Stochastic Games (Part I): Policy Improvement in Discounted (Noncompetitive) Markov Decision Processes

Paul Varkey

Multi Agent Systems Group, Department of Computer Science, UIC

4th Annual Graduate Student Probability Conference
Apr 30th – 2010
Duke University, Durham NC

Outline

1. The Model (definitions, notations and the problem statement)
2. Basic Theorems
3. The Algorithm
4. An Example

References

BLACKWELL, D. (1962): "Discrete Dynamic Programming", *The Annals of Mathematical Statistics*, Vol. 33, No. 2. (Jun. 1962), pp. 719-726.

FILAR, J.A. and VRIEZE, O.J. (1996): "Competitive Markov Decision Processes Theory, Algorithms, and Applications", SpringerVerlag, New York, 1996.

Decision Processes, States and Actions

A decision process is a *discrete stochastic process* that is observed at discrete time points $t = 0, 1, 2, 3, \ldots$ called *stages* where a *decision-maker* (or, *controller*) chooses an action at each *stage*.

- S denotes the state space – the process may be in one of (finitely) many states
- A denotes the action space – the actions may be chosen from one of (finitely) many actions
- Choosing action a in state s results in
  - (i) an immediate reward $r(s, a)$
  - (ii) a probabilistic transition to a state s' given by $p(s'|s, a)$

$\uparrow$

the (stationary) Markov assumption

Decision Processes, States and Actions

A decision process is a *discrete stochastic process* that is observed at discrete time points $t = 0, 1, 2, 3, \ldots$ called *stages* where a *decision-maker* (or, *controller*) chooses an action at each *stage*.

- S denotes the state space – the process may be in one of (finitely) many states
- A denotes the action space – the actions may be chosen from one of (finitely) many actions
- Choosing action a in state s results in
  - (i) an immediate reward $r(s, a)$
  - (ii) a probabilistic transition to a state s' given by $p(s'|s, a)$

$\uparrow$

the (stationary) Markov assumption

Decision Processes, States and Actions

A decision process is a *discrete stochastic process* that is observed at discrete time points $t = 0, 1, 2, 3, \ldots$ called *stages* where a *decision-maker* (or, *controller*) chooses an action at each *stage*.

- S denotes the state space – the process may be in one of (finitely) many states
- A denotes the action space – the actions may be chosen from one of (finitely) many actions
- Choosing action a in state s results in
  - (i) an immediate reward $r(s, a)$
  - (ii) a probabilistic transition to a state s' given by $p(s'|s, a)$

$\uparrow$

the (stationary) Markov assumption

Decision Processes, States and Actions

A decision process is a *discrete stochastic process* that is observed at discrete time points $t = 0, 1, 2, 3, \ldots$ called *stages* where a *decision-maker* (or, *controller*) chooses an action at each *stage*.

- S denotes the state space – the process may be in one of (finitely) many states
- A denotes the action space – the actions may be chosen from one of (finitely) many actions
- Choosing action a in state s results in
  - (i) an immediate reward $r(s, a)$
  - (ii) a probabilistic transition to a state s' given by $p(s'|s, a)$

$\uparrow$

the (stationary) Markov assumption

Decision Processes, States and Actions

A decision process is a *discrete stochastic process* that is observed at discrete time points $t = 0, 1, 2, 3, \ldots$ called *stages* where a *decision-maker* (or, *controller*) chooses an action at each *stage*.

- S denotes the state space – the process may be in one of (finitely) many states
- A denotes the action space – the actions may be chosen from one of (finitely) many actions
- Choosing action a in state s results in
  - (i) an immediate reward $r(s, a)$
  - (ii) a probabilistic transition to a state s' given by $p(s'|s, a)$

$\uparrow$

the (stationary) Markov assumption

Decision Rules, Strategies, Transitions & Rewards

- A decision rule $f : S \longrightarrow A$ is a function that specifies the action that a controller chooses in a given state
- Each decision rule f defines
  - a transition probability matrix $P(f)$ such that the (s, s')-th entry is given by
$$P(f)[s, s'] = p(s'|s, f(s))$$
  - a reward vector $r(f)$ such that the (s)-th entry is given by
$$r(f)[s] = r(s, f(s))$$
- A (Markov) strategy π is a sequence of decision rules $\{f_n, n = 0, 1, ...\}$ such that f_n is used by the controller at stage n
- A stationary strategy is a *stage-independent* strategy; i.e. it uses the same decision rule f at each stage, and will be denoted as $\mathbf{f}$

Decision Rules, Strategies, Transitions & Rewards

- A decision rule $f : S \rightarrow A$ is a function that specifies the action that a controller chooses in a given state
- Each decision rule f defines
  - a transition probability matrix $P(f)$ such that the (s, s')-th entry is given by
$$P(f)[s, s'] = p(s'|s, f(s))$$
  - a reward vector $r(f)$ such that the (s)-th entry is given by
$$r(f)[s] = r(s, f(s))$$
- A (Markov) strategy π is a sequence of decision rules $\{f_n, n = 0, 1, ...\}$ such that f_n is used by the controller at stage n
- A stationary strategy is a *stage-independent* strategy; i.e. it uses the same decision rule f at each stage, and will be denoted as $\mathbf{f}$

Decision Rules, Strategies, Transitions & Rewards

- A decision rule $f : S \rightarrow A$ is a function that specifies the action that a controller chooses in a given state
- Each decision rule f defines
  - a transition probability matrix $P(f)$ such that the (s, s')-th entry is given by
$$P(f)[s, s'] = p(s'|s, f(s))$$
  - a reward vector $r(f)$ such that the (s)-th entry is given by
$$r(f)[s] = r(s, f(s))$$
- A (Markov) strategy π is a sequence of decision rules $\{f_n, n = 0, 1, ...\}$ such that f_n is used by the controller at stage n
- A stationary strategy is a *stage-independent* strategy; i.e. it uses the same decision rule f at each stage, and will be denoted as $\mathbf{f}$

Decision Rules, Strategies, Transitions & Rewards

- A decision rule $f : S \longrightarrow A$ is a function that specifies the action that a controller chooses in a given state
- Each decision rule f defines
  - a transition probability matrix $P(f)$ such that the (s, s')-th entry is given by
$$P(f)[s, s'] = p(s'|s, f(s))$$
  - a reward vector $r(f)$ such that the (s)-th entry is given by
$$r(f)[s] = r(s, f(s))$$
- A (Markov) strategy π is a sequence of decision rules $\{f_n, n = 0, 1, ...\}$ such that f_n is used by the controller at stage n
- A stationary strategy is a *stage-independent* strategy; i.e. it uses the same decision rule f at each stage, and will be denoted as $\mathbf{f}$

Decision Rules, Strategies, Transitions & Rewards

- A decision rule $f : S \to A$ is a function that specifies the action that a controller chooses in a given state
- Each decision rule f defines
  - a transition probability matrix $P(f)$ such that the (s, s')-th entry is given by
$$P(f)[s, s'] = p(s'|s, f(s))$$
  - a reward vector $r(f)$ such that the (s)-th entry is given by
$$r(f)[s] = r(s, f(s))$$
- A (Markov) strategy π is a sequence of decision rules $\{f_n, n = 0, 1, ...\}$ such that f_n is used by the controller at stage n
- A stationary strategy is a *stage-independent* strategy; i.e. it uses the same decision rule f at each stage, and will be denoted as $\mathbf{f}$

Decision Rules, Strategies, Transitions & Rewards

- A decision rule $f : S \to A$ is a function that specifies the action that a controller chooses in a given state
- Each decision rule f defines
  - a transition probability matrix $P(f)$ such that the (s, s')-th entry is given by
$$P(f)[s, s'] = p(s'|s, f(s))$$
  - a reward vector $r(f)$ such that the (s)-th entry is given by
$$r(f)[s] = r(s, f(s))$$
- A (Markov) strategy π is a sequence of decision rules $\{f_n, n = 0, 1, ...\}$ such that f_n is used by the controller at stage n
- A stationary strategy is a *stage-independent* strategy; i.e. it uses the same decision rule f at each stage, and will be denoted as $\mathbf{f}$

Decision Rules, Strategies, Transitions & Rewards

- Each strategy π defines a t-stage transition probability matrix as

$$P_t(\pi) := P(f_0)P(f_1)...P(f_{t-1}) \text{ for } t = 1, 2, ...$$

where

$$P_0(\pi) := I_N$$

- Associate with each strategy π the β-discounted value vector

$$\phi_\beta(\pi) = \sum_{t=0}^{\infty} \beta^t P_t(\pi) r(f_t)$$

Decision Rules, Strategies, Transitions & Rewards

- Each strategy π defines a t-stage transition probability matrix as

$$P_t(\pi) := P(f_0)P(f_1)...P(f_{t-1}) \text{ for } t = 1, 2, ...$$

where

$$P_0(\pi) := I_N$$

- Associate with each strategy π the β-discounted value vector

$$\phi_\beta(\pi) = \sum_{t=0}^{\infty} \beta^t P_t(\pi) r(f_t)$$

$\geq$ Ordering for Value Vectors and Strategies

$\geq$ Ordering Value Vectors

- Denote $u \geq w$ whenever $[u]_s \geq [w]_s$ for all $s \in S$
- Denote $u > w$ whenever $u \geq w$ and $u \neq w$

$\geq$ Ordering for Value Vectors and Strategies

$\geq$ Ordering Value Vectors

- Denote $u \geq w$ whenever $[u]_s \geq [w]_s$ for all $s \in S$
- Denote $u > w$ whenever $u \geq w$ and $u \neq w$

$\geq$ Ordering Strategies

- $\pi^1 \geq \pi^2$ iff $\phi_\beta(\pi^1) \geq \phi_\beta(\pi^2)$

Optimal Strategies

π^0 is an optimal strategy if

$$\pi^0 \geq \pi \text{ for all } \pi$$

or, in other words,

$$\pi^0 \text{ “maximizes” } \phi_\beta(\pi)$$

The Problem

Given an MDP, find the optimal strategy

The $L(\cdot)$ operator

- Let $\pi^+ := \{f_1, f_2, ...\}$ (where $\pi = \{f_0, f_1, f_2, ...\}$)
  - Observe that

$$\begin{aligned}\phi_\beta(\pi) &= r(f_0) + \beta P(f_0) \sum_{t=1}^{\infty} \beta^{t-1} P_{t-1}(\pi^+) r(f_t) \\ &= r(f_0) + \beta P(f_0) \phi_\beta(\pi^+)\end{aligned}$$

- For every decision rule g, we define the associated operator

$$L(g) : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$$

as

$$L(g)(u) := r(g) + \beta P(g) u \ \ ; u \in \mathbb{R}^{|S|}$$

  - In this notation,

$$\phi_\beta(f, \pi) = L(f)(\phi_\beta(\pi))$$

and, more generally,

$$\phi_\beta(f_0, f_1, ..., f_{t-1}, \pi) = L(f_0) L(f_1) ... L(f_{t-1})(\phi_\beta(\pi))$$

  - $L(g)$ is a *monotone* operator

$$u \geq w \Rightarrow L(g)(u) \geq L(g)(w)$$

The $L(\cdot)$ operator

- Let $\pi^+ := \{f_1, f_2, ...\}$ (where $\pi = \{f_0, f_1, f_2, ...\}$)
  - Observe that

$$\begin{aligned}\phi_\beta(\pi) &= r(f_0) + \beta P(f_0) \sum_{t=1}^{\infty} \beta^{t-1} P_{t-1}(\pi^+) r(f_t) \\ &= r(f_0) + \beta P(f_0) \phi_\beta(\pi^+)\end{aligned}$$

- For every decision rule g, we define the associated operator

$$L(g) : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$$

as

$$L(g)(u) := r(g) + \beta P(g) u \ \ ; u \in \mathbb{R}^{|S|}$$

  - In this notation,

$$\phi_\beta(f, \pi) = L(f)(\phi_\beta(\pi))$$

and, more generally,

$$\phi_\beta(f_0, f_1, ..., f_{t-1}, \pi) = L(f_0) L(f_1) ... L(f_{t-1})(\phi_\beta(\pi))$$

  - $L(g)$ is a *monotone* operator

$$u \geq w \Rightarrow L(g)(u) \geq L(g)(w)$$

The $L(\cdot)$ operator

- Let $\pi^+ := \{f_1, f_2, ...\}$ (where $\pi = \{f_0, f_1, f_2, ...\}$)
  - Observe that

$$\begin{aligned}\phi_\beta(\pi) &= r(f_0) + \beta P(f_0) \sum_{t=1}^{\infty} \beta^{t-1} P_{t-1}(\pi^+) r(f_t) \\ &= r(f_0) + \beta P(f_0) \phi_\beta(\pi^+)\end{aligned}$$

- For every decision rule g, we define the associated operator

$$L(g) : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$$

as

$$L(g)(u) := r(g) + \beta P(g) u \ ; u \in \mathbb{R}^{|S|}$$

  - In this notation,

$$\phi_\beta(f, \pi) = L(f)(\phi_\beta(\pi))$$

and, more generally,

$$\phi_\beta(f_0, f_1, ..., f_{t-1}, \pi) = L(f_0) L(f_1) ... L(f_{t-1})(\phi_\beta(\pi))$$

  - $L(g)$ is a *monotone* operator

$$u \geq w \Rightarrow L(g)(u) \geq L(g)(w)$$

The $L(\cdot)$ operator

- Let $\pi^+ := \{f_1, f_2, ...\}$ (where $\pi = \{f_0, f_1, f_2, ...\}$)
  - Observe that

$$\begin{aligned}\phi_\beta(\pi) &= r(f_0) + \beta P(f_0) \sum_{t=1}^{\infty} \beta^{t-1} P_{t-1}(\pi^+) r(f_t) \\ &= r(f_0) + \beta P(f_0) \phi_\beta(\pi^+)\end{aligned}$$

- For every decision rule g, we define the associated operator

$$L(g) : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$$

as

$$L(g)(u) := r(g) + \beta P(g) u \ ; u \in \mathbb{R}^{|S|}$$

  - In this notation,

$$\phi_\beta(f, \pi) = L(f)(\phi_\beta(\pi))$$

and, more generally,

$$\phi_\beta(f_0, f_1, ..., f_{t-1}, \pi) = L(f_0) L(f_1) ... L(f_{t-1})(\phi_\beta(\pi))$$

  - $L(g)$ is a *monotone* operator

$$u \geq w \Rightarrow L(g)(u) \geq L(g)(w)$$

The $L(\cdot)$ operator

- Let $\pi^+ := \{f_1, f_2, ...\}$ (where $\pi = \{f_0, f_1, f_2, ...\}$)
  - Observe that

$$\begin{aligned}\phi_\beta(\pi) &= r(f_0) + \beta P(f_0) \sum_{t=1}^{\infty} \beta^{t-1} P_{t-1}(\pi^+) r(f_t) \\ &= r(f_0) + \beta P(f_0) \phi_\beta(\pi^+)\end{aligned}$$

- For every decision rule g, we define the associated operator

$$L(g) : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$$

as

$$L(g)(u) := r(g) + \beta P(g) u \ \ ; u \in \mathbb{R}^{|S|}$$

  - In this notation,

$$\phi_\beta(f, \pi) = L(f)(\phi_\beta(\pi))$$

and, more generally,

$$\phi_\beta(f_0, f_1, ..., f_{t-1}, \pi) = L(f_0) L(f_1) ... L(f_{t-1})(\phi_\beta(\pi))$$

  - $L(g)$ is a *monotone* operator

$$u \geq w \Rightarrow L(g)(u) \geq L(g)(w)$$

Theorem 1

Let $\pi^0 = \{f_n^0, n = 0, 1, ...\}$. If $\pi^0 \geq (f, \pi^0)$ for all f, then π^0 is optimal.

Proof.

Let $\pi = \{f_n, n = 0, 1, ...\}$ be an arbitrary (Markov) strategy.

$$\pi^0 \geq (f, \pi^0) \Rightarrow \phi_\beta(\pi^0) \geq \phi_\beta(f, \pi^0) = L(f)(\phi_\beta(\pi^0))$$

We use the monotonicity of the L(f) operator to obtain for e.g.

$$\phi_\beta(\pi^0) \geq L(f_{T-1})(\phi_\beta(\pi^0)) \geq L(f_{T-1})L(f_T)(\phi_\beta(\pi^0))$$

Iterating this argument using the first $(T+1)$ decision rules of π to obtain

$$\begin{aligned}\phi_\beta(\pi^0) &\geq L(f_0)L(f_1)...L(f_T)(\phi_\beta(\pi^0)) \\ &= \phi_\beta(f_0, f_1, ..., f_T, \pi^0) \\ &= \sum_{t=0}^{T} \beta^t P_t(\pi) r(f_t) + \sum_{t=T+1}^{\infty} \beta^t P_t(f_0, ...f_T, \pi^0) r(f_{t-T-1}^0)\end{aligned}$$

for every nonnegative integer T

Allowing $T \to \infty$, the last term above vanishes, leading to

$$\phi_\beta(\pi^0) \geq \sum_{t=0}^{\infty} \beta^t P_t(\pi) r(f_t) = \phi_\beta(\pi)$$

Since π was arbitrary, the proof is complete. □

Theorem 1

Let $\pi^0 = \{f_n^0, n = 0, 1, ...\}$. If $\pi^0 \geq (f, \pi^0)$ for all f, then π^0 is optimal.

Proof.

Let $\pi = \{f_n, n = 0, 1, ...\}$ be an arbitrary (Markov) strategy.

$\pi^0 \geq (f, \pi^0) \Rightarrow \phi_\beta(\pi^0) \geq \phi_\beta(f, \pi^0) = L(f)(\phi_\beta(\pi^0))$

We use the monotonicity of the L(f) operator to obtain for e.g.
$\phi_\beta(\pi^0) \geq L(f_{T-1})(\phi_\beta(\pi^0)) \geq L(f_{T-1})L(f_T)(\phi_\beta(\pi^0))$

Iterating this argument using the first $(T+1)$ decision rules of π to obtain

$$\begin{aligned}\phi_\beta(\pi^0) &\geq L(f_0)L(f_1)...L(f_T)(\phi_\beta(\pi^0)) \\ &= \phi_\beta(f_0, f_1, ..., f_T, \pi^0) \\ &= \sum_{t=0}^{T} \beta^t P_t(\pi) r(f_t) + \sum_{t=T+1}^{\infty} \beta^t P_t(f_0, ...f_T, \pi^0) r(f_{t-T-1}^0)\end{aligned}$$

for every nonnegative integer T

Allowing $T \to \infty$, the last term above vanishes, leading to
$\phi_\beta(\pi^0) \geq \sum_{t=0}^{\infty} \beta^t P_t(\pi) r(f_t) = \phi_\beta(\pi)$

Since π was arbitrary, the proof is complete. □

Theorem 1

Let $\pi^0 = \{f_n^0, n = 0, 1, ...\}$. If $\pi^0 \geq (f, \pi^0)$ for all f, then π^0 is optimal.

Proof.

Let $\pi = \{f_n, n = 0, 1, ...\}$ be an arbitrary (Markov) strategy.

$\pi^0 \geq (f, \pi^0) \Rightarrow \phi_\beta(\pi^0) \geq \phi_\beta(f, \pi^0) = L(f)(\phi_\beta(\pi^0))$

We use the monotonicity of the L(f) operator to obtain for e.g.
$\phi_\beta(\pi^0) \geq L(f_{T-1})(\phi_\beta(\pi^0)) \geq L(f_{T-1})L(f_T)(\phi_\beta(\pi^0))$

Iterating this argument using the first $(T+1)$ decision rules of π to obtain

$$\begin{aligned}\phi_\beta(\pi^0) &\geq L(f_0)L(f_1)...L(f_T)(\phi_\beta(\pi^0)) \\ &= \phi_\beta(f_0, f_1, ..., f_T, \pi^0) \\ &= \sum_{t=0}^{T} \beta^t P_t(\pi) r(f_t) + \sum_{t=T+1}^{\infty} \beta^t P_t(f_0, ...f_T, \pi^0) r(f_{t-T-1}^0)\end{aligned}$$

for every nonnegative integer T

Allowing $T \to \infty$, the last term above vanishes, leading to
$\phi_\beta(\pi^0) \geq \sum_{t=0}^{\infty} \beta^t P_t(\pi) r(f_t) = \phi_\beta(\pi)$

Since π was arbitrary, the proof is complete. □

Theorem 1

Let $\pi^0 = \{f_n^0, n = 0, 1, ...\}$. If $\pi^0 \geq (f, \pi^0)$ for all f, then π^0 is optimal.

Proof.

Let $\pi = \{f_n, n = 0, 1, ...\}$ be an arbitrary (Markov) strategy.

$\pi^0 \geq (f, \pi^0) \Rightarrow \phi_\beta(\pi^0) \geq \phi_\beta(f, \pi^0) = L(f)(\phi_\beta(\pi^0))$

We use the monotonicity of the L(f) operator to obtain for e.g.
$\phi_\beta(\pi^0) \geq L(f_{T-1})(\phi_\beta(\pi^0)) \geq L(f_{T-1})L(f_T)(\phi_\beta(\pi^0))$

Iterating this argument using the first $(T+1)$ decision rules of π to obtain

$$\begin{aligned}\phi_\beta(\pi^0) &\geq L(f_0)L(f_1)...L(f_T)(\phi_\beta(\pi^0)) \\ &= \phi_\beta(f_0, f_1, ..., f_T, \pi^0) \\ &= \sum_{t=0}^{T} \beta^t P_t(\pi) r(f_t) + \sum_{t=T+1}^{\infty} \beta^t P_t(f_0, ...f_T, \pi^0) r(f_{t-T-1}^0)\end{aligned}$$

for every nonnegative integer T

Allowing $T \to \infty$, the last term above vanishes, leading to
$\phi_\beta(\pi^0) \geq \sum_{t=0}^{\infty} \beta^t P_t(\pi) r(f_t) = \phi_\beta(\pi)$

Since π was arbitrary, the proof is complete. □

Theorem 1

Let $\pi^0 = \{f_n^0, n = 0, 1, ...\}$. If $\pi^0 \geq (f, \pi^0)$ for all f, then π^0 is optimal.

Proof.

Let $\pi = \{f_n, n = 0, 1, ...\}$ be an arbitrary (Markov) strategy.

$\pi^0 \geq (f, \pi^0) \Rightarrow \phi_\beta(\pi^0) \geq \phi_\beta(f, \pi^0) = L(f)(\phi_\beta(\pi^0))$

We use the monotonicity of the L(f) operator to obtain for e.g.
$\phi_\beta(\pi^0) \geq L(f_{T-1})(\phi_\beta(\pi^0)) \geq L(f_{T-1})L(f_T)(\phi_\beta(\pi^0))$

Iterating this argument using the first $(T+1)$ decision rules of π to obtain

$$\begin{aligned} \phi_\beta(\pi^0) &\geq L(f_0)L(f_1)...L(f_T)(\phi_\beta(\pi^0)) \\ &= \phi_\beta(f_0, f_1, ..., f_T, \pi^0) \\ &= \sum_{t=0}^{T} \beta^t P_t(\pi) r(f_t) + \sum_{t=T+1}^{\infty} \beta^t P_t(f_0, ...f_T, \pi^0) r(f_{t-T-1}^0) \end{aligned}$$

for every nonnegative integer T

Allowing $T \to \infty$, the last term above vanishes, leading to
$\phi_\beta(\pi^0) \geq \sum_{t=0}^{\infty} \beta^t P_t(\pi) r(f_t) = \phi_\beta(\pi)$

Since π was arbitrary, the proof is complete. □

Theorem 1

Let $\pi^0 = \{f_n^0, n = 0, 1, ...\}$. If $\pi^0 \geq (f, \pi^0)$ for all f, then π^0 is optimal.

Proof.

Let $\pi = \{f_n, n = 0, 1, ...\}$ be an arbitrary (Markov) strategy.

$\pi^0 \geq (f, \pi^0) \Rightarrow \phi_\beta(\pi^0) \geq \phi_\beta(f, \pi^0) = L(f)(\phi_\beta(\pi^0))$

We use the monotonicity of the L(f) operator to obtain for e.g.
$\phi_\beta(\pi^0) \geq L(f_{T-1})(\phi_\beta(\pi^0)) \geq L(f_{T-1})L(f_T)(\phi_\beta(\pi^0))$

Iterating this argument using the first $(T+1)$ decision rules of π to obtain

$$\begin{aligned} \phi_\beta(\pi^0) &\geq L(f_0)L(f_1)...L(f_T)(\phi_\beta(\pi^0)) \\ &= \phi_\beta(f_0, f_1, ..., f_T, \pi^0) \\ &= \sum_{t=0}^{T} \beta^t P_t(\pi) r(f_t) + \sum_{t=T+1}^{\infty} \beta^t P_t(f_0, ...f_T, \pi^0) r(f_{t-T-1}^0) \end{aligned}$$

for every nonnegative integer T

Allowing $T \to \infty$, the last term above vanishes, leading to
$\phi_\beta(\pi^0) \geq \sum_{t=0}^{\infty} \beta^t P_t(\pi) r(f_t) = \phi_\beta(\pi)$

Since π was arbitrary, the proof is complete. □

Theorem 1

Let $\pi^0 = \{f_n^0, n = 0, 1, ...\}$. If $\pi^0 \geq (f, \pi^0)$ for all f, then π^0 is optimal.

Proof.

Let $\pi = \{f_n, n = 0, 1, ...\}$ be an arbitrary (Markov) strategy.

$\pi^0 \geq (f, \pi^0) \Rightarrow \phi_\beta(\pi^0) \geq \phi_\beta(f, \pi^0) = L(f)(\phi_\beta(\pi^0))$

We use the monotonicity of the L(f) operator to obtain for e.g.
$\phi_\beta(\pi^0) \geq L(f_{T-1})(\phi_\beta(\pi^0)) \geq L(f_{T-1})L(f_T)(\phi_\beta(\pi^0))$

Iterating this argument using the first $(T+1)$ decision rules of π to obtain

$$\begin{aligned} \phi_\beta(\pi^0) &\geq L(f_0)L(f_1)...L(f_T)(\phi_\beta(\pi^0)) \\ &= \phi_\beta(f_0, f_1, ..., f_T, \pi^0) \\ &= \sum_{t=0}^{T} \beta^t P_t(\pi) r(f_t) + \sum_{t=T+1}^{\infty} \beta^t P_t(f_0, ...f_T, \pi^0) r(f_{t-T-1}^0) \end{aligned}$$

for every nonnegative integer T

Allowing $T \to \infty$, the last term above vanishes, leading to
$\phi_\beta(\pi^0) \geq \sum_{t=0}^{\infty} \beta^t P_t(\pi) r(f_t) = \phi_\beta(\pi)$

Since π was arbitrary, the proof is complete. □

Theorem 2

Let $\pi = \{f_n, n = 0, 1, ...\}$ be an arbitrary (Markov) strategy and f a decision rule such that $(f, \pi) > \pi$. Then $\mathbf{f} > \pi$.

Proof.

$(f, \pi) > \pi \Rightarrow \phi_\beta(f, \pi) = L(f)(\phi_\beta(\pi)) > \phi_\beta(\pi)$

As before, we use the monotonicity of $L(f)$ and repeated application of the above inequality to obtain $\phi_\beta(f, f, f, ..., f, \pi) = L^{(T)}(f)(\phi_\beta(\pi)) \geq L(f)(\phi_\beta(\pi)) = \phi_\beta(f, \pi) > \phi_\beta(\pi)$

Allowing $T \longrightarrow \infty$ yields $\phi_\beta(\mathbf{f}) = \phi_\beta(f, f, f, ...) \geq \phi_\beta(f, \pi) > \phi_\beta(\pi)$

$\Rightarrow \mathbf{f} > \pi$ □

Theorem 2

Let $\pi = \{f_n, n = 0, 1, ...\}$ be an arbitrary (Markov) strategy and f a decision rule such that $(f, \pi) > \pi$. Then $\mathbf{f} > \pi$.

Proof.

$(f, \pi) > \pi \Rightarrow \phi_\beta(f, \pi) = L(f)(\phi_\beta(\pi)) > \phi_\beta(\pi)$

As before, we use the monotonicity of $L(f)$ and repeated application of the above inequality to obtain $\phi_\beta(f, f, f, ..., f, \pi) = L^{(T)}(f)(\phi_\beta(\pi)) \geq L(f)(\phi_\beta(\pi)) = \phi_\beta(f, \pi) > \phi_\beta(\pi)$

Allowing $T \to \infty$ yields $\phi_\beta(\mathbf{f}) = \phi_\beta(f, f, f, ...) \geq \phi_\beta(f, \pi) > \phi_\beta(\pi)$

$\Rightarrow \mathbf{f} > \pi$ □

Theorem 2

Let $\pi = \{f_n, n = 0, 1, ...\}$ be an arbitrary (Markov) strategy and f a decision rule such that $(f, \pi) > \pi$. Then $\mathbf{f} > \pi$.

Proof.

$(f, \pi) > \pi \Rightarrow \phi_\beta(f, \pi) = L(f)(\phi_\beta(\pi)) > \phi_\beta(\pi)$

As before, we use the monotonicity of $L(f)$ and repeated application of the above inequality to obtain $\phi_\beta(f, f, f, ..., f, \pi) = L^{(T)}(f)(\phi_\beta(\pi)) \geq L(f)(\phi_\beta(\pi)) = \phi_\beta(f, \pi) > \phi_\beta(\pi)$

Allowing $T \to \infty$ yields $\phi_\beta(\mathbf{f}) = \phi_\beta(f, f, f, ...) \geq \phi_\beta(f, \pi) > \phi_\beta(\pi)$

$\Rightarrow \mathbf{f} > \pi$ □

Theorem 2

Let $\pi = \{f_n, n = 0, 1, ...\}$ be an arbitrary (Markov) strategy and f a decision rule such that $(f, \pi) > \pi$. Then $\mathbf{f} > \pi$.

Proof.

$(f, \pi) > \pi \Rightarrow \phi_\beta(f, \pi) = L(f)(\phi_\beta(\pi)) > \phi_\beta(\pi)$

As before, we use the monotonicity of $L(f)$ and repeated application of the above inequality to obtain $\phi_\beta(f, f, f, ..., f, \pi) = L^{(T)}(f)(\phi_\beta(\pi)) \geq L(f)(\phi_\beta(\pi)) = \phi_\beta(f, \pi) > \phi_\beta(\pi)$

Allowing $T \to \infty$ yields $\phi_\beta(\mathbf{f}) = \phi_\beta(f, f, f, ...) \geq \phi_\beta(f, \pi) > \phi_\beta(\pi)$

$\Rightarrow \mathbf{f} > \pi$ □

Theorem 2

Let $\pi = \{f_n, n = 0, 1, ...\}$ be an arbitrary (Markov) strategy and f a decision rule such that $(f, \pi) > \pi$. Then $\mathbf{f} > \pi$.

Proof.

$(f, \pi) > \pi \Rightarrow \phi_\beta(f, \pi) = L(f)(\phi_\beta(\pi)) > \phi_\beta(\pi)$

As before, we use the monotonicity of $L(f)$ and repeated application of the above inequality to obtain $\phi_\beta(f, f, f, ..., f, \pi) = L^{(T)}(f)(\phi_\beta(\pi)) \geq L(f)(\phi_\beta(\pi)) = \phi_\beta(f, \pi) > \phi_\beta(\pi)$

Allowing $T \to \infty$ yields $\phi_\beta(\mathbf{f}) = \phi_\beta(f, f, f, ...) \geq \phi_\beta(f, \pi) > \phi_\beta(\pi)$

$\Rightarrow \mathbf{f} > \pi$ □

Policy Improvement Algorithm

Step 0. (Initialization) Set $k := 0$. Select any pure strategy $\mathbf{f}$.
Set $\mathbf{f}^0 := \mathbf{f}$ and $\phi^0 := \phi_\beta(\mathbf{f}^0) = [I - \beta P(\mathbf{f}^0)]^{-1} r(\mathbf{f}^0)$

Step 1. (Check Optimality) Let a_s^k be the action selected by $\mathbf{f}^k$ in state s.
If the following "optimality equation"

$$r(s, a_s^k) + \beta \sum_{s'=1}^{N} p(s'|s, a_s^k)\phi^k(s') = \max_{a \in A} \left\{ r(s,a) + \beta \sum_{s'=1}^{N} p(s'|s, a_s^k)\phi^k(s') \right\} \qquad (\star)$$

holds for each $s \in S$, STOP.
The strategy f^k is optimal and ϕ^k is the discounted optimal value vector.

Policy Improvement Algorithm

Step 0. (Initialization) Set $k := 0$. Select any pure strategy $\mathbf{f}$.
Set $\mathbf{f}^0 := \mathbf{f}$ and $\phi^0 := \phi_\beta(\mathbf{f}^0) = [I - \beta P(\mathbf{f}^0)]^{-1} r(\mathbf{f}^0)$

Step 1. (Check Optimality) Let a_s^k be the action selected by $\mathbf{f}^k$ in state s.
If the following "optimality equation"

$$r(s, a_s^k) + \beta \sum_{s'=1}^{N} p(s'|s, a_s^k)\phi^k(s') = \max_{a \in A} \left\{ r(s,a) + \beta \sum_{s'=1}^{N} p(s'|s, a_s^k)\phi^k(s') \right\} \qquad (\star)$$

holds for each $s \in S$, STOP.
The strategy f^k is optimal and ϕ^k is the discounted optimal value vector.

Policy Improvement Algorithm

Step 2. (Policy improvement) Let $\overline{S}$ be the nonempty subset of states for which equality is violated in $(\star)$, i.e., the left side is strictly smaller than the right side. Define

$$\overline{a}_s^k = \arg\max_{a \in A} \left\{ r(s,a) + \beta \sum_{s'=1}^{N} p(s'|s, a_s^k) \phi^k(s') \right\}$$

for each s $\in \overline{S}$, and a new strategy $\mathbf{g}$ by

$$g(s,a) = \begin{cases} f(s,a) & \text{if s} \notin \overline{S}, \\ 1 & \text{if s} \in \overline{S} \text{ and a} = \overline{a}_s^k \\ 0 & \text{otherwise} \end{cases}$$

Set $\mathbf{f}^{k+1} := \mathbf{g}$ and $\phi^{k+1} := \phi_\beta(\mathbf{g})$

Step 3. (Iteration) Set $k := k+1$, $\mathbf{f}^k := \mathbf{f}^{k+1}$, $\phi^k := \phi^{k+1}$. Return to Step 1.

Policy Improvement Algorithm

Step 2. (Policy improvement) Let $\overline{S}$ be the nonempty subset of states for which equality is violated in $(\star)$, i.e., the left side is strictly smaller than the right side. Define

$$\overline{a}_s^k = \arg\max_{a \in A} \left\{ r(s,a) + \beta \sum_{s'=1}^{N} p(s'|s, a_s^k) \phi^k(s') \right\}$$

for each s $\in \overline{S}$, and a new strategy $\mathbf{g}$ by

$$g(s,a) = \begin{cases} f(s,a) & \text{if s} \notin \overline{S}, \\ 1 & \text{if s} \in \overline{S} \text{ and a} = \overline{a}_s^k \\ 0 & \text{otherwise} \end{cases}$$

Set $\mathbf{f}^{k+1} := \mathbf{g}$ and $\phi^{k+1} := \phi_\beta(\mathbf{g})$

Step 3. (Iteration) Set $k := k+1$, $\mathbf{f}^k := \mathbf{f}^{k+1}$, $\phi^k := \phi^{k+1}$. Return to Step 1.

Correctness & Termination

Improvement Lemma

$\mathbf{f}^{k+1} > \mathbf{f}^k$

Proof.

Let f^{k+1} be the decision rule corresponding to $\mathbf{f}^{k+1}$.

By its definition, we have that

$$\phi_\beta(f^{k+1}, \mathbf{f}^k) = r(f^{k+1}) + \beta P(f^{k+1}) \phi_\beta(\mathbf{f}^k) > \phi_\beta(\mathbf{f}^k)$$

i.e. that

$$(f^{k+1}, \mathbf{f}^k) > \mathbf{f}^k$$

Then, from Theorem 2, $\mathbf{f}^{k+1} > \mathbf{f}^k$. □

Correctness & Termination

Improvement Lemma

$\mathbf{f}^{k+1} > \mathbf{f}^{k}$

Proof.

Let f^{k+1} be the decision rule corresponding to $\mathbf{f}^{k+1}$.

By its definition, we have that

$$\phi_\beta(f^{k+1},\mathbf{f}^{k}) = r(f^{k+1}) + \beta P(f^{k+1})\phi_\beta(\mathbf{f}^{k}) > \phi_\beta(\mathbf{f}^{k})$$

i.e. that

$$(f^{k+1},\mathbf{f}^{k}) > \mathbf{f}^{k}$$

Then, from Theorem 2, $\mathbf{f}^{k+1} > \mathbf{f}^{k}$. □

Correctness & Termination

Improvement Lemma

$\mathbf{f}^{k+1} > \mathbf{f}^{k}$

Proof.

Let f^{k+1} be the decision rule corresponding to $\mathbf{f}^{k+1}$.

By its definition, we have that

$$\phi_\beta(f^{k+1},\mathbf{f}^{k}) = r(f^{k+1}) + \beta P(f^{k+1})\phi_\beta(\mathbf{f}^{k}) > \phi_\beta(\mathbf{f}^{k})$$

i.e. that

$$(f^{k+1},\mathbf{f}^{k}) > \mathbf{f}^{k}$$

Then, from Theorem 2, $\mathbf{f}^{k+1} > \mathbf{f}^{k}$. □

Correctness & Termination

Improvement Lemma

$\mathbf{f}^{k+1} > \mathbf{f}^{k}$

Proof.

Let f^{k+1} be the decision rule corresponding to $\mathbf{f}^{k+1}$.
By its definition, we have that

$$\phi_\beta(\mathrm{f}^{k+1},\mathbf{f}^{k}) = \mathrm{r}(\mathrm{f}^{k+1}) + \beta \mathrm{P}(\mathrm{f}^{k+1})\phi_\beta(\mathbf{f}^{k}) > \phi_\beta(\mathbf{f}^{k})$$

i.e. that

$$(\mathrm{f}^{k+1},\mathbf{f}^{k}) > \mathbf{f}^{k}$$

Then, from Theorem 2, $\mathbf{f}^{k+1} > \mathbf{f}^{k}$. □

Correctness & Termination

Improvement Lemma

$\mathbf{f}^{k+1} > \mathbf{f}^{k}$

Proof.

Let f^{k+1} be the decision rule corresponding to $\mathbf{f}^{k+1}$.
By its definition, we have that

$$\phi_\beta(\mathrm{f}^{k+1},\mathbf{f}^{k}) = \mathrm{r}(\mathrm{f}^{k+1}) + \beta \mathrm{P}(\mathrm{f}^{k+1})\phi_\beta(\mathbf{f}^{k}) > \phi_\beta(\mathbf{f}^{k})$$

i.e. that

$$(\mathrm{f}^{k+1},\mathbf{f}^{k}) > \mathbf{f}^{k}$$

Then, from Theorem 2, $\mathbf{f}^{k+1} > \mathbf{f}^{k}$. □

Correctness & Termination

Termination Lemma

The Policy Improvement Algorithm terminates in finite steps.

Proof.

At stage k, if the algorithm has not already found the optimal strategy, it computes $\mathbf{f}^{k+1}$ such that $\mathbf{f}^{k+1} > \mathbf{f}^k$ (Improvement Lemma). Therefore, the algorithm **does not cycle**. There are only finite strategies. Therefore, the algorithm terminates when it finds the (an) optimal strategy. □

Correctness & Termination

Termination Lemma

The Policy Improvement Algorithm terminates in finite steps.

Proof.

At stage k, if the algorithm has not already found the optimal strategy, it computes $\mathbf{f}^{k+1}$ such that $\mathbf{f}^{k+1} > \mathbf{f}^{k}$ (Improvement Lemma). Therefore, the algorithm **does not cycle**. There are only finite strategies. Therefore, the algorithm terminates when it finds the (an) optimal strategy. □

Correctness & Termination

Termination Lemma

The Policy Improvement Algorithm terminates in finite steps.

Proof.

At stage k, if the algorithm has not already found the optimal strategy, it computes $\mathbf{f}^{k+1}$ such that $\mathbf{f}^{k+1} > \mathbf{f}^{k}$ (Improvement Lemma). Therefore, the algorithm **does not cycle**. There are only finite strategies. Therefore, the algorithm terminates when it finds the (an) optimal strategy. □

Correctness & Termination

Termination Lemma

The Policy Improvement Algorithm terminates in finite steps.

Proof.

At stage k, if the algorithm has not already found the optimal strategy, it computes $\mathbf{f}^{k+1}$ such that $\mathbf{f}^{k+1} > \mathbf{f}^{k}$ (Improvement Lemma). Therefore, the algorithm **does not cycle**. There are only finite strategies. Therefore, the algorithm terminates when it finds the (an) optimal strategy. □

An Example

State: s1

a	r(s1,a)	p([s1,s2]|s1,a)
a1	1	[0.6,0.4]
a2	0	[1,0]

State: s2

a	r(s2,a)	p([s1,s2]|s2,a)
a1	-1	[0.6,0.4]
a2	0	[0,1]

Policy denoted as (action in state 1, action in state 2)
Corresponding value vector denoted as value of current policy starting when starting in (state 1,... state 2)

An Example

State: s1

a	r(s1,a)	p([s1,s2]|s1,a)
a1	1	[0.6,0.4]
a2	0	[1,0]

State: s2

a	r(s2,a)	p([s1,s2]|s2,a)
a1	-1	[0.6,0.4]
a2	0	[0,1]

Let $\beta = 0.9$

Policy $\mathbf{f}^0$: (a2,a2)

- Evaluate value vector $\phi_\beta^0 := (0, 0)$
- Policy improvement:
  - For state s1,
    consider a1: $1 + 0.9(0.6 * 0 + 0.4 * 0) = 1$
    consider a2: $0 + 0.9(0.6 * 0 + 0.4 * 0) = 0$
  - For state s2,
    consider a1: $-1 + 0.9(0.6 * 0 + 0.4 * 0) = -1$
    consider a2: $0 + 0.9(0 * 0 + 1 * 0) = 0$

An Example

State: s1

a	r(s1,a)	p([s1,s2]|s1,a)
a1	1	[0.6,0.4]
a2	0	[1,0]

State: s2

a	r(s2,a)	p([s1,s2]|s2,a)
a1	-1	[0.6,0.4]
a2	0	[0,1]

Let $\beta = 0.9$

Policy $\mathbf{f}^0$: (a2,a2)

- Evaluate value vector $\phi_\beta^0 := (0, 0)$
- Policy improvement:
  - For state s1,
    consider a1: $1 + 0.9(0.6 * 0 + 0.4 * 0) = 1$
    consider a2: $0 + 0.9(0.6 * 0 + 0.4 * 0) = 0$
  - For state s2,
    consider a1: $-1 + 0.9(0.6 * 0 + 0.4 * 0) = -1$
    consider a2: $0 + 0.9(0 * 0 + 1 * 0) = 0$

An Example

State: s1

a	r(s1,a)	p([s1,s2]|s1,a)
a1	1	[0.6,0.4]
a2	0	[1,0]

State: s2

a	r(s2,a)	p([s1,s2]|s2,a)
a1	-1	[0.6,0.4]
a2	0	[0,1]

Let $\beta = 0.9$

Policy $\mathbf{f}^0$: (a2,a2)

- Evaluate value vector $\phi_\beta^0 := (0, 0)$
- Policy improvement:
  - For state s1,
    consider a1: $1 + 0.9(0.6 * 0 + 0.4 * 0) = 1$
    consider a2: $0 + 0.9(0.6 * 0 + 0.4 * 0) = 0$
  - For state s2,
    consider a1: $-1 + 0.9(0.6 * 0 + 0.4 * 0) = -1$
    consider a2: $0 + 0.9(0 * 0 + 1 * 0) = 0$

An Example

State: s1

a	r(s1,a)	p([s1,s2]|s1,a)
a1	1	[0.6,0.4]
a2	0	[1,0]

State: s2

a	r(s2,a)	p([s1,s2]|s2,a)
a1	-1	[0.6,0.4]
a2	0	[0,1]

Let $\beta = 0.9$

Policy $\mathbf{f}^0$: (a2,a2)

- Evaluate value vector $\phi_\beta^0 := (0, 0)$
- Policy improvement:
  - For state s1,
    consider a1: $1 + 0.9(0.6 * 0 + 0.4 * 0) = \mathbf{1}$
    consider a2: $0 + 0.9(0.6 * 0 + 0.4 * 0) = 0$
  - For state s2,
    consider a1: $-1 + 0.9(0.6 * 0 + 0.4 * 0) = -1$
    consider a2: $0 + 0.9(0 * 0 + 1 * 0) = 0$

An Example

State: s1

a	r(s1,a)	p([s1,s2]|s1,a)
a1	1	[0.6,0.4]
a2	0	[1,0]

State: s2

a	r(s2,a)	p([s1,s2]|s2,a)
a1	-1	[0.6,0.4]
a2	0	[0,1]

Let $\beta = 0.9$

Policy $\mathbf{f}^1$: (a1,a2)

- Evaluate value vector $\phi_\beta^1 := (2.1739, 0)$
- Policy improvement:
  - For state s1,
    consider a1: $1 + 0.9(0.6 * 2.1739 + 0.4 * 0) = 2.1739$
    consider a2: $0 + 0.9(1 * 2.1739 + 0 * 0) = 1.9565$
  - For state s2,
    consider a1: $-1 + 0.9(0.6 * 2.1739 + 0.4 * 0) = 0.1739$
    consider a2: $0 + 0.9(0 * 2.1739 + 1 * 0) = 0$

An Example

State: s1

a	r(s1,a)	p([s1,s2]|s1,a)
a1	1	[0.6,0.4]
a2	0	[1,0]

State: s2

a	r(s2,a)	p([s1,s2]|s2,a)
a1	-1	[0.6,0.4]
a2	0	[0,1]

Let $\beta = 0.9$

Policy $\mathbf{f}^1$: (a1,a2)

- Evaluate value vector $\phi_\beta^1 := (2.1739, 0)$
- Policy improvement:
  - For state s1,
    consider a1: $1 + 0.9(0.6 * 2.1739 + 0.4 * 0) = 2.1739$
    consider a2: $0 + 0.9(1 * 2.1739 + 0 * 0) = 1.9565$
  - For state s2,
    consider a1: $-1 + 0.9(0.6 * 2.1739 + 0.4 * 0) = 0.1739$
    consider a2: $0 + 0.9(0 * 2.1739 + 1 * 0) = 0$

An Example

State: s1

a	r(s1,a)	p([s1,s2]|s1,a)
a1	1	[0.6,0.4]
a2	0	[1,0]

State: s2

a	r(s2,a)	p([s1,s2]|s2,a)
a1	-1	[0.6,0.4]
a2	0	[0,1]

Let $\beta = 0.9$

Policy $\mathbf{f}^1$: (a1,a2)

- Evaluate value vector $\phi_\beta^1 := (2.1739, 0)$
- Policy improvement:
  - For state s1,
    consider a1: $1 + 0.9(0.6 * 2.1739 + 0.4 * 0) = 2.1739$
    consider a2: $0 + 0.9(1 * 2.1739 + 0 * 0) = 1.9565$
  - For state s2,
    consider a1: $-1 + 0.9(0.6 * 2.1739 + 0.4 * 0) = 0.1739$
    consider a2: $0 + 0.9(0 * 2.1739 + 1 * 0) = 0$

An Example

State: s1

a	r(s1,a)	p([s1,s2]|s1,a)
a1	1	[0.6,0.4]
a2	0	[1,0]

State: s2

a	r(s2,a)	p([s1,s2]|s2,a)
a1	-1	[0.6,0.4]
a2	0	[0,1]

Let $\beta = 0.9$

Policy $\mathbf{f}^1$: (a1,a2)

- Evaluate value vector $\phi_\beta^1 := (2.1739, 0)$
- Policy improvement:
  - For state s1,
    consider a1: $1 + 0.9(0.6 * 2.1739 + 0.4 * 0) = 2.1739$
    consider a2: $0 + 0.9(1 * 2.1739 + 0 * 0) = 1.9565$
  - For state s2,
    consider a1: $-1 + 0.9(0.6 * 2.1739 + 0.4 * 0) = \mathbf{0.1739}$
    consider a2: $0 + 0.9(0 * 2.1739 + 1 * 0) = 0$

An Example

State: s1

a	r(s1,a)	p([s1,s2]|s1,a)
a1	1	[0.6,0.4]
a2	0	[1,0]

State: s2

a	r(s2,a)	p([s1,s2]|s2,a)
a1	-1	[0.6,0.4]
a2	0	[0,1]

Let $\beta = 0.9$

Policy $\mathbf{f}^2$: (a1,a1)

- Evaluate value vector $\phi^2_\beta := (2.302, 0.1637)$
- Policy improvement:
  - For state s1,
    consider a1: $1 + 0.9(0.6 * 2.302 + 0.4 * 0.1637) = 2.302$
    consider a2: $0 + 0.9(1 * 2.302 + 0 * 0.1637) = 2.0718$
  - For state s2,
    consider a1: $-1 + 0.9(0.6 * 2.302 + 0.4 * 0.1637) = 0.302$
    consider a2: $0 + 0.9(0 * 2.302 + 1 * 0.1637) = 0.14733$

An Example

State: s1

a	r(s1,a)	p([s1,s2]|s1,a)
a1	1	[0.6,0.4]
a2	0	[1,0]

State: s2

a	r(s2,a)	p([s1,s2]|s2,a)
a1	-1	[0.6,0.4]
a2	0	[0,1]

Let $\beta = 0.9$

Policy $\mathbf{f}^2$: (a1,a1)

- Evaluate value vector $\phi_\beta^2 := (2.302, 0.1637)$
- Policy improvement:
  - For state s1,
    consider a1: $1 + 0.9(0.6 * 2.302 + 0.4 * 0.1637) = 2.302$
    consider a2: $0 + 0.9(1 * 2.302 + 0 * 0.1637) = 2.0718$
  - For state s2,
    consider a1: $-1 + 0.9(0.6 * 2.302 + 0.4 * 0.1637) = 0.302$
    consider a2: $0 + 0.9(0 * 2.302 + 1 * 0.1637) = 0.14733$

An Example

State: s1

a	r(s1,a)	p([s1,s2]|s1,a)
a1	1	[0.6,0.4]
a2	0	[1,0]

State: s2

a	r(s2,a)	p([s1,s2]|s2,a)
a1	-1	[0.6,0.4]
a2	0	[0,1]

Let $\beta = 0.9$

Policy $\mathbf{f}^2$: (a1,a1)

- Evaluate value vector $\phi_\beta^2 := (2.302, 0.1637)$
- Policy improvement:
  - For state s1,
    consider a1: $1 + 0.9(0.6 * 2.302 + 0.4 * 0.1637) = 2.302$
    consider a2: $0 + 0.9(1 * 2.302 + 0 * 0.1637) = 2.0718$
  - For state s2,
    consider a1: $-1 + 0.9(0.6 * 2.302 + 0.4 * 0.1637) = 0.302$
    consider a2: $0 + 0.9(0 * 2.302 + 1 * 0.1637) = 0.14733$

An Example

State: s1

a	r(s1,a)	p([s1,s2]|s1,a)
a1	1	[0.6,0.4]
a2	0	[1,0]

State: s2

a	r(s2,a)	p([s1,s2]|s2,a)
a1	-1	[0.6,0.4]
a2	0	[0,1]

Let $\beta = 0.9$

Policy $\mathbf{f}^2$: (a1,a1)

- Evaluate value vector $\phi_\beta^2 := (2.302, 0.1637)$
- Policy improvement:
  - For state s1,
    consider a1: $1 + 0.9(0.6 * 2.302 + 0.4 * 0.1637) = \mathbf{2.302}$
    consider a2: $0 + 0.9(1 * 2.302 + 0 * 0.1637) = 2.0718$
  - For state s2,
    consider a1: $-1 + 0.9(0.6 * 2.302 + 0.4 * 0.1637) = \mathbf{0.302}$
    consider a2: $0 + 0.9(0 * 2.302 + 1 * 0.1637) = 0.14733$

No improvement! $\Rightarrow$ Optimal strategy: (a1,a1)

Thank You! Any Questions?