On the reduction of total cost and average cost MDPs to discounted MDPs

Jefferson Huang

School of Operations Research and Information Engineering
Cornell University

July 12, 2017

INFORMS Applied Probability Society Conference

Northwestern University

Evanston, IL

Based on joint work with Eugene A. Feinberg (Stony Brook University)

Overview

- Discounted MDPs are typically easier to study than undiscounted ones.
  - No need to consider structure of Markov chains induced by stationary policies.
  - Study of optimality equations, existence of optimal policies, and algorithms is often more straightforward.
- Early approach: reduce the undiscounted problem to a discounted one [Ross 1968], [Gubenko, Štatland 1975], [Dynkin, Yushkevich 1979]

This Talk: Most general known conditions under which undiscounted MDPs can be reduced to discounted ones.

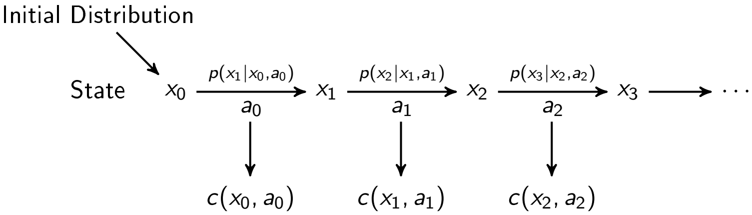
Model Description

$\mathbb{X}$ = state space; $n := |\mathbb{X}| \leqslant |\mathbb{N}|$ (will remark on uncountable case)

$A(x)$ = action space; $m := |\cup_{x \in \mathbb{X}} A(x)| \leqslant |\mathbb{R}|$

$p(y|x, a)$ = probability that the next state is y, given the current state is x and action a is taken

$c(x, a)$ = cost incurred when current state is x and action a is taken



Super-Stochastic Transition Rates

We will consider "transition rates"

$$q(y|x,a) := \alpha(x,a)p(y|x,a), \qquad \alpha(x,a) \geqslant 0.$$

Why?

- Generalization of usual discounted MDPs (constant $\alpha < 1$)
  - **This talk:** Conditions under which general discounting can be reduced to usual discounting.
- Studied by many authors since the 1960s, e.g., [Veinott 1969], [Sondik 1974], [Rothblum 1975], [Pliska 1976, 1978], [Rothblum, Veinott 1982], [Hordijk, Kallenberg 1984], [Hinderer, Waldmann 2003, 2005], [Eaves, Veinott 2014]
  - Also called e.g., "Markov branching decision chains", "Markov population decision chains"
- Applications to controlled population processes, infinite particle systems, marketing, pest eradication, multiarmed bandits with risk-seeking criteria, stochastic shortest path problems, ...

Policies

Policy = rule determining which action to take at each time step

This Talk: *deterministic stationary* policies only

- i.e., mappings ϕ on $\mathbb{X}$ where $\phi(x) \in A(x)$ for all $x \in \mathbb{X}$
- no loss of generality (wrt. randomized history-dependent policies) for models considered

Compare policies via a cost criterion $g(\phi) \in \mathbb{R}^n$

- ϕ_* is optimal if $g(\phi_*) \leqslant g(\phi)$ (component-wise) for all policies ϕ

For each policy ϕ, let

$$Q(\phi)_{x,y} := q(y|x, \phi(x)), \quad c(\phi)_x := c(x, \phi(x)).$$

Optimality Criteria

Total-Cost Criterion: For each state x,

$$g(\phi) = v(\phi) := \sum_{t=0}^{\infty} Q(\phi)^t c(\phi)$$

Average-Cost Criterion: For each state x,

$$g(\phi) = w(\phi) := \limsup_{T \to \infty} \frac{1}{T} \sum_{n=0}^{T-1} Q(\phi)^t c(\phi)$$

Complexity Estimates

An MDP is solved by computing an optimal policy.

An algorithm solves an MDP (with finite state & action sets) in strongly polynomial time if the # of arithmetic operations needed can be bounded above by a polynomial in the # of state-action pairs m.

If the # of arithmetic operations needed can be bounded above by a polynomial in m and the total bit-size of the input data, it solves the MDP in weakly polynomial time.

- Total-cost & average-cost MDPs can be formulated as linear programs $\Longrightarrow$ solvable in weakly polynomial time [Khachiyan 1979], [Karmarkar 1984]

Total-Cost MDPs: Transience Assumption

$\|Q(\phi)\|_V := \sup_{x\in\mathbb{X}} V(x)^{-1} \sum_{y\in\mathbb{X}} q(y|x,\phi(x))V(y), \qquad V \geqslant 1$

Assumption (Transience)

There is a constant K such that, for every policy ϕ,

$$\left\| \sum_{t=0}^{\infty} Q(\phi)^t \right\|_V \leqslant K < \infty.$$

- "Lifetime" of the process initiated at state x is bounded by $KV(x)$ under every policy.

[Veinott 1974]: Transience can be checked in strongly polynomial time.

Characterization of Transience

Theorem (Feinberg & H, 2017)

Transience holds if and only if there is a function $\mu : \mathbb{X} \to [1, K]$ *where*

$$\mu(x) \geqslant V(x) + \sum_{y \in \mathbb{X}} q(y|x, a)\mu(y)$$

for all $a \in A(x)$ *and* $x \in \mathbb{X}$.

E.g., let

$$\mu = \sup_{\phi} \left\{ \sum_{t=0}^{\infty} Q(\phi)^t V \right\}.$$

[Denardo 2016]: Such a μ can be computed using $O[(n^3 + mn)mK \log K]$ arithmetic operations.

Hoffman-Veinott (HV) Transformation

$\tilde{\beta} := (K-1)/K$

$\tilde{\mathbb{X}} := \mathbb{X} \cup \{\tilde{x}\}$, and $\tilde{\mathbb{A}} := \mathbb{A} \cup \{\tilde{a}\}$

$\tilde{A}(x) := A(x)$ if $x \in \mathbb{X}$ and $\tilde{A}(\tilde{x}) := \{\tilde{a}\}$.

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

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

$$\tilde{v}_{\tilde{\beta}}(\phi)_x := \tilde{\mathbb{E}}_x^{\phi} \sum_{t=0}^{\infty} \tilde{\beta}^t \tilde{c}(x_t, a_t) \qquad x \in \mathbb{X},\ \phi \in \mathbb{F}$$

Reduction to a Discounted MDP

Theorem (Feinberg & H, 2017)

Suppose transience holds, and that there is a constant $\overline{c} < \infty$ satisfying

$$|c(x,a)| \leqslant \overline{c}V(x) \qquad \forall x \in \mathbb{X},\ a \in A(x).$$

Then

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

Proof. Let $\tilde{c}_{\phi}(x) := \tilde{c}(x, \phi(x))$ and $\tilde{P}_{\phi}(x,y) := \tilde{p}(y|x, \phi(x))$. Then

$$\tilde{\beta}^{n}\tilde{P}^{t}_{\phi}\tilde{c}_{\phi}(x) = \mu(x)^{-1}Q^{t}_{\phi}c_{\phi}(x) \qquad \forall t \in \{0, 1, \dots\}$$

□

Implies that to minimize v^{ϕ}, it suffices to minimize $\tilde{v}^{\phi}_{\tilde{\beta}}$.

Leads to results on validity of optimality equation and existence and characterization of optimal policies for the original MDP [Feinberg & H, 2017].

Linear Programming Formulation

The new discounted MDP leads to the following LP.

$$
\begin{aligned}
\text{minimize} \quad & \sum_{x \in \mathbb{X}} \sum_{a \in A(x)} \frac{c(x,a)}{\mu(x)} z_{x,a} \\
\text{such that} \quad & \sum_{a \in A(x)} z_{x,a} - \sum_{x' \in \mathbb{X}} \sum_{a' \in A(x')} \frac{p(x|x',a')\mu(x)}{\mu(x')} z_{x',a'} = 1, \quad x \in \mathbb{X} \\
& z_{x,a} \geqslant 0, \quad a \in A(x),\ x \in \mathbb{X}
\end{aligned}
$$

For an optimal basic feasible solution z^*, let

$$
\phi_*(x) = \arg\max_{a \in A(x)} \left\{ z^*_{x,a} \right\}, \quad x \in \mathbb{X}.
$$

Theorem (Feinberg & H, 2017)

ϕ_* *is optimal under the total-cost criterion.*

Complexity Estimate

Theorem (Feinberg & H, 2017)

The simplex method with Dantzig's rule solves the linear program (LP) using at most

$$O(nmK \log K) \quad \textit{iterations.}$$

Also, there is a block-pivoting simplex method that solves the LP using at most

$$O(mK \log K) \quad \textit{iterations.}$$

- Via results for discounted MDPs [Scherrer 2016].
- Each iteration of the simplex method needs $O(n^3 + nm)$ arithmetic operations.
- When K is fixed, these two algorithms solve total-cost MDPs in strongly polynomial time.
- [Denardo 2016]: similar estimates, using different proof technique

Interlude: Complexity of Discounted MDPs

Discounted MDPs with a fixed discount factor are solvable in strongly polynomial time.

- [Ye 2005]: Interior-point method
- [Ye 2011], [Scherrer 2016]: simplex method with Dantzig's rule, Howard's (1960) policy iteration method
- [Hansen, Miltersen, Zwick 2013] Extension to strategy iteration for zero-sum perfect-information stochastic games

[Hollanders, Delvenne, Jungers 2012]: If discount factor isn't fixed, Howard's (1960) policy iteration may need exponential time.

[Feinberg, H 2014], [Feinberg, H, Scherrer 2014]: Modified policy iteration algorithms (e.g., value iteration, λ-policy iteration) are not strongly polynomial

Discounted MDPs with special structure can be solved in strongly polynomial time (regardless of discount factor)

- [Zadorojniy, Even, Shwartz 2009]: controlled random walks
- [Post & Ye 2015]: deterministic MDPs

Uncountable State Spaces

Need to deal with measurability and continuity issues.

- **Measurability** of new cost function $\tilde{c}$ and transition probabilities $\tilde{p}$
  - Depends on measurability of μ
  - In general, costs and transition probabilities may only be universally measurable

- **Continuity** of new cost function $\tilde{c}$ and transition probabilities $\tilde{p}$
  - Depends on continuity of μ
  - Related to existence of stationary optimal policies
  - $\tilde{c}$: semicontinuity
  - $\tilde{p}$: setwise/weak continuity

See [Feinberg & H, 2017] for details. Also relevant in the average-cost case (Slide 22).

Average-Cost MDPs: Hitting Time Assumption

$$ {}_\ell Q(\phi)_{x,y} = \begin{cases} q(y|x,\phi(x)), & y \neq \ell \\ 0, & y = \ell \end{cases} $$

Assumption (Hitting Time)

There is a state ℓ and a constant L such that, for every policy ϕ,

$$ \left\| \sum_{t=0}^{\infty} {}_\ell Q(\phi)^t \right\|_{\mathbf{1}} \leqslant L < \infty. $$

- If $q \leqslant 1$, mean recurrence time to state ℓ is bounded by L under every policy.
  - ℓ may be e.g., failed state of machine, no customers in queue
- Every such MDP is unichain.

[Feinberg & Yang 2008]: can be checked in strongly polynomial time

An Equivalent Condition

Theorem (Feinberg & H, 2017)

The hitting time assumption holds if and only if there is a function $\mu_\ell : \mathbb{X} \to [0, L]$ *satisfying*

$$\mu_\ell(x) \geqslant 1 + \sum_{y \neq \ell} q(y|x, a)\mu_\ell(y)$$

for all $a \in A(x)$ *and* $x \in \mathbb{X}$.

E.g., let

$$\mu_\ell = \sup_\phi \left\{ \sum_{t=0}^{\infty} {}_\ell Q(\phi)^t \mathbf{1} \right\}$$

where $\mathbf{1}_x = 1$ for all $x \in \mathbb{X}$.

[Denardo 2016]: Such a μ can be computed using at most $O[(n^3 + mn)mL \log L]$ arithmetic operations.

HV-AG (Akian-Gaubert) Transformation

$\bar{\beta} := (L-1)/L$

$\bar{\mathbb{X}} := \mathbb{X} \cup \{\bar{x}\}$, and $\bar{\mathbb{A}} := \mathbb{A} \cup \{\bar{a}\}$

$\bar{A}(x) := A(x)$ if $x \in \mathbb{X}$ and $\bar{A}(\bar{x}) := \{\bar{a}\}$.

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

$$\bar{c}(x,a) := \begin{cases} c(x,a)/\mu_\ell(x), & x \in \mathbb{X}, a \in A(x), \\ 0, & (x,a) = (\bar{x},\bar{a}). \end{cases}$$

$$\bar{v}^{\phi}_{\bar{\beta}}(x) := \bar{\mathbb{E}}^{\phi}_{x} \sum_{t=0}^{\infty} \bar{\beta}^t \bar{c}(x_t, a_t) \qquad x \in \bar{\mathbb{X}},\ \phi \in \mathbb{F}.$$

Reduction to a Discounted MDP

Note: $\bar{p}$ are transition *probabilities*.

Theorem (Feinberg & H)

Suppose the hitting time assumption holds, that $\sum_{y \in \mathbb{X}} q(y|x,a) = 1$ *for all* $x \in \mathbb{X}$ *and* $a \in A(x)$, *and that the constant* $\overline{c} < \infty$ *satisfies*

$$|c(x,a)| \leqslant \overline{c} V(x) \qquad \forall x \in \mathbb{X},\ a \in A(x).$$

Then

$$w^{\phi}(x) = \bar{v}_{\bar{\beta}}^{\phi}(\ell) \qquad \forall x \in \mathbb{X},\ \phi \in \mathbb{F}.$$

Proof. Show that for every ϕ, the function $h^{\phi}(x) := \mu(x)[\bar{v}_{\bar{\beta}}(x) - \bar{v}_{\bar{\beta}}(\ell)]$ satisfies

$$\bar{v}_{\bar{\beta}}^{\phi}(\ell) + h^{\phi}(x) = c_{\phi}(x) + Q_{\phi} h^{\phi}(x) \qquad \forall x \in \mathbb{X},$$

and that

$$\lim_{T \to \infty} \frac{1}{T} Q_{\phi}^{T} h^{\phi}(x) = 0.$$

□

Used to verify validity of the average-cost optimality equation and the existence of stationary optimal policies [Feinberg & H, 2017].

Linear Programming Formulation

An LP is obtained from the new discounted MDP:

$$\text{minimize} \quad \sum_{x\in\mathbb{X}} \sum_{a\in A(x)} \frac{c(x,a)}{\mu_\ell(x)} z_{x,a}$$

$$\text{such that} \quad \sum_{a\in A(x)} z_{x,a} - \sum_{x'\in\mathbb{X}} \sum_{a'\in A(x')} \frac{p(x|x',a')}{\mu_\ell(x')} \mu_\ell(x) z_{x',a'} = 1, \quad x \neq \ell$$

$$\sum_{a\in A(\ell)} z_{\ell,a} - \sum_{x'\in\mathbb{X}} \sum_{a'\in A(x')} \frac{\mu_\ell(x') - 1 - \sum_{y\neq\ell} p(y|x',a')\mu_\ell(y)}{\mu_\ell(x')} z_{x',a'} = 1$$

$$z_{x,a} \geqslant 0, \quad a \in A(x), \ x \in \mathbb{X}$$

For an optimal basic feasible solution z^*, let

$$\phi_*(x) = \arg\max_{a\in A(x)} \left\{ z^*_{x,a} \right\}, \quad x \in \mathbb{X}.$$

Theorem

ϕ_* *is optimal under the average-cost criterion.*

Complexity Estimate

Theorem (Feinberg & H, 2017)

The simplex method with Dantzig's rule solves the linear program (LP) using at most

$$O(nmL\log L) \quad \textit{iterations.}$$

Also, there is a block-pivoting simplex method that solves the LP using at most

$$O(mL\log L) \quad \textit{iterations.}$$

- Via results for discounted MDPs Scherrer 2016].
- Each iteration of the simplex method needs $O(n^3 + nm)$ arithmetic operations.
- When L is fixed, these two algorithms are strongly polynomial for average-cost MDPs.
- Result for block-pivoting is special case of result in [Akian & Gaubert 2013] for 2-player stochastic games.

Complexity of Average-Cost MDPs

Average-cost MDPs with special structure are solvable in strongly polynomial time.

- [Zadorojniy, Even, Shwartz 2009]: controlled random walk
- [Feinberg, H 2013]: replacement/maintenance problems with fixed minimal failure probability
  - [Feinberg, H 2017]: fixed upper bound on expected time to failure

[Fearnley 2010]: Howard's (1960) policy iteration may need exponential time to solve a multichain average-cost MDP.

- Not known if this is true when MDP is unichain.

[Tsitsiklis 2007]: Checking whether an MDP is unichain is NP-complete.

- Our hitting time assumption can be checked in strongly polynomial time [Feinberg, Yang 2008].

Extension to Uncountable State Spaces

- Similar issues as in the total cost case (Slide 14).
- For weak continuity of transition probabilities, the state ℓ may need to be *isolated* from $\mathbb{X}$ (i.e., the singleton $\{\ell\}$ is both open and closed)

See [H, 2016] for details.

Conclusion

This Talk:

1. Conditions under which undiscounted MDPs can be reduced to discounted ones.
   - Total Costs: Transience
   - Average Costs: Recurrence
2. Lead to validity of optimality equations, existence of optimal policies, and complexity estimates for computing optimal policies.

Questions/Extensions:

- Consequences for specific models? (e.g., queueing control, replacement & maintenance) [Feinberg, H 2013]
- More general conditions under which a reduction holds?
  - Complexity estimates for average-cost problems
- N-player stochastic games?