Reductions Of Undiscounted Markov Decision Processes and Stochastic Games To Discounted Ones

Jefferson Huang

School of Operations Research and Information Engineering

Cornell University

November 16, 2016

INFORMS Annual Meeting
Nashville, TN

Joint work with Eugene A. Feinberg (Stony Brook University)

What This Talk is About

- **Transformations** of certain undiscounted generalized two-player zero-sum stochastic games to discounted ones.
  - Undiscounted = Total or Average Costs
  - Generalized = possibly super-stochastic transition rates
  - Transition rates for the resulting discounted game are stochastic, and the one-step costs are bounded
  - General (e.g., uncountable) state and action sets
  - Special case: Markov decision processes (MDPs)
- Conditions under which these transformations lead to **reductions** of the original undiscounted problem to a discounted one.
  - Lead to results on the existence of of ϵ-optimal policies, validity of optimality equations, computational complexity estimates ...

Why?

- Discounted stochastic games are **much easier to study** than undiscounted ones!
  - Shapley's (1953) seminal paper was on the discounted case.
  - Relevant issues when costs are undiscounted:
    - Total costs: summability, convergence of value iteration
    - Average costs: structure of Markov chains induced by stationary policies
- Discounting total costs in the original model may not be desirable.
  - Discounting means we don't care about the system's behavior in the long run.
  - Costs may not have a clear economic interpretation.
- Super-stochastic transition rates are relevant to applications.
  - controlled branching processes, multi-armed bandits with risk-seeking utilities, discount factors greater than one ...

Plan of the Talk

1. Definition of generalized two-player zero-sum stochastic games, which include as special cases:
   - MDPs (one of the players can't do anything);
   - robust MDPs (see e.g., [Iyengar, 2005]).
2. Transformations of such games to discounted ones.
   - Motivated by [Veinott 1969] and [Akian Gaubert 2013].
3. Results about the original game that follow from the transformations.

Generalized Two-Player Zero-Sum Stochastic Games

Defined by *5 objects*:

1. *state* space $\mathbb{X}$
2. *action* spaces $\mathbb{A}^1, \mathbb{A}^2$ for players 1 and 2
3. for each $x \in \mathbb{X}$, sets of *available actions* $A^1(x) \subseteq \mathbb{A}^1$ and $A^2(x) \subseteq \mathbb{A}^2$ for players 1 and 2
4. for each state $x \in \mathbb{X}$ and pair of actions $(a^1, a^2) \in A^1(x) \times A^2(x)$,
   - *transition rates* $q(\cdot|x, a^1, a^2)$;
   - *one-step costs* $c(x, a^1, a^2)$.

For experts: $\mathbb{X}$, $\mathbb{A}^1$, $\mathbb{A}^2$ are Borel subsets of Polish spaces, for all $x \in \mathbb{X}$ the sets $A^1(x)$ and $A^2(x)$ are measurable, the graph of $A^1 \times A^2$ is Borel-measurable, q is a Borel-measurable transition kernel, and c is Borel-measurable.

Cost Criteria

$\Pi^1, \Pi^2 =$ set of all (randomized history-dependent) policies for players 1, 2.

For $x \in \mathbb{X}$ and $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$, let $\mathbb{E}_x^{\pi^1, \pi^2}$ denote the corresponding "expectation" operator.

For experts: $\mathbb{E}_x^{\pi^1, \pi^2}$ can be defined via the the usual definition of randomized history-dependent policies and the Ionescu-Tulcea theorem

Total cost: For $\beta \in [0, 1]$,

$$v_\beta^{\pi^1, \pi^2}(x) := \mathbb{E}_x^{\pi^1, \pi^2} \sum_{t=0}^{\infty} \beta^t c(x_t, a_t^1, a_t^2), \qquad x \in \mathbb{X}$$

and $v^{\pi^1, \pi^2} := v_1^{\pi^1, \pi^2}$.

Average cost:

$$w^{\pi^1, \pi^2}(x) := \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_x^{\pi^1, \pi^2} \sum_{t=0}^{T-1} c(x_t, a_t^1, a_t^2), \qquad x \in \mathbb{X}$$

Optimality Criteria

Player 1 wants to maximize cost, player 2 wants to minimize cost.

Consider a criterion $g \in \{v, w\}$, and $\epsilon \geq 0$.

$\pi^1_* \in \Pi^1$ is **ϵ-optimal for player 1** if

$$\inf_{\pi^2 \in \Pi^2} g^{\pi^1_*, \pi^2}(x) \geq \inf_{\pi^2 \in \Pi^2} \sup_{\pi^1 \in \Pi^1} g^{\pi^1, \pi^2}(x) - \epsilon \qquad \forall x \in \mathbb{X}.$$

$\pi^2_* \in \Pi^2$ is **ϵ-optimal for player 2** if

$$\sup_{\pi^1 \in \Pi^2} g^{\pi^1, \pi^2_*}(x) \leq \sup_{\pi^1 \in \Pi^1} \inf_{\pi^2 \in \Pi^2} g^{\pi^1, \pi^2}(x) + \epsilon \qquad \forall x \in \mathbb{X}.$$

0-optimal policies are called **optimal**.

Some Useful Definitions...

Let $\mathbb{F}^1, \mathbb{F}^2$ denote the set of all *deterministic stationary policies* for players 1,2.

Given $(\phi^1, \phi^2) \in \mathbb{F}^1 \times \mathbb{F}^2$, a Borel subset B of $\mathbb{X}$, and a Borel-measurable $u : \mathbb{X} \to \mathbb{R}$, let

$$ {}_B Q_{\phi^1,\phi^2} u(x) := \int_{\mathbb{X}\setminus B} u(y) q(dy | x, \phi^1(x), \phi^2(x)), \qquad x \in \mathbb{X}, $$

for $x \in \mathbb{X}$ let ${}_x Q_{\phi^1,\phi^2} := {}_{\{x\}} Q_{\phi^1,\phi^2}$, and let $Q_{\phi^1,\phi^2} := {}_\emptyset Q_{\phi^1,\phi^2}$.

For a *weight function* $W : \mathbb{X} \to \mathbb{R}$, given a transition kernel $B(\cdot|\cdot)$ from $\mathbb{X}$ to $\mathbb{X}$, let

$$ \|B\|_W := \sup_{x \in \mathbb{X}} W(x)^{-1} \int_{\mathbb{X}} W(y) B(dy|x). $$

Transience Assumption (for Total Costs)

Assumption (T)

There is a weight function $V : \mathbb{X} \to [1, \infty)$ such that

(i) $\| \sum_{t=0}^{\infty} Q^t_{\phi^1,\phi^2} \|_V \leq K$ for all $(\phi^1, \phi^2) \in \mathbb{F}^1 \times \mathbb{F}^2$;

(ii) there is a constant $\bar{c}$ satisfying $|c(x, a^1, a^2)| \leq \bar{c}V(x)$ for all $x \in \mathbb{X}$ and $(a^1, a^2) \in A^1(x) \times A^2(x)$;

(iii) for every $x \in \mathbb{X}$ the mapping

$$(a^1, a^2) \mapsto \int_{\mathbb{X}} V(y) q(dy|x, a^1, a^2)$$

is continuous.

Assumption (T)(i) is *equivalent* to the existence of a function μ that is upper semianalytic satisfying $V \leq \mu \leq KV$ and

$$\mu(x) \geq V(x) + \int_{\mathbb{X}} \mu(y) q(dy|x, a^1, a^2)$$

for all $x \in \mathbb{X}$,$(a^1, a^2) \in A^1(x) \times A^2(x)$.

Hitting Time Assumption (for Average Costs)

Assumption (HT)

There is a weight function $V : \mathbb{X} \to [1, \infty)$ such that

(i) $\|\sum_{t=0}^{\infty} {}_{\ell}Q^t_{\phi^1,\phi^2}\|_V \leq K < \infty$ for all $(\phi^1, \phi^2) \in \mathbb{F}^1 \times \mathbb{F}^2$;

(ii) there is a constant $\bar{c}$ satisfying $|c(x, a^1, a^2)| \leq \bar{c}V(x)$ for all $x \in \mathbb{X}$ and $(a^1, a^2) \in A^1(x) \times A^2(x)$;

(iii) for every $x \in \mathbb{X}$ the mapping

$$(a^1, a^2) \mapsto \int_{\mathbb{X}\setminus\{\ell\}} V(y)q(dy|x, a^1, a^2)$$

is continuous.

Assumption (HT)(i) is *equivalent* to the existence of a function μ that is upper semianalytic satisfying $V \leq \mu \leq KV$ and

$$\mu(x) \geq V(x) + \int_{\mathbb{X}\setminus\{\ell\}} \mu(y)q(dy|x, a^1, a^2)$$

for all $x \in \mathbb{X}$,$(a^1, a^2) \in A^1(x) \times A^2(x)$.

Transformation for Total Costs

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

$\tilde{\mathbb{X}} := \mathbb{X} \cup \{\tilde{x}\}$, and $\tilde{\mathbb{A}}^i := \mathbb{A}^i \cup \{\tilde{a}^i\}$ for $i = 1, 2$

For $i = 1, 2$, $\tilde{A}^i(x) := A^i(x)$ if $x \in \mathbb{X}$ and $\tilde{A}^i(\tilde{x}) := \{\tilde{a}\}$.

For Borel sets $B \subseteq \tilde{\mathbb{X}}$,

$$\tilde{p}(B|x, a^1, a^2) := \begin{cases} \frac{1}{\tilde{\beta}\mu(x)} \int_B \mu(y) q(dy|x, a^1, a^2), & B \subseteq \mathbb{X}, x \in \mathbb{X}, (a^1, a^2) \in A^1(x) \times A^2(x), \\ 1 - \frac{1}{\tilde{\beta}\mu(x)} \int_{\mathbb{X}} \mu(y) q(dy|x, a^1, a^2), & B = \{\tilde{x}\}, x \in \mathbb{X}, (a^1, a^2) \in A^1(x) \times A^2(x), \\ 1 & B = \{\tilde{x}\}, (x, a^1, a^2) = (\tilde{x}, \tilde{a}^1, \tilde{a}^2). \end{cases}$$

$$\tilde{c}(x, a^1, a^2) := \begin{cases} c(x, a^1, a^2)/\mu(x), & x \in \mathbb{X}, (a^1, a^2) \in A^1(x) \times A^2(x), \\ 0, & (x, a^1, a^2) = (\tilde{x}, \tilde{a}^1, \tilde{a}^2). \end{cases}$$

Transformation for Average Costs

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

$\bar{\mathbb{X}} := \mathbb{X} \cup \{\bar{x}\}$, and $\bar{\mathbb{A}}^i := \mathbb{A}^i \cup \{\bar{a}^i\}$ for $i = 1, 2$

For $i = 1, 2$, $\bar{A}^i(x) := A^i(x)$ if $x \in \mathbb{X}$ and $\bar{A}^i(\bar{x}) := \{\bar{a}\}$.

For Borel sets $B \subseteq \bar{\mathbb{X}}$,

$$\bar{p}(B|x, a^1, a^2) := \begin{cases} \frac{1}{\bar{\beta}\mu(x)} \int_B \mu(y) q(dy|x, a^1, a^2), & B \subseteq \mathbb{X} \setminus \{\ell\}, x \in \mathbb{X}; \\ \frac{1}{\bar{\beta}\mu(x)} [\mu(x) - 1 - \int_{\mathbb{X}\setminus\{\ell\}} \mu(y) q(dy|x, a^1, a^2)] & B = \{\ell\}, x \in \mathbb{X}; \\ 1 - \frac{1}{\bar{\beta}\mu(x)} [\mu(x) - 1], & B = \{\bar{x}\}, x \in \mathbb{X}; \\ 1 & B = \{\bar{x}\}, (x, a^1, a^2) = (\bar{x}, \bar{a}^1, \bar{a}^2). \end{cases}$$

$$\bar{c}(x, a^1, a^2) := \begin{cases} c(x, a^1, a^2)/\mu(x), & x \in \mathbb{X}, (a^1, a^2) \in A^1(x) \times A^2(x), \\ 0, & (x, a^1, a^2) = (\bar{x}, \bar{a}^1, \bar{a}^2). \end{cases}$$

Results for Undiscounted MDPs

Some types of results that follow from the transformation and results for discounted games, when Assumption (T) holds:

- Existence of a "value" of the game, and of a stationary ϵ-optimal policy for player 1 and optimal stationary policy for player 2, under compactness-continuity assumptions for player (by [Nowak 1985])
- Existence of stationary optimal strategies for both players, under compactness-continuity assumptions for both players (by [Nowak 1984])
- When K is fixed, the state & action sets are finite, and the game has perfect information, a pair of deterministic stationary optimal policies can be computed in strongly polynomial time (by [Hansen Miltersen Zwick 2013]).

For MDPs:

- Validity of optimality equations and characterization of stationary optimal policies (by [Schäl 1993], [Feinberg Kasyanov Zadoianchuk 2012]).
- When K is fixed and the state & action sets are finite, a deterministic stationary optimal policy can be computed in strongly polynomial time (by [Scherrer 2016]).

Summary of the Talk

- Under certain “reachability” conditions, undiscounted stochastic games (and hence MDPs) can be reduced to discounted ones.
- These reductions lead to results about the original undiscounted game.
- In particular, the reductions have implications about the complexity of algorithms for undiscounted game.