Computational Complexity Estimates for Policy and Value Iteration Algorithms for Total-Cost and Average-Cost Markov Decision Processes

Jefferson Huang

Department of Applied Mathematics and Statistics
Stony Brook University

The Fifth International Workshop in Sequential Methodologies
Columbia University
June 23, 2015

Joint work with Eugene Feinberg

Plan of the talk

1. Definitions
2. Non-strong polynomiality of the value iteration algorithm for discounted MDPs
3. Reduction of transient MDPs to discounted ones
4. Reduction of average-cost MDPs to discounted ones

Model definition

A discrete-time **Markov decision process (MDP)** is defined by:

1. $\mathbb{X}$ - state space
2. $\mathbb{A}$ - action space
3. $A(x)$ - sets of available actions
4. $c(x, a)$ - one-step costs
5. $q(y|x, a)$ - non-negative transition rates

In this talk,

1. $\mathbb{X}$ is countable
2. $\mathbb{A}$ is a Borel subset of a Polish space
3. $A(x)$ is a Borel subset of $\mathbb{A}$ $\forall x \in \mathbb{X}$.
4. c is bounded, and measurable in $a \in A(x)$ $\forall x \in \mathbb{X}$
5. q is measurable in $a \in A(x)$ $\forall x, y \in \mathbb{X}$, and $\sup\{\sum_{y \in \mathbb{X}} q(y|x, a) \mid x \in \mathbb{X}, a \in A(x)\} < \infty$

Policies

A **policy** is a mapping $\phi : \mathbb{X} \to \mathbb{A}$ where $\phi(x) \in A(x)\ \forall x \in \mathbb{X}$.

- $\mathbb{F}$ - set of all policies

Each $\phi \in \mathbb{F}$ has a corresponding **transition matrix**

$$Q_\phi(x, y) := q(y|x, \phi(x)), \quad x, y \in \mathbb{X},$$

and **cost vector**

$$c_\phi(x) := c(x, \phi(x)), \quad x \in \mathbb{X}.$$

Cost measures

Discounted costs: For $\beta \in [0,1)$,

$$v_\beta^\phi(x) := \sum_{n=0}^{\infty} \beta^n Q_\phi^n c_\phi(x).$$

Undiscounted total costs:

$$v^\phi(x) := \sum_{n=0}^{\infty} Q_\phi^n c_\phi(x).$$

Average costs:

$$w^\phi(x) := \limsup_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} Q_\phi^n c_\phi(x).$$

Optimality criteria

A policy ϕ_* is:

β-**optimal** if

$$v_\beta^{\phi_*}(x) = \inf_{\phi \in \mathbb{F}} v_\beta^{\phi}(x) =: v_\beta(x) \quad \forall x \in \mathbb{X};$$

total-cost optimal if

$$v^{\phi_*}(x) = \inf_{\phi \in \mathbb{F}} v^{\phi}(x) =: v(x) \quad \forall x \in \mathbb{X};$$

average-cost optimal if

$$w^{\phi_*}(x) = \inf_{\phi \in \mathbb{F}} w^{\phi}(x) =: w(x) \quad \forall x \in \mathbb{X}.$$

Computing optimal policies

There are 3 main approaches:

1. **Value iteration**
   - discounted: Shapley (1953)
   - undiscounted total: Bellman (1957), Blackwell (1961, 1967), Strauch (1966)
   - average: White (1963), Schweitzer & Federgruen (1977, 1979)
2. **Policy iteration**
   - discounted: Howard (1960)
   - undiscounted total: Veinott (1969), van der Wal (1981)
   - average: Howard (1960), Veinott (1966)
3. **Linear programming**
   - discounted: D'Epenoux (1963)
   - undiscounted total: Veinott (1969), Kallenberg (1983)
   - average: de Ghellinck (1960) and Manne (1960); Denardo and Fox (1968), Hordijk and Kallenberg (1979, 1980)

Complexity of algorithms

Finite $\mathbb{X}$ and $\mathbb{A}$

$m :=$ number of state-action pairs (x, a), $x \in \mathbb{X}$, $a \in A(x)$

Two classes of "efficient" algorithms:

- **weakly polynomial**: number of *arithmetic operations* needed is bounded above by a polynomial in m & the bit-size L of the input data;
- **strongly polynomial**: number of arithmetic operations needed is bounded above by a polynomial in m only.

Plan of the talk

1. Definitions
2. Non-strong polynomiality of the value iteration algorithm for discounted MDPs
3. Reduction of transient MDPs to discounted ones
4. Reduction of average-cost MDPs to discounted ones

Complexity of algorithms - discounted costs

Take β to be a constant.

Weakly polynomial algorithms exist for all 3 approaches.

1. Value iteration: Tseng (1990)
2. Policy iteration: Meister & Holzbaur (1986)
3. Linear programming: Khachiyan (1979), Karmarkar (1984)

Ye (2011): strongly polynomial algorithms exist for the latter two approaches.

Feinberg & H. (2014): value iteration algorithm is **not** strongly polynomial

Value iteration - discounted costs

For $\beta \in [0,1)$ and $f : \mathbb{X} \to \mathbb{R}$, define the **optimality operator**

$$T_\beta f(x) := \min_{A(x)} \left[c(x,a) + \beta \sum_{y \in \mathbb{X}} q(y|x,a) f(y) \right], \quad x \in \mathbb{X}.$$

Step 0: Pick $V_0 : \mathbb{X} \to \mathbb{R}$, and set $k = 1$.

Step 1: Pick any $\phi^k \in \mathbb{F}$ satisfying $c_{\phi^k} + \beta Q_{\phi^k} V_{k-1} = T_\beta V_{k-1}$.

Step 2:

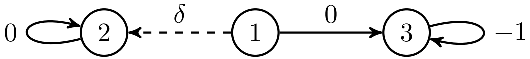
- If $V_{k-1} = T_\beta V_{k-1}$, then ϕ^k is β-optimal.
- Else, set $V_k = T_\beta V_{k-1}$, increase k by 1 and go to **Step 1**.

If $\mathbb{X}$ and $\mathbb{A}$ are finite, and the $q(y|x,a)$'s are transition probabilities, then

$$V_k \to v_\beta \text{ and } \phi^k \text{ is } \beta\text{-optimal for some } k < \infty.$$

The example

Deterministic MDP with $m = 4$ state-action pairs:



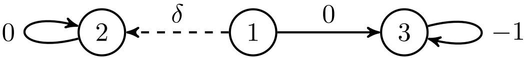
Arcs correspond to actions, and are labeled with their one-step costs.

Note: Suppose $V_0 \equiv 0$. Then at state 1, the solid arc is selected on iteration k only if

$$\delta \geq \beta V_{k-1}(3).$$

Use δ to control the required number of iterations.

The example



Theorem

Let $\beta \in (0, 1)$ and $V_0 \equiv 0$. Then for any positive integer N, there is a $\delta \in \mathbb{R}$ such that at least N iterations are required to find the optimal policy.

Proof. Let δ satisfy

$$-\frac{\beta}{1-\beta} < \delta < -\frac{\beta(1-\beta^{N-1})}{1-\beta}.$$

Then at state 1, the solid arc is the unique optimal action. Also, for $k = 1, \ldots, N$

$$\delta < -\frac{\beta(1-\beta^{N-1})}{1-\beta} \leq -\frac{\beta(1-\beta^{k-1})}{1-\beta} = \beta V_{k-1}(3).$$

□

Non-strong polynomiality

Corollary

The value iteration algorithm is not strongly polynomial.

Proof. By the preceding theorem, the required number of iterations cannot be bounded by a polynomial in m only. □

Feinberg, H., and Scherrer (2014): the same example shows that many **optimistic policy iteration** algorithms are not strongly polynomial.

- Includes Puterman & Shin's (1978) modified policy iteration and Bertsekas & Tsitsiklis's (1996) λ-policy iteration.

Plan of the talk

1. Definitions
2. Non-strong polynomiality of the value iteration algorithm for discounted MDPs
3. Reduction of transient MDPs to discounted ones
4. Reduction of average-cost MDPs to discounted ones

Transient MDPs

For a nonnegative matrix B with entries $B(x, y)$, $x, y \in \mathbb{X}$, let

$$\|B\| := \sup_{x\in\mathbb{X}} \sum_{y\in\mathbb{X}} B(x, y).$$

Assumption T

The MDP is **transient**, i.e., there is a constant K satisfying

$$\|\sum_{n=0}^{\infty} Q_\phi^n\| \leq K < \infty \quad \forall \phi \in \mathbb{F}.$$

There's a strongly polynomial algorithm, due to Eric Denardo, for checking Assumption T - see Veinott (1969).

A preliminary result

Proposition

Suppose the MDP is transient. Then there is a $\mu : \mathbb{X} \to [0, \infty)$ *that is bounded above by* K *and satisfies*

$$\mu(x) \geq 1 + \sum_{y \in \mathbb{X}} q(y|x, a)\mu(y), \quad x \in \mathbb{X}, \; a \in A(x). \tag{1}$$

Proof. When the MDP is transient, the operator

$$\mathcal{U}f(x) := \sup_{A(x)} \left[1 + \sum_{y \in \mathbb{X}} q(y|x, a)f(y) \right], \quad x \in \mathbb{X},$$

has a nonnegative fixed point bounded above by K. □

The Hoffman-Veinott transformation

Extension of an idea attributed to Alan Hoffman by Veinott (1969):

State space: $\tilde{\mathbb{X}} := \mathbb{X} \cup \{\tilde{x}\}$

Action space: $\tilde{\mathbb{A}} := \mathbb{A} \cup \{\tilde{a}\}$

Available actions:

$$\tilde{A}(x) := \begin{cases} A(x), & x \in \mathbb{X}, \\ \{\tilde{a}\}, & x = \tilde{x} \end{cases}$$

One-step costs:

$$\tilde{c}(x, a) := \begin{cases} \mu(x)^{-1} c(x, a), & x \in \mathbb{X}, a \in A(x), \\ 0, & (x, a) = (\tilde{x}, \tilde{a}) \end{cases}$$

The Hoffman-Veinott transformation (continued)

Choose a discount factor

$$\tilde{\beta} \in \left[\frac{K-1}{K}, 1\right).$$

Transition probabilities:

$$\tilde{p}(y|x,a) := \begin{cases} \frac{1}{\tilde{\beta}\mu(x)} q(y|x,a)\mu(y), & x, y \in \mathbb{X}, \\ 1 - \frac{1}{\tilde{\beta}\mu(x)} \sum_{y \in \mathbb{X}} q(y|x,a)\mu(y), & y = \tilde{x},\ x \in \mathbb{X}, \\ 1, & y = x = \tilde{x} \end{cases}$$

Representation of total costs

Proposition

Suppose the MDP is transient, and the one-step costs are bounded. Then

$$v^{\phi}(x) = \mu(x)\tilde{v}^{\phi}_{\tilde{\beta}}(x), \quad \phi \in \mathbb{F},\ x \in \mathbb{X}.$$

Proof. Use the fact that $\tilde{x}$ is a cost-free absorbing state to rewrite $\tilde{v}^{\phi}_{\tilde{\beta}}$ in terms of the original problem data. □

Compactness conditions

Our main results use the following conditions:

Compactness Conditions

(i) $A(x)$ is compact $\forall x \in \mathbb{X}$;

(ii) $c(x, a)$ is:

- bounded in (x, a) where $x \in \mathbb{X}$ and $a \in A(x)$, and
- lower semicontinuous in $a \in A(x)$ $\forall x \in \mathbb{X}$;

(iii) $q(y|x, a)$ is continuous in $a \in A(x)$ $\forall x, y \in \mathbb{X}$;

(iv) $q(\mathbb{X}|x, a) := \sum_{y \in \mathbb{X}} q(y|x, a)$ is continuous in $a \in A(x)$ $\forall x \in \mathbb{X}$.

For a discounted MDP, the Compactness Conditions imply the existence of an optimal policy - see e.g., Feinberg Kasyanov & Zadoianchuk (2012).

Main result for transient MDPs

$$A^*(x) := \{a \in A(x) \mid v(x) = c(x,a) + \sum_{y \in \mathbb{X}} q(y|x,a)v(y)\},\ x \in \mathbb{X}.$$

Theorem - cf. Pliska (1978)

Suppose the MDP is transient, and satisfies the Compactness Conditions. Then:

(i) the value function $v = \mu \tilde{v}_\beta$ is the unique bounded function satisfying

$$v(x) = \min_{A(x)}[c(x,a) + \sum_{y \in \mathbb{X}} q(y|x,a)v(y)], \quad x \in \mathbb{X};$$

(ii) there is a stationary total-cost optimal policy;

(iii) $\phi \in \mathbb{F}$ is total-cost optimal iff. $\phi(x) \in A^*(x)\ \forall x \in \mathbb{X}$, and for $x \in \mathbb{X}$

$$A^*(x) = \{a \in A(x) \mid \tilde{v}_{\tilde{\beta}}(x) = \tilde{c}(x,a) + \tilde{\beta} \sum_{y \in \tilde{\mathbb{X}}} \tilde{p}(y|x,a)\tilde{v}_{\tilde{\beta}}(y)\}.$$

A strongly polynomial algorithm

To compute a total-cost optimal policy for a transient MDP, **solve the LP**

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

When $\tilde{\beta} = (K-1)/K$ and $K > 1$, Scherrer's (2013) results imply that this LP can be solved using

$$\boxed{O(mK \log K) \quad \text{iterations}}$$

of a block-pivoting simplex method corresponding to Howard's policy iteration.

- Ye (2011) and Denardo (2015) also provide complexity estimates for transient MDPs.

Plan of the talk

1. Definitions
2. Non-strong polynomiality of the value iteration algorithm for discounted MDPs
3. Reduction of transient MDPs to discounted ones
4. Reduction of average-cost MDPs to discounted ones

An assumption for average-cost MDPs

For $z \in \mathbb{X}$ and $\phi \in \mathbb{F}$, consider the matrix ${}_zQ_\phi$ with entries

$$ {}_zQ_\phi(x,y) := \begin{cases} q(y|x,\phi(x)), & \text{if } x \in \mathbb{X},\ y \neq z, \\ 0, & \text{if } x \in \mathbb{X},\ y = z. \end{cases} $$

Assumption HT

There is a state $\ell \in \mathbb{X}$ and a constant K^* satisfying

$$ \| \sum_{n=0}^{\infty} {}_\ell Q_\phi^n \| \leq K^* < \infty \quad \text{for all } \phi \in \mathbb{F}. $$

Feinberg & Yang (2008): there's a strongly polynomial algorithm for checking Assumption HT when the $q(y|x,a)$'s are transition probabilities.

The HV-AG transformation

- modification of Akian & Gaubert's (2013) transformation for turn-based zero-sum stochastic games with finite state & action sets
- can be viewed as an extension of the Hoffman-Veinott transformation
- Ross's (1968) transformation can be viewed as a special case

Note: If Assumption HT holds, then there's a $\mu : \mathbb{X} \to [0, \infty)$ that's bounded above by K^* and satisfies

$$\mu(x) \geq 1 + \sum_{y \in \mathbb{X} \setminus \{\ell\}} q(y|x, a)\mu(y), \quad x \in \mathbb{X}, \; a \in A(x);$$

cf. (1).

The HV-AG transformation

State space: $\bar{\mathbb{X}} := \mathbb{X} \cup \{\bar{x}\}$

Action space: $\bar{\mathbb{A}} := \mathbb{A} \cup \{\bar{a}\}$

Available actions:

$$\bar{A}(x) := \begin{cases} A(x), & x \in \mathbb{X}, \\ \{\bar{a}\}, & x = \bar{x} \end{cases}$$

One-step costs:

$$\bar{c}(x, a) := \begin{cases} \mu(x)^{-1} c(x, a), & x \in \mathbb{X}, a \in A(x), \\ 0, & (x, a) = (\bar{x}, \bar{a}) \end{cases}$$

(So far, it's the same as the Hoffman-Veinott transformation.)

The HV-AG transformation (continued)

Choose a discount factor

$$\bar{\beta} \in \left[\frac{K^* - 1}{K^*}, 1\right).$$

Transition probabilities:

$$\bar{p}(y|x,a) := \begin{cases} \frac{1}{\bar{\beta}\mu(x)} q(y|x,a)\mu(y), & y \in \mathbb{X} \setminus \{\ell\},\ x \in \mathbb{X}, \\ \frac{1}{\bar{\beta}\mu(x)}[\mu(x) - 1 - \sum_{y \in \mathbb{X}\setminus\{\ell\}} q(y|x,a)\mu(y)], & y = \ell,\ x \in \mathbb{X}, \\ 1 - \frac{1}{\bar{\beta}\mu(x)}[\mu(x) - 1], & y = \bar{x},\ x \in \mathbb{X}, \\ 1, & y = x = \bar{x} \end{cases}$$

Representation result for average costs

Proposition

For $\phi \in \mathbb{F}$, let $h^\phi(x) := \mu(x)[\bar{v}^\phi_{\bar{\beta}}(x) - \bar{v}^\phi_{\bar{\beta}}(\ell)]$, $x \in \mathbb{X}$. Then

$$\bar{v}^\phi_{\bar{\beta}}(\ell) + h^\phi(x) = c(x, \phi(x)) + \sum_{y \in \mathbb{X}} q(y|x, \phi(x)) h^\phi(y), \quad x \in \mathbb{X}.$$

If the one-step costs c are bounded and the $q(y|x,a)$'s are transition probabilities, then $w^\phi \equiv \bar{v}^\phi_{\bar{\beta}}(\ell)$.

Proof. Rewrite

$$\bar{v}^\phi_{\bar{\beta}}(x) = \bar{c}(x, \phi(x)) + \bar{\beta} \sum_{y \in \bar{\mathbb{X}}} \bar{p}(y|x, \phi(x)) \bar{v}^\phi_{\bar{\beta}}(y), \quad x \in \mathbb{X},$$

in terms of the original problem data. □

Main result for average-cost MDPs

Theorem - cf. Derman (1966), Derman & Veinott (1967), Federgruen & Tijms (1978), Dynkin & Yushkevich (1979)

Suppose the original MDP with transition probabilities q satisfies Assumption HT and the Compactness Conditions. Then:

(i) $w = \bar{v}_{\bar{\beta}}(\ell)$ and $h(x) = \mu(x)[\bar{v}_{\bar{\beta}}(x) - \bar{v}_{\bar{\beta}}(\ell)]$, $x \in \mathbb{X}$, satisfy the optimality equation

$$w + h(x) = \min_{A(x)} \left[c(x,a) + \sum_{y \in \mathbb{X}} q(y|x,a) h(y) \right], \ x \in \mathbb{X};$$

(ii) there is a stationary average-cost optimal policy;

(iii) any $\phi \in \mathbb{F}$ satisfying

$$\phi(x) \in A^*_{\text{av}}(x) := \{a \in A(x) \mid w + h(x) = c(x,a) + \sum_{y \in \mathbb{X}} q(y|x,a) h(y)\}$$

for all $x \in \mathbb{X}$ is average-cost optimal, and for $x \in \mathbb{X}$

$$A^*_{\text{av}}(x) = \{a \in A(x) \mid \bar{v}_{\bar{\beta}}(x) = \bar{c}(x,a) + \bar{\beta} \sum_{y \in \bar{\mathbb{X}}} \bar{p}(y|x,a) \bar{v}_{\bar{\beta}}(y)\}.$$

A strongly polynomial algorithm

To compute an average-cost optimal policy for an MDP with transition probabilities that satisfy Assumption HT, **solve the LP**

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

When $\bar{\beta} = (K^* - 1)/K^*$ and $K^* > 1$, Scherrer's (2013) results imply that this LP can be solved using

$$\boxed{O(mK^* \log K^*) \quad \text{iterations}}$$

of the block-pivoting simplex method corresponding to Howard's policy iteration - see also Akian & Gaubert (2013).

Summary

1. A simple deterministic MDP shows that the value iteration algorithm is not strongly polynomial.
2. Transient MDPs satisfying the Compactness Conditions can be reduced to discounted ones.
3. Average-cost MDPs satisfying Assumption HT and the Compactness Conditions can be reduced to discounted ones.
4. The reductions lead to strongly polynomial algorithms.