Strongly Polynomial Algorithms for Transient and Average-Cost MDPs

Jefferson Huang

School of Operations Research and Information Engineering
Cornell University

June 5, 2017

Workshop on MAthematical performance Modeling and Analysis (MAMA)

University of Illinois at Urbana-Champaign

Urbana, IL

Joint work with Eugene A. Feinberg (Stony Brook University)

Overview

Markov decision processes (MDPs): model of sequential decision-making under uncertainty

- Boucherie & van Dijk (2017): applications to healthcare, transportation, production systems, communications, finance

Alternative "good" linear programming formulations of certain total-cost and average-cost MDPs.

- Total-cost: should be transient.
- Average-cost: hitting time to a certain state should be bounded uniformly in initial states & policies.
- Conditions under which they are solvable in strongly polynomial time using classic methods.
- Based on recent results on discounted MDPs.

Discrete-Time Markov Decision Process (MDP)

$\mathbb{X}$ = finite state set; $|\mathbb{X}| = n$

$A(x)$ = set of actions available at state x; $\sum_x |A(x)| = m$

$p(y|x,a)$ = probability that the next state is y, given the current state is x and action a is taken

$c(x,a)$ = cost incurred when current state is x and action a is taken

$$\begin{tikzcd}
\text{Initial Distribution} \arrow[dr] & & & & & & & & \\
\text{State} & x_0 \arrow[rr, "p(x_1|x_0,a_0)"] & & x_1 \arrow[rr, "p(x_2|x_1,a_1)"] & & x_2 \arrow[rr, "p(x_3|x_2,a_2)"] & & x_3 \arrow[r] & \cdots \\
& & a_0 \arrow[d] & & a_1 \arrow[d] & & a_2 \arrow[d] & & \\
& & c(x_0,a_0) & & c(x_1,a_1) & & c(x_2,a_2) & &
\end{tikzcd}$$

Policies

Policy = rule determining which action to take at each time step

In this talk: deterministic stationary policies only

- i.e., mappings ϕ on $\mathbb{X}$ where $\phi(x) \in A(x)$ for all $x \in \mathbb{X}$
- no loss of generality (wrt. randomized history dependent policies) for models considered

Compare policies via a cost criterion $g(\phi) \in \mathbb{R}^n$

- ϕ_* is optimal if $g(\phi_*) \leqslant g(\phi)$ (component-wise) for all policies ϕ

For each policy ϕ, let

$$P(\phi)_{x,y} := p(y|x, \phi(x)), \quad c(\phi)_x := c(x, \phi(x)).$$

Optimality Criteria

Total-Cost Criterion: For each state x,

$$g(\phi) = v(\phi) := \sum_{n=0}^{\infty} P(\phi)^n c(\phi)$$

Average-Cost Criterion: For each state x,

$$g(\phi) = w(\phi) := \limsup_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} P(\phi)^n c(\phi)$$

Complexity Estimates

An MDP is solved by computing an optimal policy.

An algorithm solves an MDP in strongly polynomial time if the # of arithmetic operations needed can be bounded above by a polynomial in the # of state-action pairs m.

If the # of arithmetic operations needed can be bounded above by a polynomial in m and the total bit-size of the input data, it solves the MDP in weakly polynomial time.

- Total-cost & average-cost MDPs can be formulated as linear programs $\implies$ solvable in weakly polynomial time (Khachiyan, 1979)

Total-Cost MDPs: Transience Assumption

$\|P(\phi)\| := \max_{x \in \mathbb{X}} \sum_{y \in \mathbb{X}} p(y|x, \phi(x))$

- $\sum_y p(y|x,a) < 1 \implies$ positive probability that process ends

Assumption (Transience)

There is a constant K such that, for every policy ϕ,

$$\left\| \sum_{n=0}^{\infty} P(\phi)^n \right\| \leqslant K < \infty.$$

- Lifetime of the process is bounded by K under every policy.

Veinott (1974): Transience can be checked in strongly polynomial time.

A Condition Equivalent to Transience

Theorem (Feinberg & H, 2017)

Transience holds if and only if there is a function $\mu : \mathbb{X} \to [0, K]$ *where*

$$\mu(x) \geqslant 1 + \sum_{y \in \mathbb{X}} p(y|x, a)\mu(y)$$

for all $a \in A(x)$ *and* $x \in \mathbb{X}$.

E.g., let

$$\mu = \max_{\phi} \left\{ \sum_{n=0}^{\infty} P(\phi)^n \mathbf{1} \right\}$$

where $\mathbf{1}_x = 1$ for all $x \in \mathbb{X}$.

Denardo (2016): Such a μ can be computed using at most $O[(n^3 + mn)mK \log K]$ arithmetic operations.

Linear Programming Formulation

$$
\begin{aligned}
\text{minimize} \quad & \sum_{x \in \mathbb{X}} \sum_{a \in A(x)} \frac{c(x,a)}{\mu(x)} z_{x,a} \\
\text{such that} \quad & \sum_{a \in A(x)} z_{x,a} - \sum_{x' \in \mathbb{X}} \sum_{a' \in A(x')} \frac{p(x|x',a')\mu(x)}{\mu(x')} z_{x',a'} = 1, \quad x \in \mathbb{X} \\
& z_{x,a} \geqslant 0, \quad a \in A(x), \ x \in \mathbb{X}
\end{aligned}
$$

For an optimal basic feasible solution z^*,

$$\phi_*(x) = \arg\max_{a \in A(x)} \left\{ z^*_{x,a} \right\}, \quad x \in \mathbb{X}.$$

Theorem (Feinberg & H, 2017)

ϕ_* *is optimal under the total-cost criterion.*

Complexity Estimate

Theorem (Feinberg & H, 2017)

The simplex method with Dantzig's rule solves the linear program (LP) using at most

$$O(nmK \log K) \quad \textit{iterations.}$$

Also, there is a block-pivoting simplex method that solves the LP using at most

$$O(mK \log K) \quad \textit{iterations.}$$

- Each iteration of the simplex method needs $O(n^3 + nm)$ arithmetic operations.
- When K is fixed, these two algorithms solve total-cost MDPs in strongly polynomial time.
- Denardo (2016): similar estimates, using different proof technique

Proof Sketch

The LP and the results about it come from a reduction to a discounted MDP with cost-free absorbing state $\tilde{x} \notin \mathbb{X}$, based on Veinott (1968).

- discount factor $\tilde{\beta} = (K-1)/K$
- scaled transition matrices

$$\tilde{P}(\phi)_{x,y} \begin{cases} \tilde{\beta}^{-1}\mathrm{diag}(\mu^{-1})P(\phi)\mathrm{diag}(\mu)_{x,y}, & x, y \neq \tilde{x} \\ 1 - \sum_{y \neq \tilde{x}} \tilde{P}(\phi)_{x,y} & x \neq \tilde{x},\ y = \tilde{x} \\ 1, & x = y = \tilde{x} \end{cases}$$

  and one-step costs

$$\tilde{c}(\phi)_x = \begin{cases} \mathrm{diag}(\mu^{-1})c(\phi)_x, & x \neq \tilde{x} \\ 0, & x = \tilde{x} \end{cases}$$

- minimize $\tilde{v}(\phi) = \sum_{n=0}^{\infty} \tilde{\beta}^n \tilde{P}(\phi)^n \tilde{c}(\phi)$

Feinberg & Huang (2017): For every policy ϕ, $v(\phi) = \mathrm{diag}(\mu)\tilde{v}(\phi)$.

Use complexity estimates in Scherrer (2016) for discounted MDPs.

Interlude: Discounted MDPs

An optimal policy for a discounted MDP with discount factor $\beta \in (0,1)$ can be computed by solving

$$\begin{aligned}
\text{minimize} \quad & \sum_{x \in \mathbb{X}} \sum_{a \in A(x)} c(x,a) z_{x,a} \\
\text{such that} \quad & \sum_{a \in A(x)} z_{x,a} - \beta \sum_{x' \in \mathbb{X}} \sum_{a' \in A(x')} p(x|x',a') z_{x',a'} = 1, \quad x \in \mathbb{X} \\
& z_{x,a} \geqslant 0, \quad a \in A(x), \ x \in \mathbb{X}
\end{aligned}$$

- z is a basic feasible solution (BFS) $\implies$ for every state x, exactly one $z_{x,a}$ is positive
- z^* is optimal BFS $\implies$ policy $\phi_*(x) = \arg\max_a \{z_{x,a}\}$ is optimal

Interlude: Complexity of Discounted MDPs

Discounted MDPs with a fixed discount factor are solvable in strongly polynomial time.

- Ye (2005): Interior-point method
- Ye (2011), Scherrer (2016): simplex method with Dantzig's rule, Howard's (1960) policy iteration method

Hollanders, Delvenne, Jungers (2012): If discount factor isn't fixed, Howard's (1960) policy iteration may need exponential time.

Discounted MDPs with special structure can be solved in strongly polynomial time (regardless of discount factor).

- Zadorojniy, Even, Shwartz (2009): M/M/1 queue with service rate control
- Post & Ye (2015): deterministic MDPs

Average-Cost MDPs: Hitting Time Assumption

$$
{}_{\ell}P(\phi)_{x,y} = \begin{cases} p(y|x,\phi(x)), & y \neq \ell \\ 0, & y = \ell \end{cases}
$$

Assumption (Hitting Time)

There is a state ℓ and a constant L such that, for every policy ϕ,

$$
\left\| \sum_{n=0}^{\infty} {}_{\ell}P(\phi)^{n} \right\| \leqslant L < \infty.
$$

- Mean recurrence time to state ℓ is bounded by L under every policy.
  - E.g., failed state of machine, no customers in queue
- Every such MDP is unichain.

Feinberg & Yang (2008): can be checked in strongly polynomial time

An Equivalent Condition

Theorem (Feinberg & H, 2017)

The hitting time assumption holds if and only if there is a function $\mu_\ell : \mathbb{X} \to [0, L]$ *satisfying*

$$\mu_\ell(x) \geqslant 1 + \sum_{y \neq \ell} p(y|x, a)\mu_\ell(y)$$

for all $a \in A(x)$ *and* $x \in \mathbb{X}$*.*

E.g., let

$$\mu_\ell = \max_{\phi} \left\{ \sum_{n=0}^{\infty} {}_\ell P(\phi)^n \mathbf{1} \right\}$$

where $\mathbf{1}_x = 1$ for all $x \in \mathbb{X}$.

Denardo (2016): Such a μ can be computed using at most $O[(n^3 + mn)mL \log L]$ arithmetic operations.

Linear Programming Formulation

$$
\begin{aligned}
\text{minimize} \quad & \sum_{x \in \mathbb{X}} \sum_{a \in A(x)} \frac{c(x,a)}{\mu_\ell(x)} z_{x,a} \\
\text{such that} \quad & \sum_{a \in A(x)} z_{x,a} - \sum_{x' \in \mathbb{X}} \sum_{a' \in A(x')} \frac{p(x|x',a')}{\mu_\ell(x')} \mu_\ell(x) z_{x',a'} = 1, \quad x \neq \ell \\
& \sum_{a \in A(\ell)} z_{\ell,a} - \sum_{x' \in \mathbb{X}} \sum_{a' \in A(x')} \frac{\mu_\ell(x') - 1 - \sum_{y \neq \ell} p(y|x',a') \mu_\ell(y)}{\mu_\ell(x')} z_{x',a'} = 1 \\
& z_{x,a} \geqslant 0, \quad a \in A(x), \; x \in \mathbb{X}
\end{aligned}
$$

For an optimal basic feasible solution z^*,

$$\phi_*(x) = \arg\max_{a \in A(x)} \left\{ z^*_{x,a} \right\}, \quad x \in \mathbb{X}.$$

Theorem

ϕ_* is optimal under the average-cost criterion.

Complexity Estimate

Theorem (Feinberg & H, 2017)

The simplex method with Dantzig's rule solves the linear program (LP) using at most

$$O(nmL \log L) \quad \textit{iterations.}$$

Also, there is a block-pivoting simplex method that solves the LP using at most

$$O(mL \log L) \quad \textit{iterations.}$$

- Each iteration of the simplex method needs $O(n^3 + nm)$ arithmetic operations.
- When L is fixed, these two algorithms are strongly polynomial for average-cost MDPs.
- Result for block-pivoting is special case of result in Akian & Gaubert (2013) for 2-player stochastic games.

Proof Sketch

The LP and the results about it come from a reduction to a discounted MDP with cost-free absorbing state $\bar{x} \notin \mathbb{X}$, based on Akian & Gaubert (2013).

- discount factor $\bar{\beta} = (L-1)/L$
- scaled transition matrices

$$\bar{P}(\phi)_{x,y} = \begin{cases} \bar{\beta}^{-1}\mathrm{diag}(\mu_\ell^{-1})P(\phi)\mathrm{diag}(\mu_\ell)_{x,y}, & x \in \mathbb{X},\ y \in \mathbb{X} \setminus \{\ell\} \\ \bar{\beta}^{-1}\mathrm{diag}(\mu_\ell^{-1})(\mu_\ell - \mathbf{1} - {}_\ell P(\phi)\mu)_{x,y}, & x \in \mathbb{X},\ y = \ell \\ 1 - \bar{\beta}^{-1}\mathrm{diag}(\mu_\ell^{-1})(\mu - \mathbf{1})_x, & x \in \mathbb{X},\ y = \bar{x}, \\ 1, & x = y = \bar{x} \end{cases}$$

  and one-step costs

$$\bar{c}(\phi)_x = \begin{cases} \mathrm{diag}(\mu_\ell^{-1})c(\phi)_x, & x \neq \bar{x} \\ 0, & x = \bar{x} \end{cases}$$

- minimize $\bar{v}(\phi) = \sum_{n=0}^{\infty} \bar{\beta}^n \bar{P}(\phi)^n \bar{c}(\phi)$

Feinberg & Huang (2017): For every policy ϕ, $w(\phi) = \bar{v}(\phi)_\ell \cdot \mathbf{1}$.

Use complexity estimates in Scherrer (2016) for discounted MDPs.

Complexity of Average-Cost MDPs

Average-cost MDPs with special structure are solvable in strongly polynomial time.

- Zadorojniy, Even, Shwartz (2009): M/M/1 queue with service rate control
- Feinberg & H (2013): replacement/maintenance problems with fixed minimal failure probability
  - Feinberg & H (2017): fixed upper bound on expected time to failure

Fearnley (2010): Howard's (1960) policy iteration may need exponential time to solve a multichain average-cost MDP.

- Not known if this is true when MDP is unichain.

Extensions

For total costs, the numbers $p(y|x, a)$ need not be at most one.

- controlled multitype branching processes: Pliska (1976)
- multi-armed bandit problems with risk-seeking utilities: Denardo, Feinberg, Rothblum (2013)

Feinberg & H (2017): For both criteria, the reductions to discounting can be generalized to infinite state and action sets to verify e.g.,

- existence of optimal policies
- validity of optimality equations

The reductions can also be formulated for stochastic games.

- model robust control

Summary

Complexity estimates for certain total-cost and average-cost MDPs

- Conditions under which optimal policies for total-cost and average-cost MDPs can be computed in strongly polynomial time.

Future work:

- Do reductions to discounting hold under more general conditions?
- Generalize to N-player stochastic games.