Computational complexity estimates for value and policy iteration algorithms for total-cost and average-cost Markov decision processes

Jefferson Huang

Dept. Applied Mathematics and Statistics
Stony Brook University

AI Seminar
University of Alberta
May 5, 2016

Joint work with Eugene A. Feinberg

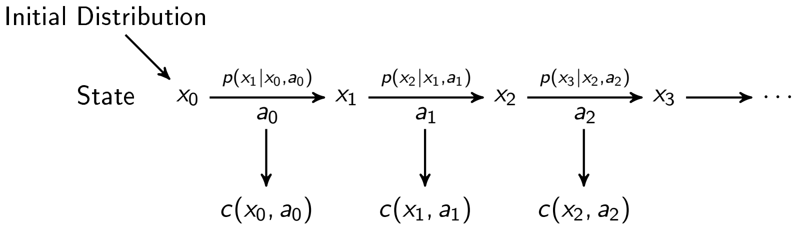
Outline

1. Background on Markov decision processes (MDPs) & complexity of algorithms
2. Complexity of optimistic policy iteration (e.g., value iteration, λ-policy iteration) for discounted MDPs
3. Reductions of total & average-cost MDPs to discounted ones

Markov decision process (MDP)

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



Policies & cost criteria

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

Common cost criteria for policies:

- **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.

Examples of MDPs

- **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:** model checking, robot motion planning, playing classic games

Computing optimal policies

3 main (and related) approaches:

1. **Value iteration (VI)** (Shapley 1953)
   - Iteratively approximate the optimal cost function.
2. **Policy iteration (PI)** (Howard 1960)
   - Iteratively improve a starting policy.
3. **Linear programming (LP)** (early 1960s)
   - Compute the optimal frequencies with which each state-action pair should be used.

Complexity of computing optimal policies

Optimal policies can be computed in (weakly) **polynomial time**:

- for discounted MDPs, via **value iteration** (Tseng 1990), **policy iteration** (Meister & Holzbaur 1986), or **linear programming** (Khachiyan 1979);
- for average-cost MDPs and certain undiscounted total-cost MDPs, via linear programming.

Computing an optimal policy is **P-complete**: Papadimitriou & Tsitsiklis (1987).

Solving *constrained MDPs* and *partially observable MDPs* is harder: Feinberg (2000), Papadimitriou & Tsitsiklis (1987)

Applications to the complexity of the simplex method

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

This has been **used to show that**:

- many simplex pivoting rules may need a **super-polynomial** number of iterations: Melekopoglou & Condon (1994), Friedmann (2011, 2012), Friedmann Hansen & Zwick (2011);
- for certain problems, **classic simplex pivoting rules** (e.g., Dantzig, Gass-Saaty) are **strongly polynomial**: Ye (2011), Kitahara & Mizuno (2011), Even & Zadorojniy (2012), Feinberg & H. (2013)

Complexity of computing optimal policies

This talk:

- Value iteration and many 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**.
  - Discounted MDPs are **generally easier to study**
  - Leads to attractive iteration bounds for algorithms

Outline

1. Background on Markov decision processes (MDPs) & complexity of algorithms
2. Complexity of optimistic policy iteration (e.g., value iteration, λ-policy iteration) for discounted MDPs
3. Reductions of total & average-cost MDPs to discounted ones

Notation

Here, the state & action sets are finite.

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: $v_\beta(x) := \min_\phi v_\beta^\phi(x)$

Value iteration for discounted MDPs

A policy ϕ 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.

$$\begin{array}{ccccccccc}
V_0 & \longrightarrow & V_1 = TV_0 & \longrightarrow & V_2 = TV_1 & \longrightarrow \cdots \longrightarrow & V_j = TV_{j-1} & \longrightarrow & \cdots \\
\downarrow & & \downarrow & & \downarrow & & \downarrow & & \\
\phi^1 \in \mathcal{G}(V_0) & & \phi^2 \in \mathcal{G}(V_1) & & \phi^3 \in \mathcal{G}(V_2) & & \phi^{j+1} \in \mathcal{G}(V_j) & &
\end{array}$$

Well-known that 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.

Strong polynomiality

m $:=$ **number of state-action pairs** (x, a).

Definition

An algorithm for computing an optimal policy is **strongly polynomial** if there's an upper bound on the required number of arithmetic operations that's a polynomial in m only.

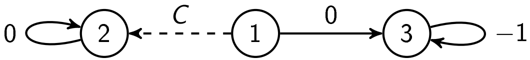
Ye (2011): When the discount factor is fixed, **Howard's PI** and the simplex method with **Dantzig's pivoting rule** are strongly polynomial.

Feinberg & H. (2014): **VI** is not strongly polynomial.

Feinberg H. & Scherrer (2014): many **generalizations of VI** are not strongly polynomial.

The example

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



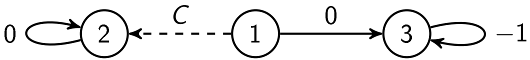
Arcs: correspond to actions, labeled with their one-step costs.

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

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

Idea: Use C 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 $C \in \mathbb{R}$ such that VI needs at least N iterations to return the optimal policy.

Corollary

VI is not strongly polynomial.

Policy iteration for discounted MDPs

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

$$
\begin{array}{ccccccccc}
V_0 & \longrightarrow & V_1 = v_\beta^{\phi^1} & \longrightarrow & V_2 = v_\beta^{\phi^2} & \longrightarrow \cdots \longrightarrow & V_j = v_\beta^{\phi^j} & \longrightarrow & \cdots \\
\rightsquigarrow & & \rightsquigarrow & & \rightsquigarrow & & \rightsquigarrow & & \\
\phi^1 \in \mathcal{G}(V_0) & & \phi^2 \in \mathcal{G}(V_1) & & \phi^3 \in \mathcal{G}(V_2) & & \phi^{j+1} \in \mathcal{G}(V_j) & &
\end{array}
$$

$v_\beta^{\phi^j}$ 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).

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.

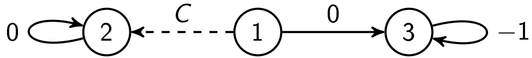
Generalized Optimistic PI: Select any $V_0 : \mathbb{X} \to \mathbb{R}$ and iteratively generate $\{V_j\}_{j=1}^{\infty}$ as follows:

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

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

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

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 generalized optimistic PI needs at least N iterations to return the optimal policy.

Corollary

VI, modified PI, λ-PI, and optimistic PI are not strongly polynomial.

Outline

1. Background on Markov decision processes (MDPs) & complexity of algorithms
2. Complexity of optimistic policy iteration (e.g., value iteration, λ-policy iteration) for discounted MDPs
3. Reductions of total & average-cost MDPs to discounted ones

Reductions to discounted MDPs

Discounted MDPs: 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

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

For a matrix $B = [B(x,y)]_{x,y\in\mathbb{X}}$, let $\|B\| := \sup_{x\in\mathbb{X}} \sum_{y\in\mathbb{X}} |B(x,y)|$.

Assumption T (Transience)

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

$$\|\sum_{n=0}^{\infty} P_\phi^n\| \le K < \infty \quad \text{for all policies } \phi.$$

Interpretation: the "lifetime" of the process is bounded over all policies and initial states.

Veinott (1969): There's a strongly polynomial algorithm for **checking** if Assumption T holds.

Transient MDPs

Note: Here, $p(y|x,a) \geq 0$ may satisfy $\sum_{y \in \mathbb{X}} p(y|x,a) \neq 1$.

Can be **used to model**:

- stochastic shortest path problems (e.g., Bertsekas 2005)
- controlled multitype branching processes (e.g., Pliska 1976, Rothblum & Veinott 1992)

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

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

Characterization of transience

Proposition

An MDP is transient iff there's a $\mu : \mathbb{X} \to [1, \infty)$ that's bounded above by K and satisfies

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

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

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:

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

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. Then for any policy ϕ,

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

Idea: Rewrite $\tilde{v}^{\phi}_{\tilde{\beta}}$ in terms of the original problem data, and use the fact that $\tilde{x}$ is a cost-free absorbing state.

Corollary

A policy is optimal for the new discounted MDP iff it's optimal for the original transient MDP.

Computing an optimal policy

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

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

Scherrer's (2016) results imply that this linear program can be solved using

$\boldsymbol{O(mK \log K)}$ iterations

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

- Ye (2011) and Denardo (2016) also provide complexity estimates for transient MDPs.

Computing the function μ

Choice of μ affects the iteration bound!

When $\sum_{y \in \mathbb{X}} p(y|x,a) \leq 1$ for all (x,a), a μ and K can be computed using $\boldsymbol{O(mn + n^3)}$ arithmetic operations. ($n =$ number of states)

- Idea: Construct a "dominating" Markov chain.

In general, a suitable $\mu \leq \sup_\phi \| \sum_{n \geq 0} P_\phi^n \| =: K^*$ can be computed using $\boldsymbol{O((mn + n^2)mK^* \log K^*)}$ arithmetic operations.

- Idea: Replace all costs with -1 and solve the LP for the resulting total-cost MDP. For the complexity result, follow the proofs in Scherrer (2016) using a weighted norm instead of the max-norm.

Theorem

Suppose $\sup_\phi \| \sum_{n \geq 0} P_\phi^n \| < \infty$ is fixed. Then there's a strongly polynomial algorithm that returns a total-cost optimal policy for the transient MDP, which involves the solution of two linear programs.

Outline

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

Assumption for average-cost MDPs

$\tau_x := \inf\{n \geq 1 | x_n = x\}$ = **hitting time** to x

Assumption HT (Hitting Time)

There's a state ℓ and a constant L such that for any policy ϕ,

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

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

Feinberg & Yang (2008): There's a strongly polynomial algorithm for **checking** if Assumption HT holds.

Sufficient condition for Assumption HT

Assumption

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**, i.e. $\exists$ 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 & Tijms (1978).
- average cost w^ϕ is **constant** $\forall \phi$.

HV-AG transformation

- Modification of Akian & Gaubert's (2013) transformation for zero-sum turn-based 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**.

Proposition

If Assumption HT holds, then there's a $\mu : \mathbb{X} \to [1, \infty)$ *that's bounded above by L and satisfies*

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

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:

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

HV-AG transformation (continued)

Choose a discount factor

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

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 of average costs

Proposition

If the one-step costs c are bounded, then any policy ϕ satisfies $w^{\phi} \equiv \bar{v}^{\phi}_{\bar{\beta}}(\ell)$.

Idea: Use the fact that $h^{\phi}(x) := \mu(x)[\bar{v}^{\phi}_{\bar{\beta}}(x) - \bar{v}^{\phi}_{\bar{\beta}}(\ell)]$, $x \in \mathbb{X}$, satisfies

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

Corollary

If c is bounded, then 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 that satisfies Assumption HT, solve

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

Scherrer's (2016) results imply that this LP can be solved using

$$\boxed{\boldsymbol{O(mL \log L)} \quad \text{iterations}}$$

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

Computing the function μ

If Assumption HT holds, a state ℓ satisfying Assumption HT can be found using $\boldsymbol{O(mn^2)}$ arithmetic operations (Feinberg & Yang 2008).

A suitable $\mu \le \sup_{x\in\mathbb{X}} \sup_\phi \mathbb{E}_x^\phi \tau_\ell =: L^*$ can then be computed using $\boldsymbol{O((mn+n^2)mL^* \log L^*)}$ arithmetic operations.

- Idea: Remove state ℓ, set $p(\ell|\cdot) \equiv 0$, set all one-step costs to -1, and consider the LP for the resulting transient total-cost MDP.

Theorem

Suppose $\sup_{x\in\mathbb{X}} \sup_\phi \mathbb{E}_x^\phi \tau_\ell < \infty$ *is fixed. Then there's a strongly polynomial algorithm that returns an optimal policy for the average-cost MDP, which involves the solution of two linear programs.*

Summary

- Any member of a large class of optimistic PI algorithms (e.g., VI, λ-PI) is **not strongly polynomial**.
- Transient MDPs, and average-cost MDPs satisfying a hitting time assumption, can be **reduced to discounted ones**.
- These reductions lead to **alternative algorithms** with attractive complexity estimates.