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

Jefferson Huang

Department of Applied Mathematics and Statistics
Stony Brook University

AP for Lunch Seminar
IBM T. J. Watson Research Center
July 29, 2015

Joint work with Eugene A. Feinberg

Plan of the talk

1. Background on Markov decision processes (MDPs) & computational complexity theory
2. Value iteration & optimistic policy iteration for discounted MDPs
3. Reductions of total & average-cost MDPs to discounted ones

Markov decision processes

Defined by **4 objects**:

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

Initial Distribution

State $\quad x_0 \xrightarrow[a_0]{p(x_1|x_0,a_0)} x_1 \xrightarrow[a_1]{p(x_2|x_1,a_1)} x_2 \xrightarrow[a_2]{p(x_3|x_2,a_2)} x_3 \longrightarrow \cdots$

Policies & cost criteria

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

Common cost criteria for policies are:

- Total (discounted) costs: for $\beta \in [0, 1]$,

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

- Average costs:

$$w^\phi(x) := \limsup_{N \to \infty} \frac{1}{N} \mathbb{E}_x^\phi \sum_{n=0}^{N} c(x_n, a_n)$$

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

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)

Applications of MDPs

First (?) application of MDPs: Sears mail-order catalogs ($\sim$1958)

Ronald A. Howard (1978):

> *... my one successful application was the original application that sparked my interest in this whole research area.*

Some others:

- **Operations Research:** inventory control, control of queues, vehicle routing, job shop scheduling
- **Finance:** Option pricing, portfolio selection, credit granting
- **Healthcare:** medical decision making, epidemic control
- **Power Systems:** Voltage & reactive power control, economic dispatch, bidding in electricity markets with storage, charging electric vehicles
- **Computer Science:** robot motion planning, model checking, playing video games

MDPs and pure mathematics

Ronald A. Howard (1978):

> *The Markov decision process and its extensions have now become principally the province of mathematicians.*

Borel-space MDPs: Blackwell (1965), Strauch (1966)

> $\rightarrow$ connections to **descriptive set theory**: see e.g. Bertsekas & Shreve (1978), Dynkin & Yushkevich (1979)

Motivated **counterexamples** on:

- theory of Borel sets, semicontinuity of minimum functions

and **new results** on:

- extensions of Berge's Theorems & Fatou's Lemma
- convergence of probability measures, solutions of Kolmogorov's equations

MDPs and computational complexity theory

Optimal policies can be computed in **polynomial time**.

- For discounted MDPs, this can be done with **value iteration** (Tseng 1990), **policy iteration** (Meister & Holzbaur 1986), or via **linear programming** (Khachiyan 1979).
- For average-cost MDPs and certain undiscounted total-cost MDPs, this can be done via linear programming.
- Computing an optimal policy is **P-complete**: Papadimitriou & Tsitsiklis (1987).

It gets harder for partially observable MDPs and constrained MDPs: see e.g. Papadimitriou & Tsitsiklis (1987), Madani Hanks & Condon (1999), Feinberg (2000).

MDPs and computational complexity theory

Policy iteration (PI) is closely related to the **simplex method** for linear programming.

This has been used to show that:

- many simplex pivoting rules can require a **super-polynomial** number of iterations: Melekopoglou & Condon (1994), Friedmann (2011, 2012), Friedmann Hansen & Zwick (2011);
- certain decision problems associated with the simplex method are **PSPACE-complete**: Fearnley & Savani (2015);
- for certain problems, classic simplex pivoting rules (e.g. Dantzig's, Gass-Saaty) are **strongly polynomial**: Ye (2011), Kitahara & Mizuno (2011), Even & Zadorojniy (2012), Feinberg & H. (2013)

MDPs and computational complexity theory

This talk:

- Value iteration and some of its generalizations aren't strongly polynomial for discounted MDPs.
- Under certain conditions, undiscounted total-cost and average-cost MDPs can be reduced to discounted ones.
  - Leads to attractive iteration bounds for algorithms

Plan of the talk

1. Background on Markov decision processes (MDPs) & computational complexity theory
2. Value iteration & optimistic policy iteration 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

Idea: Approximate the value function by iterating the DP operator

Value Iteration (VI)

1: Select a function $V_0 : \mathbb{X} \to \mathbb{R}$, and set $j = 1$.
2: Select a policy ϕ^j satisfying $T_{\phi^j} V_{j-1} = TV_{j-1}$.
3: **if** $V_{j-1} = TV_{j-1}$ **then**
4: Stop.
5: **else**
6: Set $V_j = TV_{j-1}$, and set $j = j + 1$.
7: **go to** 2.

It's well-known that:

- $V_j(x) \to v_\beta(x)$ for all $x \in \mathbb{X}$.
- After a finite number of iterations, VI terminates with an optimal policy.

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

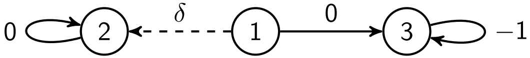
1. is a polynomial in m, and
2. holds for any particular MDP.

Ye (2011): When the discount factor 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.

The example

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



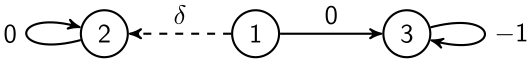
Arcs: correspond to actions, labeled with their one-step costs.

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

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

Idea: 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.

Corollary

Value iteration is not strongly polynomial.

Proof of the Theorem

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 $j = 1, \ldots, N$

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

However, the optimal policy is selected only if $\delta \ge \beta V_{j-1}(3)$. $\square$

Optimistic policy iteration

Howard's PI converges at least as quickly as value iteration; see e.g. Puterman (1994).

Howard's Policy Iteration (PI)

1: Select a function $V_0 : \mathbb{X} \to \mathbb{R}$, and set $j = 1$.
2: Select a policy ϕ^j satisfying $T_{\phi^j} V_{j-1} = T V_{j-1}$.
3: **if** $V_{j-1} = T V_{j-1}$ **then**
4: Stop.
5: **else**
6: Set $V_j = v_\beta^{\phi^j} = \lim_{N\to\infty} T_{\phi^j}^N V_{j-1}$, and set $j = j + 1$.
7: **go to** 2.

The vector $v_\beta^{\phi^j}$ is the solution of a linear system of equations.

Idea: Replace $v_\beta^{\phi^j}$ with an approximation (be **"optimistic"** about the need to evaluate ϕ^j exactly).

Optimistic policy iteration algorithms

$v_\beta^{\phi^j} = \lim_{N\to\infty} T_{\phi^j}^N V_{j-1}$

Value iteration: Replace $v_\beta^{\phi^j}$ with $TV_{j-1} = T_{\phi^j} V_{j-1}$.

Modified policy iteration: Replace $v_\beta^{\phi^j}$ with $T_{\phi^j}^{n_j} V_{j-1}$. (Puterman & Shin 1978)

λ-**policy iteration:** Replace $v_\beta^{\phi^j}$ with $(1-\lambda_j)\sum_{n=1}^{\infty} \lambda_j^{n-1} T_{\phi^j}^n V_{j-1}$, where $\lambda_j \in [0,1)$. (Bertsekas & Tsitsiklis 1996)

Optimistic policy iteration: Replace $v_\beta^{\phi^j}$ with $\sum_{n=1}^{\infty} \lambda_{j,n} T_{\phi^j}^n V_{j-1}$, where $\lambda_{j,n} \geq 0$ for all n and $\sum_{n=1}^{\infty} \lambda_{j,n} = 1$. (Thiéry & Scherrer 2010)

Feinberg, H., and Scherrer (2014): the preceding example shows that none of these are strongly polynomial.

Generalized optimistic policy iteration

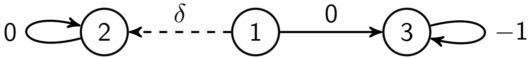
$\bar{\mathbb{N}} := \{1, 2, \dots\} \cup \{\infty\}$

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

Generalized Optimistic Policy Iteration

1: Select a function $V_0 : \mathbb{X} \to \mathbb{R}$, and set $j = 1$.
2: Select a policy ϕ^j satisfying $T_{\phi^j} V_{j-1} = T V_{j-1}$.
3: **if** $V_{j-1} = T V_{j-1}$ **then**
4:     Stop.
5: **else**
6:     Set $V_j = E[T_{\phi^j}^{N_j} V_{j-1}]$, and set $j = j + 1$.
7:     **go to** 2.

Generalized optimistic policy iteration



Theorem

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 $\delta \in \mathbb{R}$ 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.

Proof of the Theorem

Let δ satisfy

$$-\frac{\beta}{1-\beta} < \delta < -\frac{\beta(1-\prod_{\ell=1}^{N-1} E[\beta^{N_\ell}])}{1-\beta}.$$

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

$$\begin{aligned}\delta &< -\frac{\beta(1-\prod_{\ell=1}^{N-1} E[\beta^{N_\ell}])}{1-\beta} \\ &\leq -\frac{\beta(1-\prod_{\ell=1}^{j-1} E[\beta^{N_\ell}])}{1-\beta} = \beta V_{j-1}(3).\end{aligned}$$

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

Plan of the talk

1. Background on Markov decision processes (MDPs) & computational complexity theory
2. Value iteration & optimistic policy iteration for discounted MDPs
3. Reductions of total & average-cost MDPs to discounted ones

Reductions to discounted MDPs

Discounted-cost MDPs are generally easier to study than undiscounted ones.

This talk: Reductions to discounted MDPs of:

1. undiscounted total-cost MDPs that are **transient**;
2. average-cost MDPs satisfying a uniform **hitting time assumption**.

Transient MDPs

Here the numbers $p(y|x,a)$ are allowed to not correspond to transition probabilities.

Can be used to model:

- stochastic shortest path problems (e.g. Bertsekas 2005)
- controlled multitype branching processes (e.g. Pliska 1978, Rothblum & Veinott 1992)

It's well-known that **discounted MDPs can be reduced to transient ones** (e.g. Altman 1999).

Feinberg & H. (2015): conditions under which the **converse** is true for infinite-state MDPs.

Transient MDPs

$P_\phi := [p(y|x,\phi(x))]_{x,y\in\mathbb{X}}$ = nonnegative transition matrix associated with policy ϕ.

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

$$\|\sum_{n=0}^{\infty} P_\phi^n\| \le K < \infty \quad \forall\phi.$$

A preliminary result

Proposition

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

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

Idea: Use μ to transform the transient MDP into a discounted one with transition probabilities.

The Hoffman-Veinott transformation

Extension of an idea attributed to Alan Hoffman (an IBM Fellow Emeritus) 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:

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

Representation of total costs

Proposition

Suppose the MDP is transient, and the one-step costs are bounded. Then for any policy ϕ,

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

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. $\square$

Corollary

Any optimal policy for the new discounted MDP is optimal for the original transient MDP.

Computing an optimal policy

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. Background on Markov decision processes (MDPs) & computational complexity theory
2. Value iteration & optimistic policy iteration for discounted MDPs
3. Reductions of total & average-cost MDPs to discounted ones

An assumption for average-cost MDPs

Back to transition probabilities $p(y|x, a)$

ℓ is a fixed state

$\tau_\ell := \inf\{n \geq 1 | x_n = \ell\}$ = hitting time to ℓ

Assumption HT

There's a constant K^* such that for any policy ϕ,

$$\mathbb{E}_x^\phi \tau_\ell \leq K^* < \infty \quad \forall x \in \mathbb{X}.$$

Holds for replacement & maintenance problems. (e.g. ℓ = machine is broken)

Sufficient condition for Assumption HT

Assumption D

There's a positive integer N & constant α where, for all policies ϕ,

$$\mathbb{P}_x^{\phi}\{x_N = \ell\} \geq \alpha > 0 \quad \forall x \in \mathbb{X}.$$

- Special case of Hordijk's (1974) *simultaneous Doeblin condition*.
- Ross's (1968) assumption: $N = 1$.
- Implies that for all policies ϕ

$$\mathbb{E}_x^{\phi} \tau_\ell \leq N/\alpha < \infty \quad \forall x \in \mathbb{X}.$$

Implications of Assumption HT

$P_\phi :=$ Markov chain corresponding to policy ϕ

Assumption HT implies:

- state ℓ is *positive recurrent* $\forall \phi$.
- MDP is **unichain**, i.e. P_ϕ has a single recurrent class $\forall \phi$.
- If P_ϕ is aperiodic $\forall \phi$,
  - each P_ϕ has a stationary distribution π_ϕ;
  - each P_ϕ is **fast mixing** - $\exists$ positive integer N and $\rho < 1$ where

$$\sup_{B \subseteq \mathbb{X}} \left| \sum_{y \in B} P_\phi^n(x, y) - \sum_{y \in B} \pi_\phi(y) \right| \leq \rho^{\lfloor n/N \rfloor} \quad \forall x \in \mathbb{X}, n \geq 1;$$

    see Federgruen, Hordijk, and Tijms (1978).

- average cost w^ϕ is **constant** $\forall \phi$.

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

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

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:

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

Representation result for average costs

Proposition

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

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

If the one-step costs c *are bounded, then* $w^\phi \equiv \bar{v}^\phi_{\bar{\beta}}(\ell)$.

Corollary

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

Computing an optimal policy

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

- Value iteration and many optimistic PI algorithms are not strongly polynomial.
- Transient MDPs can be reduced to discounted ones.
- Average-cost MDPs satisfying a hitting time assumption can be reduced to discounted ones.
- These reductions lead to alternative algorithms with attractive complexity estimates.