# Reduction of average-cost Markov Decision Processes to discounting under an accessibility condition


Jefferson Huang

Department of Applied Mathematics and Statistics
Stony Brook University



INFORMS Annual Meeting
November 10, 2014


Joint work with Eugene Feinberg

# Plan of the talk

1. Definitions
2. Review: complexity of discounted MDPs
3. Review: complexity of average-cost MDPs
4. Reducing average-cost MDPs to discounting
   - Complexity of policy iteration
   - Existence of optimal policies - infinite state spaces

# **Definitions:** The model

**Markov Decision Process (MDP)**: defined by $(\mathbb{X}, A(\cdot), p, c)$ where

1. $\mathbb{X}$ - state space
2. $A(x)$ - sets of actions available at $x \in \mathbb{X}$
3. $p(y|x, a)$ - transition probabilities, where
   - $x$ - current state
   - $a$ - current action
   - $y$ - next state
4. $c(x, a)$ - one-step costs

**Assume:** $\mathbb{X}$ is **discrete**, $A(x)$ is **finite** $\forall x \in \mathbb{X}$.

# **Definitions:** Policies

Policy $\pi$ - history-dependent and randomized in general.

- $\Pi :=$ set of all policies.

**Stationary policy** $\phi$: selects action $\phi(x) \in A(x)$ whenever the state is $x \in \mathbb{X}$.

- $\mathbb{F} :=$ set of all stationary policies.

For $\pi \in \Pi$ & initial $x \in \mathbb{X}$, the **average cost** is

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

for $\beta \in [0,1)$ the $\beta$-**discounted cost** is

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

# **Definitions:** Optimality

$\pi_* \in \Pi$ is **average-cost optimal** if

$$w^{\pi_*}(x) = \inf_{\pi \in \Pi} w^{\pi}(x) \quad \forall x \in \mathbb{X}$$

and $\beta$-**discount optimal** if

$$v_\beta^{\pi_*}(x) = \inf_{\pi \in \Pi} v_\beta^{\pi}(x) \quad \forall x \in \mathbb{X}.$$

*Main questions*:

1. When do (stationary) optimal policies exist?
2. How can optimal policies be computed (and how quickly)?

# Computing optimal policies

*Main methods*:

1. Value Iteration
   - discounted: Shapley (1953)
   - undiscounted: Bellman (1957)
   - average-cost: White (1963)
2. Policy Iteration
   - discounted & average-cost: Howard (1960)
3. Linear Programming Algorithms via LP formulation
   - discounted: D'Epenoux (1963)
   - average-cost: de Ghellinck (1960) and Manne (1960); Denardo and Fox (1968), Hordijk and Kallenberg (1979, 1980), Kallenberg (1983)
   - Wolfe and Dantzig (1962)

**Focus of talk**: Policy Iteration & Simplex Method

# **Definitions:** Complexity of algorithms

$m :=$ 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.

# Complexity results: Discounted costs (fixed $\beta$)

**Value Iteration**:

- weakly polynomial - Tseng (1990)
- not strongly polynomial - Feinberg and H. (2014)

**Howard's Policy Iteration**:

- weakly polynomial - Meister and Holzbaur (1986)
- strongly polynomial - Ye (2011), sharper bounds by Hansen, Miltersen, and Zwick (2013) and Scherrer (2013)

**LP Algorithms**:

- weakly polynomial - Khachiyan (1979) (ellipsoid), Karmarkar (1984) (interior point)
- strongly polynomial - Ye (2005) (interior point); Ye (2011) (simplex + Dantzig's rule), sharper bound by Scherrer (2013)

Many **modified policy iteration** algorithms are not strongly polynomial - Feinberg, H., and Scherrer (2014).

- Puterman and Shin's (1978) algorithm, Bertsekas and Tsitsiklis's (1996) $\lambda$-policy iteration

# **Complexity results:** Discounted costs - particular models

**Simplex**: **strongly polynomial**, regardless of $\beta$, for:

- **deterministic MDPs** - Post and Ye (2013) (Dantzig's rule), sharper bound by Hansen, Kaplan, and Zwick (2014)
- **controlled random walks** (e.g. M/M/1 queues) - Zadorojniy, Even, and Schwartz (2009) & Even and Zadorojniy (2012) (Gass-Saaty rule)

# **Complexity results:** Average costs - particular models

**Simplex**: **strongly polynomial** for

- controlled random walks - Zadorojniy, Even and Schwartz (2009), Even and Zadorojniy (2012) (Gass-Saaty rule)
- problems with a state $\ell$ that's reached under any action with probability at least $\alpha > 0$ - Feinberg and H. (2013) (Dantzig's rule)

**Howard's policy iteration**: **strongly polynomial** for

- problems with a state $\ell$ that's reached under any action with probability at least $\alpha > 0$ - Feinberg and H. (2013)
- problems where the hitting time to a state $\ell$ is uniformly bounded in starting state & policy - Akian and Gaubert (2013)
  - shown for Hoffman and Karp's (1966) algorithm for mean-payoff games

# Methods for studying complexity of average costs

Two approaches:

1. New algorithms - Zadorojniy, Even, and Schwartz (2009)
2. **Reduction** to discounted problem
   - Feinberg and H. (2013) - Ross's (1968a,b) transformation
   - Akian and Gaubert (2013) - non-linear Perron-Frobenius theory

**Akian and Gaubert's (2013) transformation**: generalization of Ross's (1968a,b) transformation.

- see also Gubenko and Štatland (1975), Dynkin and Yushkevich (1979).

# Rest of the talk

1. Sufficient conditions for & implications of Akian and Gaubert's (2013) hitting time assumption;
2. Their reduction for MDPs without non-linear Perron-Frobenius theory;
3. Infinite $\mathbb{X}$ - obtaining existence of a stationary optimal policy.

# The assumption

$\ell \in \mathbb{X}$ - fixed state

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

**Assumption HT (Hitting Time)**

There's a constant $K$ where

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

**Equivalent:** $\exists$ *bounded nonnegative function* $\xi$ on $\mathbb{X}$ satisfying

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

# Sufficient condition for Assumption HT

## Assumption D

There's a positive integer $N$ & constant $\alpha$ where

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

- Special case of Hordijk's (1974) *simultaneous Doeblin condition*.
- Implies
$$\mathbb{E}_x^\phi \tau_\ell \leq N/\alpha < \infty \quad \forall x \in \mathbb{X}, \phi \in \mathbb{F}.$$
- Ross's (1968a,b) assumption: $N = 1$.

# Implications of Assumption HT

$P(\phi) :=$ Markov chain corresponding to $\phi \in \mathbb{F}$

- state $\ell$ is *positive recurrent* $\forall \phi \in \mathbb{F}$.
- MDP is **unichain**, i.e. $P(\phi)$ has a single recurrent class $\forall \phi \in \mathbb{F}$.
- If $P(\phi)$ is aperiodic $\forall \phi \in \mathbb{F}$,
  - each $P(\phi)$ has a stationary distribution $\pi(\phi)$;
  - each $P(\phi)$ is **fast mixing** - $\exists$ positive integer $N$ and $\rho < 1$ where

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

    see Federgruen, Hordijk, and Tijms (1978).
- average cost $w^{\phi}$ is **constant** $\forall \phi \in \mathbb{F}$.

# Reduction to discounting under Assumption HT

$\xi$ - bounded nonnegative function satisfying

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

$K$ - upper bound for $\xi$

**Step 1:** Use $\xi$ to construct MDP with **state-dependent discount** factors

$$\frac{\xi(x) - 1}{\xi(x)}, \quad x \in \mathbb{X}.$$

**Step 2:** Construct MDP with **uniform discount** factor

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

## **Step 1:** State-dependent discounting - Akian and Gaubert (2013)

1. State space $\mathbb{X}$
2. Action sets $A(x)$, $x \in \mathbb{X}$
3. Transition probabilities

$$p_\xi(y|x,a) := \begin{cases} \frac{1}{\xi(x)-1} p(y|x,a)\xi(y), & y \neq \ell, \\ 1 - \frac{1}{\xi(x)-1} \sum_{y \neq \ell} p(y|x,a)\xi(y), & y = \ell \end{cases}$$

4. One-step costs $c_\xi(x,a) := c(x,a)/\xi(x)$
5. Current state is $x \implies$ next period's cost discounted by

$$\gamma_\xi(x) := \frac{\xi(x)-1}{\xi(x)}$$

# **Step 2:** Uniform discounting - Feinberg (2002)

*"Grave state"* - $\bar{x} \notin \mathbb{X}$

1. State space $\bar{\mathbb{X}} := \mathbb{X} \cup \{\bar{x}\}$
2. Action sets
$$\bar{A}(x) := \begin{cases} A(x), & x \in \mathbb{X} \\ \{\bar{a}\}, & x = \bar{x} \end{cases}$$
3. Transition probabilities
$$\bar{p}(y|x,a) := \begin{cases} \frac{\gamma(x)}{\beta} p_\xi(y|x,a), & x, y \in \mathbb{X} \\ 1 - \frac{\gamma(x)}{\beta}, & x \in \mathbb{X}, y = \bar{x} \\ 1, & x = y = \bar{x} \end{cases}$$
4. One-step costs
$$\bar{c}(x,a) := \begin{cases} c_\xi(x,a), & x \in \mathbb{X} \\ 0, & x = \bar{x} \end{cases}$$
5. Discount factor $\beta = (K-1)/K$

# Howard's & simple policy iteration

Policy iteration (both discounted and average-cost):

0. Select $\phi \in \mathbb{F}$.
1. Evaluate $\phi$.
2. Improve $\phi$ if possible and go to step 1; otherwise $\phi$ is optimal.

Improvement rule $\iff$ simplex pivoting rule for LP formulation.

$\rho_{xa}$ - decision variables, $a \in A(x)$, $x \in \mathbb{X}$.

**Howard's policy iteration:** For each $x$, variable $\rho_{xa}$ with most negative reduced cost enters the basis. (**block pivoting**)

**Simple policy iteration:** Variable $\rho_{xa}$ with most negative reduced cost enters basis. (**Dantzig's rule**)

# Correspondence of policy iterations

**Lemma**

*A sequence of policies is generated by discounted policy iteration for the MDP* $(\bar{\mathbb{X}}, \bar{A}(\cdot), \bar{p}, \bar{c})$ *with discount factor* $\beta = (K-1)/K$

*if and only if*

*that sequence is generated by average-cost policy iteration for the MDP* $(\mathbb{X}, A(\cdot), p, c)$*.*

**Idea:**

- Write evaluation and improvement steps for $(\bar{\mathbb{X}}, \bar{A}(\cdot), \bar{p}, \bar{c})$ in terms of $\xi$ and the original transition probabilities & costs.
- Use the uniqueness of the solutions obtained in the evaluation step for policy iterations under both criteria.

# **Complexity estimates:** Average-cost policy iterations

**Theorem**

*If Assumption HT holds, then for average costs Howard's policy iteration needs*

$$O\left(m \cdot K \log K\right)$$

*iterations, and simple policy iteration needs*

$$O(nm \cdot K \log K)$$

*iterations.*

Theorem follows from the Lemma and Scherrer's (2013) iteration bounds for Howard's and simple policy iteration.

# **Complexity estimates:** Average-cost policy iterations

## Assumption D

There's a positive integer $N$ & constant $\alpha$ where

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

## Corollary

*If Assumption D holds, then for average costs Howard's policy iteration needs*

$$O(m \cdot (N/\alpha) \log(N/\alpha))$$

*iterations, and simple policy iteration needs*

$$O(nm \cdot (N/\alpha) \log(N/\alpha))$$

*iterations.*

For $N = 1$, Corollary was proved by Feinberg and H. (2013).

# Existence of stationary optimal policies: Infinite $\mathbb{X}$

Bounded one-step costs $c \not\Rightarrow$ average-cost optimal policy exists when state space $\mathbb{X}$ is countably infinite.

- Ross (1970)

**Theorem**

*If $c$ is bounded, and Assumption HT holds, then there's a stationary average-cost optimal policy.*

- Theorem follows from Akian and Gaubert's (2013) reduction.
- Theorem was proved by Federgruen and Tijms (1978) using a different method.
- Theorem follows from a much more general result covering uncountable state spaces, noncompact action sets, and possibly no special state $\ell$, proved by Feinberg, Kasyanov, and Zadoianchuk (2012).

## Existence of stationary optimal policies: Infinite $\mathbb{X}$

**Idea:** Obtain a *bounded* solution $(g, h)$ to the average-cost optimality equation

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

(Derman (1966) showed this suffices) by showing that

$$T_\xi v(x) :=$$

$$\min_{A(x)} \left[ \frac{c(x, a)}{\xi(x)} + \frac{1}{\xi(x)} \sum_{y \in \mathbb{X}} p(y|x, a) \xi(y)(v(y) - v(\ell)) + \frac{\xi(x) - 1}{\xi(x)} v(\ell) \right]$$

is a contraction mapping on the space of bounded functions on $\mathbb{X}$.

# Summary

1. Akian and Gaubert (2013) proposed a new reduction of mean-payoff games to discounted games.
2. For MDPs, the complexity results it implies can be proved without non-linear Perron-Frobenius theory.
3. It can also be used to verify the existence of stationary optimal policies.