Policy Iteration: Computational Complexity and Relation to the Simplex Algorithm

COMP 767

Matthew Smith & Pascale Gourdeau

Summary of: *The Simplex and Policy-Iteration Methods are Strongly Polynomial for the Markov Decision Problem with a Fixed Discount Rate* by Yinyu Ye

January 27th, 2017

Overview

Introduction

- Problem Setting
- Main Results

Background

- Linear Programming
- LP Representation of MDP

Results

Notation and Problem Setting

- The MDP has m states and a discount factor γ.

Notation and Problem Setting

- The MDP has m states and a discount factor γ.
- For each transition $s \xrightarrow{a} s'$, there is an associated *cost*, instead of a reward, and so the goal is to *minimize* the expected discounted cost.

Notation and Problem Setting

- The MDP has m states and a discount factor γ.
- For each transition $s \xrightarrow{a} s'$, there is an associated *cost*, instead of a reward, and so the goal is to *minimize* the expected discounted cost.
- We assume that at each state s there is a specific set of actions $\mathcal{A}_s$ that can be performed, with $|\mathcal{A}_s| = k_s$ and

$$n := \sum_{s=1}^{m} k_s$$

.

Notation and Problem Setting

- The MDP has m states and a discount factor γ.
- For each transition $s \xrightarrow{a} s'$, there is an associated *cost*, instead of a reward, and so the goal is to *minimize* the expected discounted cost.
- We assume that at each state s there is a specific set of actions $\mathcal{A}_s$ that can be performed, with $|\mathcal{A}_s| = k_s$ and

  $$n := \sum_{s=1}^{m} k_s$$

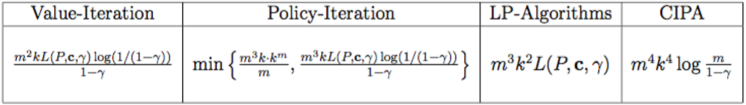
  .
- We assume that the transition probabilities, represented as a matrix $\mathbf{P}$, and the cost function $c(s', a|s)$ are both known.

Polynomial vs Strongly Polynomial

- *Polynomial:* The number of arithmetic operations needed to compute an (optimal) solution is bounded by a polynomial function in the number of input data and their bit-sizes.
- *Strongly Polynomial:* The number of arithmetic operations needed to compute an (optimal) solution is bounded by a polynomial function in the number of input data only.
- In our case, finding the optimal policy is done in $poly(n, m)$ time.

Previous results

Given that there are exactly k actions in each state, and $L(\mathbf{P}, \mathbf{c}, \gamma)$ is the bit-size of the MDP input data to the LP,



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

Main Results

- When framing the discounted MDP problem of finding an an optimal policy π^* as an LP, the Simplex method that solves the LP is exactly policy iteration.
- The Simplex/policy iteration method is polynomial in the number of states and actions:

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

Linear Programming Refresher

- We seek to optimize some linear function subject to linear constraints.
- This can be written as:

$$
\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{n} c_i x_i \\
\text{subject to} \quad & \sum_{i=1}^{n} a_{ji} x_i = b_j \forall i \in 1, ..., n \text{ and } j \in 1, ...m \\
& x_i \geq 0 \qquad\qquad \forall i \in 1, ..., n
\end{aligned}
$$

- or equivalently, in vector form:

$$
\begin{aligned}
\text{minimize} \quad & \mathbf{c}^\top \mathbf{x} \\
\text{subject to} \quad & \mathbf{A}\mathbf{x} = b \\
& \mathbf{x} \geq 0
\end{aligned}
$$

Important Ideas In LP

Feasible Solution:

An allocation of values $\mathbf{x}$, such that all constraints are satisfied.

Basic Feasible Solution:

A feasible solution, such that there are only $rank(\mathbf{A})$ nonzero values of $\mathbf{x}$. These correspond to extreme points in the convex set defined by the constraints.

Dual Form:

Using Lagrange method, rewrite optimization as a function of constraints. Using notation from earlier, this gives us:

$$\begin{array}{ll} \text{maximize} & -\mathbf{b}^\top \mathbf{y} \\ \text{subject to} & \mathbf{A}^\top \mathbf{y} + \mathbf{c} \geq \mathbf{0} \end{array}$$

The Simplex Method

- Every optimal solution to the LP corresponds to a BFS
- Starting from a BFS, we note that the columns of $\mathbf{A}$ associated with nonzero values of $\mathbf{x}$ form a basis for $\mathbb{R}^m$
- We express some other column of $\mathbf{A}$, $\mathbf{a}_\nu$ as a linear combination of these basis vectors, and use the coefficients, $\mathbf{y}_\nu$ to compute:
$$\bar{c}_\nu = c_\nu - \mathbf{c}^\top \mathbf{y}_\nu$$
- And replace some existing vector of the basis with the vector $\mathbf{a}_f$ associated with $f = \operatorname{argmin}_i \bar{\mathbf{c}}$ if $\bar{c}_f < 0$
- We do this by choosing
$$\theta = \min . \frac{x_i}{y_{if}}$$
and modifying all variables according to
$$x_i^{t+1} = x_i^t - \theta y_{if}$$

Linear Program Representation of the MDP

- Primal form:

$$
\begin{array}{ll}
\text{minimize} & \mathbf{c}^\top \mathbf{x} \\
\text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b} \\
 & \mathbf{x} \geq \mathbf{0}
\end{array}
$$

Linear Program Representation of the MDP

- Primal form:

$$\begin{array}{ll} \text{minimize} & \mathbf{c}^\top \mathbf{x} \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{array}$$

- Dual form:

$$\begin{array}{ll} \text{maximize} & \mathbf{b}^\top \mathbf{y} \\ \text{subject to} & \mathbf{s} = \mathbf{c} - \mathbf{A}^\top \mathbf{y} \geq \mathbf{0} \end{array}$$

Linear Program Representation of the MDP

- Primal form:

$$\begin{aligned} \text{minimize} \quad & \mathbf{c}^\top \mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{aligned}$$

Linear Program Representation of the MDP

- Primal form:

$$\begin{aligned} \text{minimize} \quad & \mathbf{c}^\top \mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{aligned}$$

- $\mathbf{b} = \mathbf{1} \in \mathbb{R}^m$

Linear Program Representation of the MDP

- Primal form:

$$\begin{array}{ll} \text{minimize} & \mathbf{c}^\top \mathbf{x} \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{array}$$

- $\mathbf{b} = \mathbf{1} \in \mathbb{R}^m$
- $\mathbf{c} \in \mathbb{R}^n$ such that if $a \in \mathcal{A}_s$ then $c_a = \sum_{s'} p(s'|s,a)c(s'|s,a)$

Linear Program Representation of the MDP

- Primal form:

$$\begin{array}{ll} \text{minimize} & \mathbf{c}^\top \mathbf{x} \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{array}$$

- $\mathbf{b} = \mathbf{1} \in \mathbb{R}^m$
- $\mathbf{c} \in \mathbb{R}^n$ such that if $a \in \mathcal{A}_s$ then $c_a = \sum_{s'} p(s'|s,a)c(s'|s,a)$
- $\mathbf{A} = \mathbf{E} - \gamma \mathbf{P} \in \mathbb{R}^{m \times n}$ with rank m, where
  - $P_{s',a} = p(s'|s,a)$ for $s' = 1, \dots, m$.
  - $E_{s,a} = \begin{cases} 1 & \text{if } a \in \mathcal{A}_s \\ 0 & \text{otherwise} \end{cases}$

Lemma 1

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

Lemma 1

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

   Proof: We need the policy to both: 1) specify an action in every state, and 2) each state must have exactly one action. Let π index the basis set of a basic feasible solution for the primal problem $|\pi| = m$.
   By the constraint, the sth row of $\mathbf{A}$ satisfies

   $$1 = \sum_{d \in \pi} a_{sm} x_d^{\pi}$$

   but for all actions $d \notin \mathcal{A}_s$, $a_{sm} \leq 0$, so if there is some state s such that $\forall d \in \pi : d \notin \mathcal{A}_s$, we have

   $$\sum_{d \in \pi, d \notin \mathcal{A}_s} a_{sm} x_d^{\pi} \leq 0$$

   which must violate one of the constraints.

Lemma 1

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

   Proof (contd.):
   If we have an action in every state (as just shown), then since $|\pi| = m$, each state specifies only one action.

Lemma 1

1. *There is a one-to-one correspondence between a (stationary) policy of the original discounted MDP and a basic feasible solution of the discounted MDP primal.*
2. *Let $\mathbf{x}^{\pi}$ be a basic feasible solution of the discounted MDP primal. Then any basic variables, say $\mathbf{x}^{\pi}_{s}$, has its value*

$$1 \leq \mathbf{x}^{\pi}_{s} \leq \frac{m}{1-\gamma} \quad .$$

Lemma 1

1. *There is a one-to-one correspondence between a (stationary) policy of the original discounted MDP and a basic feasible solution of the discounted MDP primal.*
2. *Let $\mathbf{x}^{\pi}$ be a basic feasible solution of the discounted MDP primal. Then any basic variables, say $\mathbf{x}^{\pi}_{s}$, has its value*

$$1 \leq \mathbf{x}^{\pi}_{s} \leq \frac{m}{1-\gamma} \quad .$$

3. *The feasible set of the discounted MDP primal is bounded. More precisely, for every feasible $\mathbf{x} \geq 0$,*

$$\mathbf{e}^{\top}\mathbf{x} = \frac{m}{1-\gamma} \quad .$$

Lemma 2

Let both the primal and dual LPs be feasible. Then the is a unique partition $\{\mathcal{P}, \mathcal{O}\}$ of the actions [1] *such that for all optimal solution pairs $(\mathbf{x}^*, \mathbf{s}^*)$,*

$$x_a^* = 0 \; \forall j \in \mathcal{O}, \text{ and } s_{a'}^* = 0 \; \forall a' \in \mathcal{P},$$

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

$$x_a^*, \; \forall j \in \mathcal{P}, \text{ and } s_{a'}^* > 0 \; \forall a' \in \mathcal{O},$$

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

[1] $\mathcal{P} \subseteq \{1, 2 \ldots, n\}$ and $\mathcal{O} \subseteq \{1, 2, \ldots, n\}$, with $\mathcal{P} \cup \mathcal{O} = \{1, 2, \ldots, n\}$ and$\mathcal{P} \cap \mathcal{O} = \emptyset$

Simplex as Policy Iteration

We reframe the primal as the equivalent LP:

$$\begin{array}{ll} \text{minimize} & \bar{\mathbf{c}}_\nu^\top \mathbf{x}_\nu + \bar{\mathbf{c}}_\pi^\top (A_\pi)^{-1}\mathbf{e} \\ \text{subject to} & \mathbf{A}_\pi \mathbf{x}_\pi + \mathbf{A}_\nu \mathbf{x}_\nu = \mathbf{e} \\ & \mathbf{x} \geq \mathbf{0} \end{array}$$

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

$$\bar{\mathbf{c}}_\pi = \mathbf{0} \qquad \bar{\mathbf{c}}_\nu = \mathbf{c}_\nu - \mathbf{A}_\nu^\top \mathbf{y}_\pi \qquad \mathbf{y}_\pi = (\mathbf{A}_\pi^\top)^{-1}\mathbf{c}_\pi$$

If $\bar{\mathbf{c}} \geq \mathbf{0}$, we are at the optimal policy. Otherwise there is some minimum reduced cost variable that is negative, meaning that under the current policy, some other action in a specific state leads to lower future discounted cost, which suggests this action should be taken instead. We substitute the action corresponding to the minimum reduced cost and substitute it for the existing one at that state in the basis $\mathbf{A}_\pi$. This corresponds with the minimum reduced cost pivot rule for the simplex algorithm, or one iteration of minimum reduced cost policy improvement.

Lemma 3

Let z^* be the optimal objective value of the primal. Then, in any iteration of the Simplex method from current policy $\pi$to new policy π^+

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

where $\Delta := -\min \bar{c}_\nu$. Moreover,

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

Therefore the simplex method generates a sequence of policies, indexed by t, such that

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

Proof of Lemma 3

We observe that:

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

which shows the first inequality of the lemma for all $\mathbf{x}$, including the optimal solution.
Further, since the value of the new basic variable is greater than or equal to 1, the simplex objective is changed by at least Δ, which gives us

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

solving for Δ in the first inequality and substituting here gives the second. The third follows by induction.

Lemma 4

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

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

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

Lemma 4

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

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

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

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

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

Proof of Lemma 4.1

Since all non-basic variables of $\mathbf{x}^{\pi}$ have zero values:

$$\begin{aligned}
\mathbf{c}^{\top}\mathbf{x} - z^{*} &= \mathbf{c}^{\top}\mathbf{x} - \mathbf{c}^{\top}\mathbf{x}^{*} && \text{(by definition)} \\
&= \mathbf{c}^{\top}\mathbf{x} - \mathbf{e}^{\top}\mathbf{y}^{*} && \text{(strong duality)} \\
&= (\mathbf{s}^{*})^{\top}\mathbf{x}^{\pi} && \text{(dual definition)} \\
&= \sum_{j \in \pi} s_j^{*} x_j^{\pi}
\end{aligned}$$

Now, there must be action $j \in \pi$ s.t.

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

Proof of Lemma 4.1 Cont'd

(Lemma 1 $\implies$) $1 \leq x_j^\pi \leq \frac{m}{1-\gamma}$. Then

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

and $j \in \mathcal{O}$ from Lemma 2.

Proof of Lemma 4.2

Suppose π_0 is non-optimal. Then, using Lemma 4.1, let $j^0 \in \pi^0 \cap \mathcal{O}$ be the state-action pair index such that

$$s^*_{j_0} \geq \frac{1-\gamma}{m^2} \left(\mathbf{c}^\top \mathbf{x}^{\pi^0} - z^* \right) \quad . \ .$$

If $j^0 \in \pi^t$ for some $t \geq 1$, where π^t is generated by the Simplex method,

$$\mathbf{c}^\top \mathbf{x}^{\pi^t} - z^* = (\mathbf{s}^*)^\top \mathbf{x}^{\pi^t} \geq s^*_{j^0} x^{\pi^t}_{j_0} \quad .$$

Then,

$$x^{\pi^t}_{j_0} \leq \frac{\mathbf{c}^\top \mathbf{x}^{\pi^t} - z^*}{s^*_{j^0}} \leq \frac{m^2}{1-\gamma} \cdot \frac{\mathbf{c}^\top \mathbf{x}^{\pi^t} - z^*}{\mathbf{c}^\top \mathbf{x}^{\pi^0} - z^*} \quad ,$$

as required.

Theorem 1

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

Proof of Theorem 1

(Lemma 3 $\implies$) after t iterations of the simplex method:

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

So, after t iterations from the initial policy π^0 with $j^0 \in \pi^t$,

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

Now, using the identity $\log(1-y) \leq -y$ for $y < 1$ and letting $t > T$, we get that the RHS of $(\dagger)$ is < 1, implying that $x_{j^0}^{\pi^t} < 1$ and so contradicting Lemma 1. Then $j^0 \notin \pi^t$ for all $t > T$.

Theorem 2

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

Proof of Theorem 2

- The state-action j^0 from theorem 1 will never be part of a policy after time T.
- Performing iterations, if π^{T+1} is not optimal, there exists $j^1 \in \pi^{T+1} \cap \mathcal{O}$ with $j^1 \neq j^0$ s.t. $j^1 \notin \pi^t$ for all $t > 2T$, etc.
- For each Simplex iteration, we get a better policy.
- There are at most $|\mathcal{O}| \leq n - m$ such iterations by Lemma 2.
- This gives running time $\leq \frac{m(n-m)}{1-\gamma} \cdot \log \left(\frac{m^2}{1-\gamma} \right)$.

Corollary 1

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

Proof of Corollary 1

- Lemma 1 and 2 are independent of the method being used.

Proof of Corollary 1

- Lemma 1 and 2 are independent of the method being used.
- Given policy π, the incoming Simplex basic variable $j^{+} = \arg\min(\mathbf{c})$ is always an incoming basic variable for policy-iteration.

Proof of Corollary 1

- Lemma 1 and 2 are independent of the method being used.
- Given policy π, the incoming Simplex basic variable $j^+ = \arg\min(\mathbf{c})$ is always an incoming basic variable for policy-iteration.
- Lemma 4 holds for policy iteration as well: its consequences hold as long as the state-action with the most-negative-reduced-cost is in the next policy.

Proof of Corollary 1

- Lemma 1 and 2 are independent of the method being used.
- Given policy π, the incoming Simplex basic variable $j^{+} = \arg\min(\mathbf{c})$ is always an incoming basic variable for policy-iteration.
- Lemma 4 holds for policy iteration as well: its consequences hold as long as the state-action with the most-negative-reduced-cost is in the next policy.
- Therefore there always exists an action in a non-optimal policy that will never appear after T iterations when doing policy-iteration.

$\implies$ Theorem 1 holds, and the proof is complete.