Computational Complexity Estimates for Policy and Value Iteration Algorithms for Total-Cost and Average-Cost MDPs

Jefferson Huang

Department of Applied Mathematics and Statistics
Stony Brook University

INFORMS Annual Meeting
Philadelphia, PA
November 2, 2015

Joint work with Eugene A. Feinberg

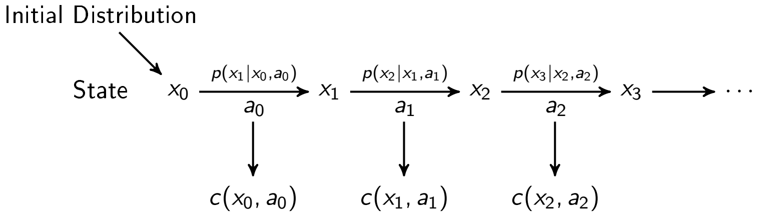
Plan of the talk

1. MDPs & strong polynomiality
2. Value iteration & its generalizations for discounted MDPs
3. Reductions of total & average-cost MDPs to discounted ones

Markov decision processes

Defined by:

1. **state** space $\mathbb{X}$
2. sets of available **actions** $A(x)$ at each state x
3. one-step **costs** $c(x, a)$: incurred whenever the state is x and action $a \in A(x)$ is performed
4. transition **probabilities** $p(y|x, a)$: probability that the next state is y, given that the current state is x & action $a \in A(x)$ is performed



This talk: $\mathbb{X}$ and $A(x)$'s are **finite**.

Policies & cost criteria

A **policy** ϕ prescribes an action for every state.

Common criteria for policies:

- Discounted costs: for $\beta \in (0, 1)$,

$$v_\beta^\phi(x) := \mathbb{E}_x^\phi \sum_{t=0}^{\infty} \beta^n c(x_t, a_t)$$

- Undiscounted total costs: discounted costs with $\beta = 1$.
- Average costs:

$$w^\phi(x) := \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}_x^\phi \sum_{t=0}^{T-1} c(x_t, a_t)$$

A policy is **optimal** if it minimizes the chosen criterion for every initial state.

Computing optimal policies

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)

Strong polynomiality

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

Definition

An algorithm for computing an optimal policy is **strongly polynomial** if there exists an upper bound on the required number of arithmetic operations that

1. is a polynomial in m, and
2. holds for any particular MDP.

Ye (2011): When the discount factor $\beta \in (0, 1)$ is fixed, **Howard's PI** and the simplex method with **Dantzig's pivoting rule** are strongly polynomial.

Feinberg & H. (2014): **Value iteration** is *not* strongly polynomial, even when $\beta \in (0, 1)$ is fixed.

Plan of the talk

1. MDPs & strong polynomiality
2. Value iteration & its generalizations for discounted MDPs
3. Reductions of total & average-cost MDPs to discounted ones

Notation

One-step operator:

$$T_\phi f(x) := c(x,\phi(x)) + \beta \sum_{y\in\mathbb{X}} p(y|x,\phi(x)) f(y)$$

Dynamic Programming (DP) operator:

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

Value function: $v_\beta(x) := \inf_\phi v_\beta^\phi(x)$

Value iteration for discounted MDPs

A policy $\phi \in \mathbb{F}$ is **greedy** with respect to $f : \mathbb{X} \to \mathbb{R}$ if

$$\phi \in \mathcal{G}(f) := \{\varphi \in \mathbb{F} \mid T_\varphi f = Tf\}.$$

Value Iteration (VI): Select any $V_0 : \mathbb{X} \to \mathbb{R}$, and iteratively apply the DP operator.

$$V_0 \longrightarrow V_1 = TV_0 \longrightarrow V_2 = TV_1 \longrightarrow \cdots \longrightarrow V_j = TV_{j-1} \longrightarrow \cdots$$

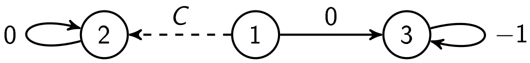
$$\phi^1 \in \mathcal{G}(V_0) \qquad \phi^2 \in \mathcal{G}(V_1) \qquad \phi^3 \in \mathcal{G}(V_2) \qquad \phi^{j+1} \in \mathcal{G}(V_j)$$

For $\beta \in (0, 1)$,

- $\lim_{j\to\infty} V_j(x) = v_\beta(x)$ for all $x \in \mathbb{X}$.
- For some $j < \infty$, ϕ^j is optimal.

The example

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



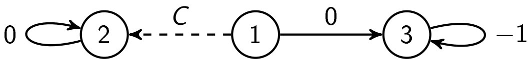
Arcs: correspond to actions, labeled with their one-step costs.

Note: Suppose $V_0 \equiv 0$. Then at state 1, the solid arc is selected for the j^{th} policy only if

$$C \geq \beta V_{j-1}(3).$$

Idea: Use C to control the required number of iterations.

The example



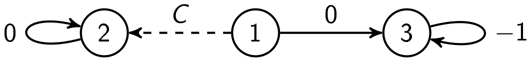
Theorem (Feinberg & H. 2014)

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

Corollary

Value iteration is not strongly polynomial.

Proof of the Theorem



Let C satisfy

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

Then at state 1, the solid arc is the unique optimal action. Also, $C < 0 = V_0(3)$, and for $j = 1, \ldots, N$

$$C < -\frac{\beta(1-\beta^N)}{1-\beta} \leq -\frac{\beta(1-\beta^j)}{1-\beta} = \beta V_j(3).$$

But, the optimal policy is selected only if $C \geq \beta V_{j-1}(3)$. $\square$

Generalized optimistic policy iteration

$\bar{\mathbb{N}} := \{1, 2, \dots\} \cup \{\infty\}$

Let $\{N_j\}_{j=1}^{\infty}$ be a $\bar{\mathbb{N}}$-valued stochastic sequence with associated probability measure P and expectation operator E.

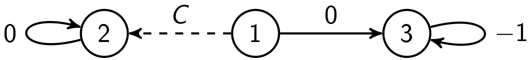
Generalized Optimistic PI: Select any $V_0 : \mathbb{X} \to \mathbb{R}$ and iteratively generate $\{V_j\}_{j=1}^{\infty}$ as follows:

$$V_0 \to V_1 = E[T_{\phi^1}^{N_1} V_0] \to V_2 = E[T_{\phi^2}^{N_2} V_1] \to \cdots \to V_j = E[T_{\phi^j}^{N_j} V_{j-1}] \to \cdots$$

$$\phi^1 \in \mathcal{G}(V_0) \quad \phi^2 \in \mathcal{G}(V_1) \quad \phi^3 \in \mathcal{G}(V_2) \quad \phi^{j+1} \in \mathcal{G}(V_j)$$

Special cases: **VI** (N_j's $\equiv 1$), **modified PI** (Puterman & Shin 1978), λ-**PI** (Bertsekas & Tsitsiklis 1996), **optimistic PI** (Thiéry & Scherrer 2010), **Howard's PI** (N_j's $\equiv \infty$)

Generalized optimistic policy iteration



Theorem (Feinberg, H., & Scherrer 2014)

Let $\beta \in (0, 1)$ and $V_0 \equiv 0$. Suppose $P\{N_j < \infty\} > 0$ for all j. Then for any positive integer N, there is a $C \in \mathbb{R}$ such that at least N iterations are required by Generalized Optimistic PI to find the optimal policy.

Corollary

Value iteration, modified policy iteration, λ-policy iteration, and optimistic policy iteration are not strongly polynomial.

Plan of the talk

1. MDPs & strong polynomiality
2. Value iteration & its generalizations for discounted MDPs
3. Reductions of total & average-cost MDPs to discounted ones

Reductions to discounted MDPs

For $x \in \mathbb{X}$, let $\tau_x := \inf\{t \geq 1 \mid x_t = x\}$.

Theorem

Suppose there's a state $\ell \in \mathbb{X}$ and a constant K satisfying

$$\mathbb{E}_x^\phi \tau_\ell \leq K < \infty \quad \text{for all } x \in \mathbb{X},\ \phi \in \mathbb{F}.$$

Then:

(i) *an* **average-cost** *optimal policy can be found by solving a discounted MDP;*

(ii) *if ℓ is a cost-free absorbing state, then an* **undiscounted total-cost** *optimal policy can be found by solving a discounted MDP.*

Feinberg & H. (2015): Conditions under which

- the Theorem holds for MDPs with infinite $\mathbb{X}$ and $A(x)$'s, and
- (ii) holds for a more general model.

Checking the assumption

Let $m := |\cup_{x\in\mathbb{X}} |A(x)||$ and $n := |\mathbb{X}|$.

The assumption that

$$\mathbb{E}_x^{\phi}\tau_\ell \leq K < \infty \text{ for all } x \in \mathbb{X},\ \phi \in \mathbb{F}. \tag{1}$$

can be checked using $O(mn^2)$ arithmetic operations.

- For average costs, (1) holds iff. the MDP is unichain and has a recurrent state ℓ, which can be checked with $O(mn^2)$ arithmetic operations (Feinberg & Yang 2008).
- For undiscounted total costs, (1) can be checked for a given cost-free absorbing state using $O(mn)$ arithmetic operations (Veinott 1974).

Construction of the discounted MDPs

Proposition

For $\ell \in \mathbb{X}$,

$$\mathbb{E}_x^{\phi}\tau_\ell \le K < \infty \quad \textit{for all } x \in \mathbb{X},\ \phi \in \mathbb{F}.$$

if and only if there's a $\mu : \mathbb{X} \to [0, \infty)$ that's bounded above by K and satisfies

$$\mu(x) \ge 1 + \sum_{y \in \mathbb{X}\setminus\{\ell\}} p(y|x, a)\mu(y) \quad \textit{for all } x \in \mathbb{X},\ a \in A(x).$$

Use μ to construct the discounted MDPs, by extending ideas of Alan Hoffman (Veinott 1969) and Akian & Gaubert (2013).

Computing μ

For $x \in \mathbb{X}$, let $\tau(x) := \max_{\phi \in \mathbb{F}} \mathbb{E}_x^{\phi} \tau_\ell$. Then

$$\tau(x) = \max_{a \in A(x)} \left[1 + \sum_{y \in \mathbb{X} \setminus \{\ell\}} p(y|x,a)\tau(y) \right], \quad x \in \mathbb{X}.$$

It follows from Denardo (2015) that τ can be computed using $O(mn \cdot mnK \log(nK))$ arithmetic operations.

It's also possible to use ideas from Veinott (1974) to compute a $\mu \geq \tau$ using $O(n^3 + mn)$ arithmetic operations.

Construction of the discounted MDPs

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

Action sets: for $x \in \tilde{\mathbb{X}}$,

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

One-step costs: for $x \in \tilde{\mathbb{X}}$ and $a \in \tilde{A}(x)$,

$$\tilde{c}(x, a) := \begin{cases} c(x, a)/\mu(x), & \text{if } x \in \mathbb{X}, \\ 0, & \text{if } x = \tilde{x}. \end{cases}$$

Discount factor:

$$\tilde{\beta} := \frac{K - 1}{K}.$$

Transition probabilities for the discounted MDPs

When the original criterion is **average costs**, use the transition probabilities

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

For **undiscounted total costs**, the transition probabilities are

$$\tilde{p}_{\mathrm{tot}}(y|x,a) := \begin{cases} \frac{1}{\tilde{\beta}\mu(x)} p(y|x,a)\mu(y), & y, x \in \mathbb{X} \setminus \{\ell\}, \\ 0, & y = \ell,\ x \in \mathbb{X} \setminus \{\ell\}, \\ 1 - \frac{1}{\tilde{\beta}\mu(x)} \sum_{y \in \mathbb{X} \setminus \{\ell\}} p(y|x,a)\mu(y), & y = \tilde{x},\ x \in \mathbb{X} \setminus \{\ell\}, \\ 1, & y = x \in \{\ell, \tilde{x}\} \end{cases}$$

Representation of average costs

Let $\tilde{v}^{\phi}_{\tilde{\beta}}$ be the discounted cost function under $\phi \in \mathbb{F}$ for the MDP $(\tilde{\mathbb{X}}, \tilde{A}(\cdot), \tilde{c}, \tilde{p}_{\mathrm{av}})$.

Proposition

Let $h^{\phi}(x) := \mu(x)[\tilde{v}^{\phi}_{\tilde{\beta}}(x) - \tilde{v}^{\phi}_{\tilde{\beta}}(\ell)]$, $x \in \mathbb{X}$. *Then*

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

and $w^{\phi} \equiv \tilde{v}^{\phi}_{\tilde{\beta}}(\ell)$.

Corollary

Any optimal policy for the new discounted MDP is average-cost optimal for the original MDP.

Representation of undiscounted total costs

Now let $\tilde{v}_{\tilde{\beta}}^{\phi}$ be the discounted cost function under $\phi \in \mathbb{F}$ for the MDP $(\tilde{\mathbb{X}}, \tilde{A}(\cdot), \tilde{c}, \tilde{p}_{\text{tot}})$.

Proposition

If ℓ is a cost-free absorbing state, then

$$v_1^{\phi}(x) = \mu(x)\tilde{v}_{\tilde{\beta}}^{\phi}(x), \quad x \in \mathbb{X}.$$

Corollary

Any optimal policy for the new discounted MDP is undiscounted total-cost optimal for the original MDP.

Computing an optimal policy

To compute an **average-cost** optimal policy, solve the LP

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

To compute an **undiscounted total-cost** optimal policy, solve the above LP with $\tilde{p}_{\text{av}}$ replaced by $\tilde{p}_{\text{tot}}$.

When $K > 1$, for both $\tilde{p}_{\text{av}}$ and $\tilde{p}_{\text{tot}}$ Scherrer's (2013) results imply the LP can be solved using the **simplex method** with

- **Dantzig's** rule, using $O(mnK \log K)$ iterations, or
- the **block-pivoting** rule corresponding to Howard's PI, using $O(mK \log K)$ iterations.

Summary

1. Unlike Howard's PI and the simplex method with Dantzig's rule, value iteration and many of its generalizations are **not strongly polynomial**.

2. If there's a state ℓ satisfying

$$\mathbb{E}_x^{\phi}\tau_\ell \le K < \infty \quad \text{for all } x \in \mathbb{X},\ \phi \in \mathbb{F},$$

   then both an average-cost optimal policy, and an undiscounted total-cost optimal policy when ℓ is cost-free and absorbing, can be computed by:

   (1) computing a function μ using $O(m^2 n^2 K \log nK)$ arithmetic operations;
   (2) constructing a discounted MDP using $O(mn)$ arithmetic operations;
   (3) computing an optimal policy for the discounted MDP using $O(mn \cdot mK \log K)$ arithmetic operations.