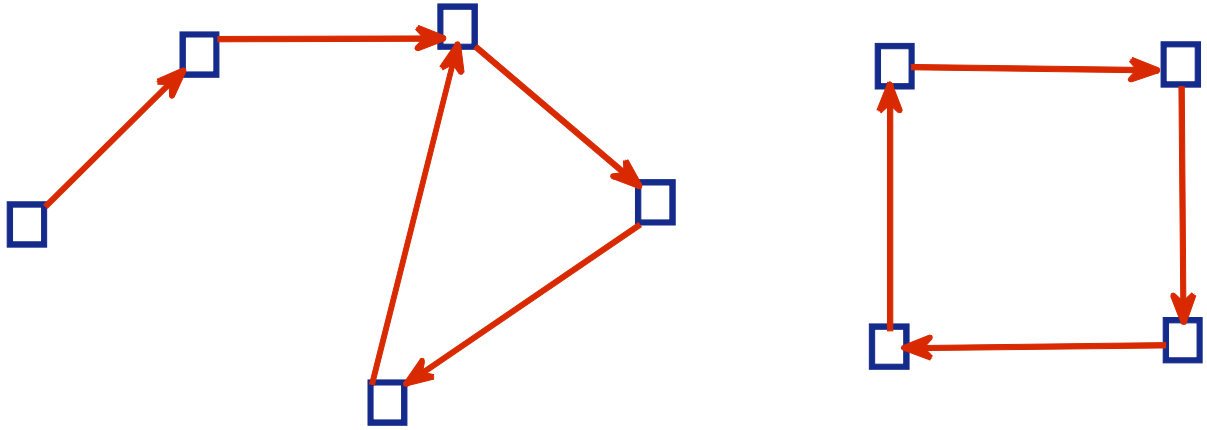
pivot steps when discount factors are uniform, or in at most

$$0(m^5 n^3 \log^2 m)$$

pivot steps with non-uniform discounts.

- Hansen, Miltersen and Zwick (2013) reduced the bound by a factor of *m*.
- Not yet able to prove such results for the policy iteration method.

Policy structures with uniform factors



Each chosen action can be either a path-edge or cycle-edge.

x_j in [1, m] if it is a path-action,
x_j in [1/(1-γ), m/(1-γ)] if it is a cycle-action, so that they form two possible polynomial layers.

Roadmap of proof

- There two types of pivots: the newly chosen action is either on a path or on a cycle of the new policy.
- In every $m^2 n \log(m)$ consecutive pivot steps, there must be at least one step that is a cycle pivot.
- After every $m \log(m)$ cycle pivot steps, there is an action that would never re-enter as a cycle or path action.
- There are at most n action for such a down-grade.
- Item 2 result remains true when discounts are not uniform, but others do not hold.

General non-degenerate LP

- Kitahara and Mizuno (2011) extended the bound to solving general non-degenerate and bounded LPs:

$$\min \quad \sum_{i=1}^{n} c_j x_j$$

$$\text{s.t.} \quad \sum_{j=1}^{n} a_{ij} x_j = b_i, \forall\, i;\ x_j \geq 0, \forall\, j.$$

- The simplex method terminates in at most

$$\frac{mn}{\sigma} \log\left(\frac{m^2}{\sigma}\right)$$

pivot steps, when the ratio of the minimum value over the maximum value, in all basic feasible solution entries, is bounded below by σ.

General non-degenerate LP

- What about for general non-degenerate LPs with possible unboundedness:

$$\min \quad \sum_{i=1}^{n} c_j x_j$$

$$\text{s.t.} \quad \sum_{j=1}^{n} a_{ij} x_j = b_i, \forall\, i; \;\; x_j \geq 0, \forall\, j.$$

- The simplex method terminates in at most

$$\frac{mn}{\sigma} \log\left(\frac{m^2}{\sigma}\right)$$

pivot steps, either finds an optimal basic feasible solution or detects the unboundedness.

Proof sketch I

- Let the objective value of the last basic feasible solution be z*, and consider the "shadow" LP problem

$$\min \quad \sum_{i=1}^{n} c_j x_j$$

$$\text{s.t.} \quad \sum_{j=1}^{n} a_{ij} x_j = b_i, \forall\ i;$$

$$\sum_{i=1}^{n} c_j x_j - x_{n+1} = z^*, x_j \geq 0, \forall\ j.$$

- Obviously, the shadow LP is bounded with a minimal value z*.

Proof sketch II

$$\min \quad \sum_{i=1}^{n} c_j x_j$$

$$\text{s.t.} \quad \sum_{j=1}^{n} a_{ij} x_j = b_i, \forall\, i;\ x_j \geq 0, \forall\, j.$$

$$\min \quad \sum_{i=1}^{n} c_j x_j$$

$$\text{s.t.} \quad \sum_{j=1}^{n} a_{ij} x_j = b_i, \forall\, i;$$

$$\sum_{i=1}^{n} c_j x_j - x_{n+1} = z^*, x_j \geq 0, \forall\, j.$$

- The simplex method with the greedy pivoting rule, applied to the original LP, would generate the identical solution and reduced cost sequence as it is applied to the "shadow" LP in which X_{n+1} remains a basic variable before detects unboundedness.
- In the shadow LP, the basic variable values (excluding X_{n+1}) satisfy the σ property.
- In at most $\frac{mn}{\sigma}\log\left(\frac{m^2}{\sigma}\right)$ pivoting steps, the shadow LP find the optimal basic feasible solution that is the last basic feasible solution of the original LP before detecting unboundedness.

Remarks and Open Problems

- Other pivoting rules?

- Is the policy iteration method a strongly polynomial time algorithm for deterministic MDP?

- Is there strongly polynomial time algorithm for MDP with variable discounts or even general LP?

- Solve LPs with a huge size (billion-dimension) in practice?

The Simplex Method Story Continues …