CS 747, Autumn 2022: Lecture 10

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

Autumn 2022

Markov Decision Problems

1. Action value function
2. Policy iteration
   - Policy improvement
   - Policy improvement theorem and proof
   - Policy iteration algorithm
3. History-dependent and stochastic policies

Markov Decision Problems

1. Action value function
2. Policy iteration
   - Policy improvement
   - Policy improvement theorem and proof
   - Policy iteration algorithm
3. History-dependent and stochastic policies

Action Value Function

- For $\pi \in \Pi$, $s \in S$, $a \in A$:

$$Q^\pi(s, a) \stackrel{\text{def}}{=} \mathbb{E}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s; a^0 = a; a^t = \pi(s^t) \text{ for } t \geq 1].$$

Action Value Function

- For $\pi \in \Pi, s \in S, a \in A$:

$$Q^{\pi}(s, a) \stackrel{\text{def}}{=} \mathbb{E}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s; a^0 = a; a^t = \pi(s^t) \text{ for } t \geq 1].$$

$Q^{\pi}(s, a)$ is the expected long-term reward from starting at state s, taking action a at $t = 0$, and following policy π for $t \geq 1$.

Action Value Function

- For $\pi \in \Pi, s \in S, a \in A$:
  $$Q^\pi(s, a) \stackrel{\text{def}}{=} \mathbb{E}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s; a^0 = a; a^t = \pi(s^t) \text{ for } t \geq 1].$$

  $Q^\pi(s, a)$ is the expected long-term reward from starting at state s, taking action a at $t = 0$, and following policy π for $t \geq 1$.
  $Q^\pi : S \times A \to \mathbb{R}$ is called the action value function of π.

Action Value Function

- For $\pi \in \Pi$, $s \in S$, $a \in A$:

  $$Q^\pi(s, a) \stackrel{\text{def}}{=} \mathbb{E}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s; a^0 = a; a^t = \pi(s^t) \text{ for } t \geq 1].$$

  $Q^\pi(s, a)$ is the expected long-term reward from starting at state s, taking action a at $t = 0$, and following policy π for $t \geq 1$.

  $Q^\pi : S \times A \to \mathbb{R}$ is called the action value function of π.

  Observe that Q^π satisfies, for $s \in S$, $a \in A$:

  $$Q^\pi(s, a) = \sum_{s' \in S} T(s, a, s')\{R(s, a, s') + \gamma V^\pi(s')\}.$$

Action Value Function

- For $\pi \in \Pi, s \in S, a \in A$:
  $Q^{\pi}(s, a) \stackrel{\text{def}}{=} \mathbb{E}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s; a^0 = a; a^t = \pi(s^t) \text{ for } t \geq 1]$.

  $Q^{\pi}(s, a)$ is the expected long-term reward from starting at state s, taking action a at $t = 0$, and following policy π for $t \geq 1$.
  $Q^{\pi} : S \times A \to \mathbb{R}$ is called the action value function of π.
  Observe that Q^{π} satisfies, for $s \in S, a \in A$:

$$Q^{\pi}(s, a) = \sum_{s' \in S} T(s, a, s')\{R(s, a, s') + \gamma V^{\pi}(s')\}.$$

  For $\pi \in \Pi, s \in S$: $Q^{\pi}(s, \pi(s)) = V^{\pi}(s)$.

Action Value Function

- For $\pi \in \Pi, s \in S, a \in A$:

  $Q^\pi(s, a) \stackrel{\text{def}}{=} \mathbb{E}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s; a^0 = a; a^t = \pi(s^t) \text{ for } t \geq 1]$.

  $Q^\pi(s, a)$ is the expected long-term reward from starting at state s, taking action a at $t = 0$, and following policy π for $t \geq 1$.

  $Q^\pi : S \times A \to \mathbb{R}$ is called the action value function of π.

  Observe that Q^π satisfies, for $s \in S, a \in A$:

$$Q^\pi(s, a) = \sum_{s' \in S} T(s, a, s')\{R(s, a, s') + \gamma V^\pi(s')\}.$$

  For $\pi \in \Pi, s \in S$: $Q^\pi(s, \pi(s)) = V^\pi(s)$.

- Q^π needs $O(n^2 k)$ operations to compute if V^π is available.

Action Value Function

- For $\pi \in \Pi, s \in S, a \in A$:

  $$Q^\pi(s,a) \stackrel{\text{def}}{=} \mathbb{E}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s; a^0 = a; a^t = \pi(s^t) \text{ for } t \geq 1].$$

  $Q^\pi(s,a)$ is the expected long-term reward from starting at state s, taking action a at $t = 0$, and following policy π for $t \geq 1$.

  $Q^\pi : S \times A \to \mathbb{R}$ is called the action value function of π.

  Observe that Q^π satisfies, for $s \in S, a \in A$:

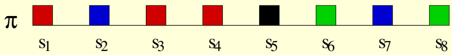
  $$Q^\pi(s,a) = \sum_{s' \in S} T(s,a,s')\{R(s,a,s') + \gamma V^\pi(s')\}.$$

  For $\pi \in \Pi, s \in S$: $Q^\pi(s, \pi(s)) = V^\pi(s)$.
- Q^π needs $O(n^2 k)$ operations to compute if V^π is available.
- All optimal policies have the same (optimal) action value function $Q^\star$.

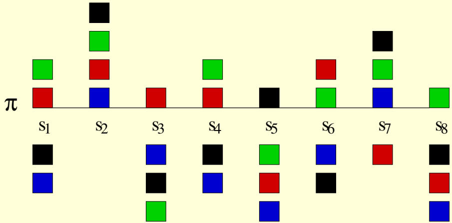
Markov Decision Problems

1. Action value function
2. Policy iteration
   - Policy improvement
   - Policy improvement theorem and proof
   - Policy iteration algorithm
3. History-dependent and stochastic policies

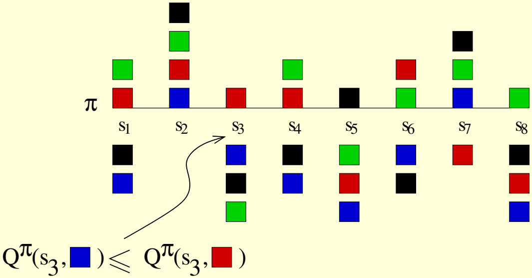
Policy Improvement



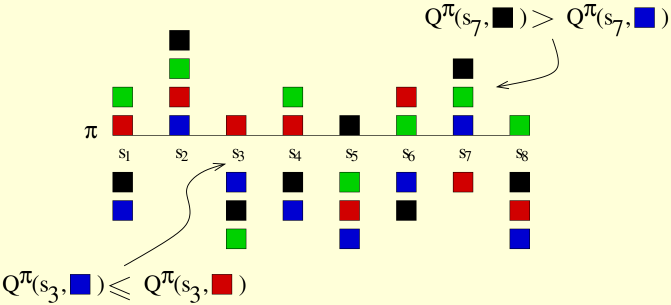
Policy Improvement



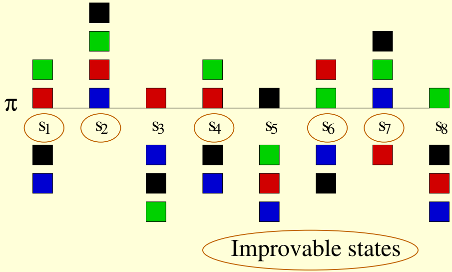
Policy Improvement



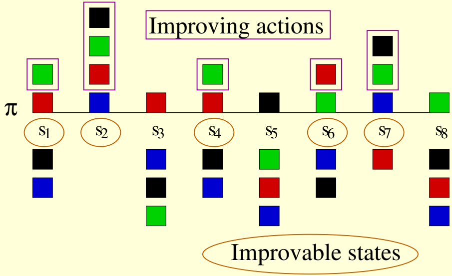
Policy Improvement



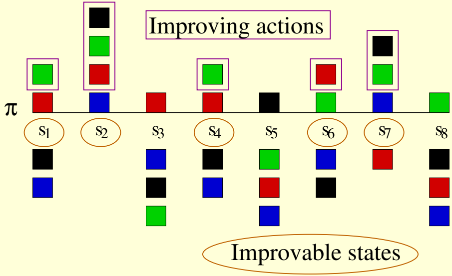
Policy Improvement



Policy Improvement



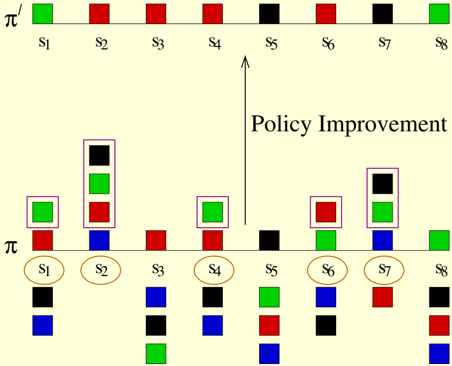
Policy Improvement



Given π,
- Pick one or more improvable states, and in these states,
- Switch to an arbitrary improving action.

Let the resulting policy be π'.

Policy Improvement



Given π,
- Pick one or more improvable states, and in these states,
- Switch to an arbitrary improving action.

Let the resulting policy be π'.

Markov Decision Problems

1. Action value function

2. Policy iteration
   - Policy improvement
   - Policy improvement theorem and proof
   - Policy iteration algorithm

3. History-dependent and stochastic policies

Policy Improvement Theorem

- For $\pi \in \Pi$, $s \in S$,

$$\mathbf{IA}(\pi, s) \stackrel{\text{def}}{=} \{a \in A : Q^{\pi}(s, a) > V^{\pi}(s)\}.$$

Policy Improvement Theorem

- For $\pi \in \Pi$, $s \in S$,

$$\mathbf{IA}(\pi, s) \stackrel{\text{def}}{=} \{a \in A : Q^{\pi}(s, a) > V^{\pi}(s)\}.$$

- For $\pi \in \Pi$,

$$\mathbf{IS}(\pi) \stackrel{\text{def}}{=} \{s \in S : |\mathbf{IA}(\pi, s)| \geq 1\}.$$

Policy Improvement Theorem

- For $\pi \in \Pi, s \in S$,

$$\mathbf{IA}(\pi, s) \stackrel{\text{def}}{=} \{a \in A : Q^\pi(s, a) > V^\pi(s)\}.$$

- For $\pi \in \Pi$,

$$\mathbf{IS}(\pi) \stackrel{\text{def}}{=} \{s \in S : |\mathbf{IA}(\pi, s)| \geq 1\}.$$

- Suppose $\mathbf{IS}(\pi) \neq \emptyset$ and $\pi' \in \Pi$ is obtained by policy improvement on π. Thus, π' satisfies

$$\forall s \in S : [\pi'(s) = \pi(s) \text{ or } \pi'(s) \in \mathbf{IA}(\pi, s)] \text{ and } \exists s \in S : \pi'(s) \in \mathbf{IA}(\pi, s).$$

Policy Improvement Theorem

- For $\pi \in \Pi, s \in S$,

$$\mathbf{IA}(\pi, s) \stackrel{\text{def}}{=} \{a \in A : Q^\pi(s, a) > V^\pi(s)\}.$$

- For $\pi \in \Pi$,

$$\mathbf{IS}(\pi) \stackrel{\text{def}}{=} \{s \in S : |\mathbf{IA}(\pi, s)| \geq 1\}.$$

- Suppose $\mathbf{IS}(\pi) \neq \emptyset$ and $\pi' \in \Pi$ is obtained by policy improvement on π. Thus, π' satisfies

$$\forall s \in S : [\pi'(s) = \pi(s) \text{ or } \pi'(s) \in \mathbf{IA}(\pi, s)] \text{ and } \exists s \in S : \pi'(s) \in \mathbf{IA}(\pi, s).$$

Policy Improvement Theorem:
(1) If $\mathbf{IS}(\pi) = \emptyset$, then π is optimal, else
(2) if π' is obtained by policy improvement on π, then $\pi' \succ \pi$.

Implication of Policy Improvement Theorem

Policy Improvement Theorem:
(1) If **IS**$(\pi) = \emptyset$, then π is optimal, else
(2) if π' is obtained by policy improvement on π, then $\pi' \succ \pi$.

Implication of Policy Improvement Theorem

Policy Improvement Theorem:
(1) If **IS**$(\pi) = \emptyset$, then π is optimal, else
(2) if π' is obtained by policy improvement on π, then $\pi' \succ \pi$.

- If $\pi \in \Pi$ is such that **IS**$(\pi) \neq \emptyset$, then there exists $\pi' \in \Pi$ such that $\pi' \succ \pi$.

Implication of Policy Improvement Theorem

> **Policy Improvement Theorem**:
> (1) If **IS**$(\pi) = \emptyset$, then π is optimal, else
> (2) if π' is obtained by policy improvement on π, then $\pi' \succ \pi$.

- If $\pi \in \Pi$ is such that **IS**$(\pi) \neq \emptyset$, then there exists $\pi' \in \Pi$ such that $\pi' \succ \pi$.
- But Π has a finite number of policies (k^n).

Implication of Policy Improvement Theorem

> **Policy Improvement Theorem**:
> (1) If $\mathbf{IS}(\pi) = \emptyset$, then π is optimal, else
> (2) if π' is obtained by policy improvement on π, then $\pi' \succ \pi$.

- If $\pi \in \Pi$ is such that $\mathbf{IS}(\pi) \neq \emptyset$, then there exists $\pi' \in \Pi$ such that $\pi' \succ \pi$.
- But Π has a finite number of policies (k^n).
- Hence, there must exist a policy $\pi^\star \in \Pi$ such that $\mathbf{IS}(\pi^\star) = \emptyset$.

Implication of Policy Improvement Theorem

Policy Improvement Theorem:
(1) If $\textbf{IS}(\pi) = \emptyset$, then π is optimal, else
(2) if π' is obtained by policy improvement on π, then $\pi' \succ \pi$.

- If $\pi \in \Pi$ is such that $\textbf{IS}(\pi) \neq \emptyset$, then there exists $\pi' \in \Pi$ such that $\pi' \succ \pi$.
- But Π has a finite number of policies (k^n).
- Hence, there must exist a policy $\pi^\star \in \Pi$ such that $\textbf{IS}(\pi^\star) = \emptyset$.
- The theorem itself also tells us that $\pi^\star$ must be optimal.

Implication of Policy Improvement Theorem

> **Policy Improvement Theorem**:
> (1) If **IS**$(\pi) = \emptyset$, then π is optimal, else
> (2) if π' is obtained by policy improvement on π, then $\pi' \succ \pi$.

- If $\pi \in \Pi$ is such that **IS**$(\pi) \neq \emptyset$, then there exists $\pi' \in \Pi$ such that $\pi' \succ \pi$.
- But Π has a finite number of policies (k^n).
- Hence, there must exist a policy $\pi^\star \in \Pi$ such that **IS**$(\pi^\star) = \emptyset$.
- The theorem itself also tells us that $\pi^\star$ must be optimal.
- Observe that **IS**$(\pi^\star) = \emptyset \iff B^\star(V^{\pi^\star}) = V^{\pi^\star}$.

Implication of Policy Improvement Theorem

Policy Improvement Theorem:
(1) If **IS**$(\pi) = \emptyset$, then π is optimal, else
(2) if π' is obtained by policy improvement on π, then $\pi' \succ \pi$.

- If $\pi \in \Pi$ is such that **IS**$(\pi) \neq \emptyset$, then there exists $\pi' \in \Pi$ such that $\pi' \succ \pi$.
- But Π has a finite number of policies (k^n).
- Hence, there must exist a policy $\pi^\star \in \Pi$ such that **IS**$(\pi^\star) = \emptyset$.
- The theorem itself also tells us that $\pi^\star$ must be optimal.
- Observe that **IS**$(\pi^\star) = \emptyset \iff B^\star(V^{\pi^\star}) = V^{\pi^\star}$.
- In other words, $V^{\pi^\star}$ satisfies the Bellman optimality equations—which we know has a unique solution. It is a convention to denote $V^{\pi^\star} = V^\star$.

Bellman Operator B^π

- For $\pi \in \Pi$, we define $B^\pi : \mathbb{R}^n \to \mathbb{R}^n$ as follows.
  For $X : S \to \mathbb{R}$ and for $s \in S$,

$$(B^\pi(X))(s) \stackrel{\text{def}}{=} \sum_{s' \in S} T(s, \pi(s), s') \left(R(s, \pi(s), s') + \gamma X(s') \right).$$

Bellman Operator B^{π}

- For $\pi \in \Pi$, we define $B^{\pi} : \mathbb{R}^n \to \mathbb{R}^n$ as follows.
  For $X : S \to \mathbb{R}$ and for $s \in S$,

$$(B^{\pi}(X))(s) \stackrel{\text{def}}{=} \sum_{s' \in S} T(s, \pi(s), s') \left(R(s, \pi(s), s') + \gamma X(s')\right).$$

- One Bellman operator for each $\pi \in \Pi$. No "max" like $B^{\star}$.

Bellman Operator B^π

- For $\pi \in \Pi$, we define $B^\pi : \mathbb{R}^n \to \mathbb{R}^n$ as follows.
  For $X : S \to \mathbb{R}$ and for $s \in S$,
  $$(B^\pi(X))(s) \stackrel{\text{def}}{=} \sum_{s' \in S} T(s, \pi(s), s') \left(R(s, \pi(s), s') + \gamma X(s')\right).$$
- One Bellman operator for each $\pi \in \Pi$. No "max" like $B^\star$.
- Some facts about B^π for all $\pi \in \Pi$. Similar proofs as for $B^\star$.
  - B^π is a contraction mapping with contraction factor γ.
  - For $X : S \to \mathbb{R}$: $\lim_{l \to \infty} (B^\pi)^l(X) = V^\pi$.
  - For $X : S \to \mathbb{R}$, $Y : S \to \mathbb{R}$: $X \succeq Y \implies B^\pi(X) \succeq B^\pi(Y)$.

Bellman Operator B^π

- For $\pi \in \Pi$, we define $B^\pi : \mathbb{R}^n \to \mathbb{R}^n$ as follows.
  For $X : S \to \mathbb{R}$ and for $s \in S$,

$$(B^\pi(X))(s) \stackrel{\text{def}}{=} \sum_{s' \in S} T(s, \pi(s), s') \left(R(s, \pi(s), s') + \gamma X(s') \right).$$

- One Bellman operator for each $\pi \in \Pi$. No "max" like $B^\star$.

- Some facts about B^π for all $\pi \in \Pi$. Similar proofs as for $B^\star$.
  - B^π is a contraction mapping with contraction factor γ.
  - For $X : S \to \mathbb{R}$: $\lim_{l \to \infty} (B^\pi)^l(X) = V^\pi$.
  - For $X : S \to \mathbb{R}$, $Y : S \to \mathbb{R}$: $X \succeq Y \implies B^\pi(X) \succeq B^\pi(Y)$.

- Observe that for $\pi, \pi' \in \Pi, \forall s \in S$: $B^{\pi'}(V^\pi)(s) = Q^\pi(s, \pi'(s))$.

Proof of Policy Improvement Theorem

$\textbf{IS}(\pi) = \emptyset$

Proof of Policy Improvement Theorem

$$\textbf{IS}(\pi) = \emptyset \implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi})$$

Proof of Policy Improvement Theorem

$$\mathbf{IS}(\pi) = \emptyset \implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi})$$
$$\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^{2}(V^{\pi})$$

Proof of Policy Improvement Theorem

$$
\begin{aligned}
\textbf{IS}(\pi) = \emptyset &\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \\
&\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \\
&\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \succeq \cdots \succeq \lim_{l \to \infty} (B^{\pi'})^l(V^{\pi})
\end{aligned}
$$

Proof of Policy Improvement Theorem

$$
\begin{aligned}
\textbf{IS}(\pi) = \emptyset \implies & \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \\
\implies & \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \\
\implies & \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \succeq \cdots \succeq \lim_{l \to \infty} (B^{\pi'})^l(V^{\pi}) \\
\implies & \forall \pi' \in \Pi : V^{\pi} \succeq V^{\pi'}.
\end{aligned}
$$

Proof of Policy Improvement Theorem

$$\begin{aligned}
\mathbf{IS}(\pi) = \emptyset \implies & \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \\
\implies & \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \\
\implies & \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \succeq \cdots \succeq \lim_{l \to \infty} (B^{\pi'})^l(V^{\pi}) \\
\implies & \forall \pi' \in \Pi : V^{\pi} \succeq V^{\pi'}.
\end{aligned}$$

$\mathbf{IS}(\pi) \neq \emptyset; \pi \xrightarrow{\text{P.I.}} \pi'$

Proof of Policy Improvement Theorem

$$\begin{aligned}
\mathbf{IS}(\pi) = \emptyset \implies & \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \\
\implies & \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \\
\implies & \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \succeq \cdots \succeq \lim_{l \to \infty} (B^{\pi'})^l(V^{\pi}) \\
\implies & \forall \pi' \in \Pi : V^{\pi} \succeq V^{\pi'}.
\end{aligned}$$

$$\mathbf{IS}(\pi) \neq \emptyset; \pi \xrightarrow{\text{P.I.}} \pi' \implies B^{\pi'}(V^{\pi}) \succ V^{\pi}$$

Proof of Policy Improvement Theorem

$$
\begin{aligned}
\textbf{IS}(\pi) = \emptyset &\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \\
&\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \\
&\implies \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \succeq \cdots \succeq \lim_{l \to \infty} (B^{\pi'})^l(V^{\pi}) \\
&\implies \forall \pi' \in \Pi : V^{\pi} \succeq V^{\pi'}.
\end{aligned}
$$

$$
\begin{aligned}
\textbf{IS}(\pi) \neq \emptyset; \pi \xrightarrow{\text{P.I.}} \pi' &\implies B^{\pi'}(V^{\pi}) \succ V^{\pi} \\
&\implies (B^{\pi'})^2(V^{\pi}) \succeq B^{\pi'}(V^{\pi}) \succ V^{\pi}
\end{aligned}
$$

Proof of Policy Improvement Theorem

$$\begin{aligned}
\mathbf{IS}(\pi) = \emptyset \implies & \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \\
\implies & \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \\
\implies & \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \succeq \cdots \succeq \lim_{l \to \infty} (B^{\pi'})^l(V^{\pi}) \\
\implies & \forall \pi' \in \Pi : V^{\pi} \succeq V^{\pi'}.
\end{aligned}$$

$$\begin{aligned}
\mathbf{IS}(\pi) \neq \emptyset; \pi \xrightarrow{\text{P.I.}} \pi' \implies & B^{\pi'}(V^{\pi}) \succ V^{\pi} \\
\implies & (B^{\pi'})^2(V^{\pi}) \succeq B^{\pi'}(V^{\pi}) \succ V^{\pi} \\
\implies & \lim_{l \to \infty} (B^{\pi'})^l(V^{\pi}) \succeq \cdots \succeq (B^{\pi'})^2(V^{\pi}) \succeq B^{\pi'}(V^{\pi}) \succ V^{\pi}
\end{aligned}$$

Proof of Policy Improvement Theorem

$$
\begin{aligned}
\mathbf{IS}(\pi) = \emptyset \implies & \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \\
\implies & \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \\
\implies & \forall \pi' \in \Pi : V^{\pi} \succeq B^{\pi'}(V^{\pi}) \succeq (B^{\pi'})^2(V^{\pi}) \succeq \cdots \succeq \lim_{l \to \infty} (B^{\pi'})^l(V^{\pi}) \\
\implies & \forall \pi' \in \Pi : V^{\pi} \succeq V^{\pi'}.
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{IS}(\pi) \neq \emptyset; \pi \xrightarrow{\text{P.I.}} \pi' \implies & B^{\pi'}(V^{\pi}) \succ V^{\pi} \\
\implies & (B^{\pi'})^2(V^{\pi}) \succeq B^{\pi'}(V^{\pi}) \succ V^{\pi} \\
\implies & \lim_{l \to \infty} (B^{\pi'})^l(V^{\pi}) \succeq \cdots \succeq (B^{\pi'})^2(V^{\pi}) \succeq B^{\pi'}(V^{\pi}) \succ V^{\pi} \\
\implies & V^{\pi'} \succ V^{\pi}.
\end{aligned}
$$

Markov Decision Problems

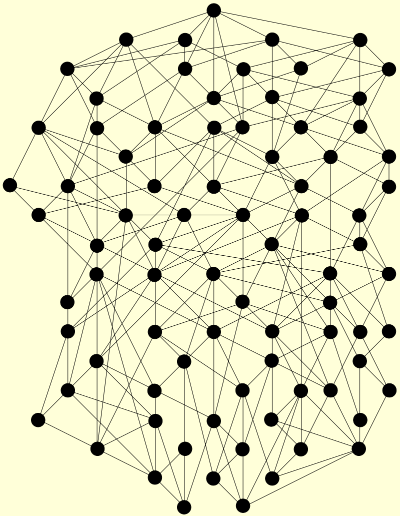
1. Action value function
2. Policy iteration
    - Policy improvement
    - Policy improvement theorem and proof
    - Policy iteration algorithm
3. History-dependent and stochastic policies

Policy Iteration Algorithm

$\pi \leftarrow$ Arbitrary policy.
While π has improvable states:
    $\pi' \leftarrow$ PolicyImprovement(π).
    $\pi \leftarrow \pi'$.
Return π.

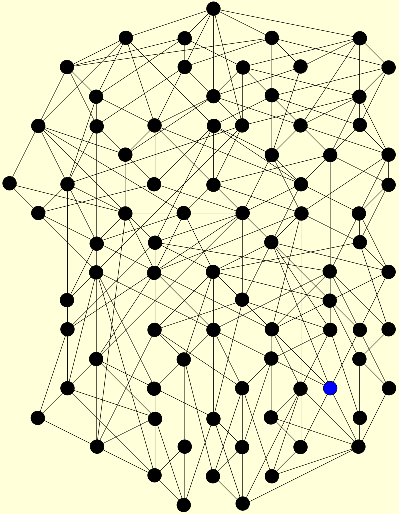
Policy Iteration Algorithm

$\pi \leftarrow$ Arbitrary policy.
While π has improvable states:
  $\pi' \leftarrow$ PolicyImprovement(π).
  $\pi \leftarrow \pi'$.
Return π.



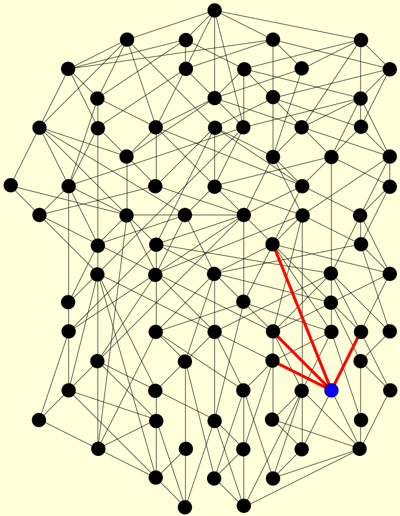
Policy Iteration Algorithm

$\pi \leftarrow$ Arbitrary policy.
While π has improvable states:
  $\pi' \leftarrow$ PolicyImprovement(π).
  $\pi \leftarrow \pi'$.
Return π.



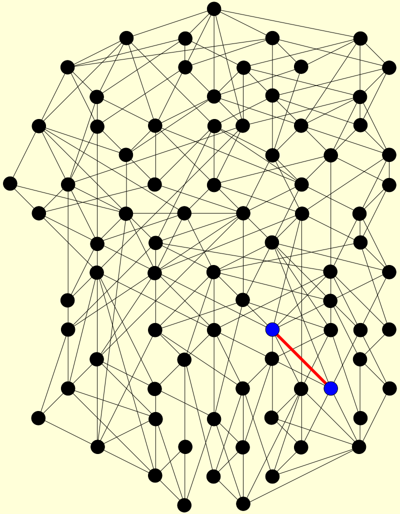
Policy Iteration Algorithm

$\pi \leftarrow$ Arbitrary policy.
While π has improvable states:
  $\pi' \leftarrow$ PolicyImprovement(π).
  $\pi \leftarrow \pi'$.
Return π.



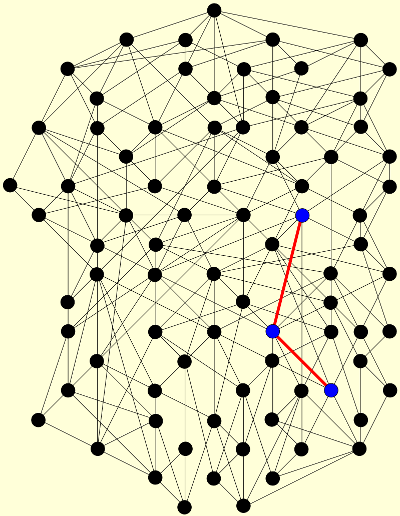
Policy Iteration Algorithm

$\pi \leftarrow$ Arbitrary policy.
While π has improvable states:
  $\pi' \leftarrow$ PolicyImprovement(π).
  $\pi \leftarrow \pi'$.
Return π.



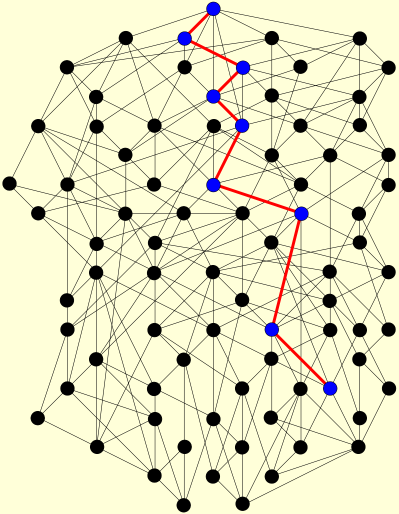
Policy Iteration Algorithm

$\pi \leftarrow$ Arbitrary policy.
While π has improvable states:
  $\pi' \leftarrow$ PolicyImprovement(π).
  $\pi \leftarrow \pi'$.
Return π.



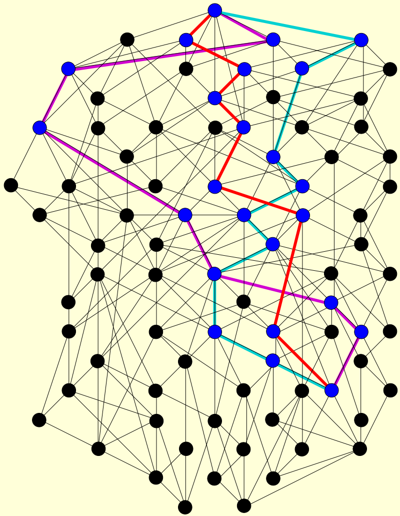
Policy Iteration Algorithm

$\pi \leftarrow$ Arbitrary policy.
While π has improvable states:
  $\pi' \leftarrow$ PolicyImprovement(π).
  $\pi \leftarrow \pi'$.
Return π.



Policy Iteration Algorithm

$\pi \leftarrow$ Arbitrary policy.
While π has improvable states:
  $\pi' \leftarrow$ PolicyImprovement(π).
  $\pi \leftarrow \pi'$.
Return π.



Path taken (and hence the number of iterations) in general depends on the switching strategy.

Markov Decision Problems

1. Action value function

2. Policy iteration
   - Policy improvement
   - Policy improvement theorem and proof
   - Policy iteration algorithm

3. History-dependent and stochastic policies

A More General Class of Policies

- In principle, an agent can follow a policy λ that maps every possible history $s^0, a^0, r^0, s^1, a^1, r^1, \dots, s^t$ for $t \geq 0$ to a probability distribution over A.
- Let Λ be the set of such policies λ (which are in general non-Markovian, non-stationary, and stochastic).

A More General Class of Policies

- In principle, an agent can follow a policy λ that maps every possible history $s^0, a^0, r^0, s^1, a^1, r^1, \ldots, s^t$ for $t \geq 0$ to a probability distribution over A.
- Let Λ be the set of such policies λ (which are in general non-Markovian, non-stationary, and stochastic).

- Recall that we only considered Π, the set of all policies $\pi : S \to A$ (which are Markovian, stationary, and deterministic). Observe that $\Pi \subset \Lambda$.
- We have shown that there exists $\pi^\star \in \Pi$ such that for all $\pi \in \Pi$, $\pi^\star \succeq \pi$.

A More General Class of Policies

- In principle, an agent can follow a policy λ that maps every possible history $s^0, a^0, r^0, s^1, a^1, r^1, \dots, s^t$ for $t \geq 0$ to a probability distribution over A.
- Let Λ be the set of such policies λ (which are in general non-Markovian, non-stationary, and stochastic).

- Recall that we only considered Π, the set of all policies $\pi : S \to A$ (which are Markovian, stationary, and deterministic). Observe that $\Pi \subset \Lambda$.
- We have shown that there exists $\pi^\star \in \Pi$ such that for all $\pi \in \Pi$, $\pi^\star \succeq \pi$.

Could there exist $\lambda \in \Lambda \setminus \Pi$ such that $\neg(\pi^\star \succeq \lambda)$?

A More General Class of Policies

- In principle, an agent can follow a policy λ that maps every possible history $s^0, a^0, r^0, s^1, a^1, r^1, \dots, s^t$ for $t \geq 0$ to a probability distribution over A.
- Let Λ be the set of such policies λ (which are in general non-Markovian, non-stationary, and stochastic).

- Recall that we only considered Π, the set of all policies $\pi : S \to A$ (which are Markovian, stationary, and deterministic). Observe that $\Pi \subset \Lambda$.
- We have shown that there exists $\pi^\star \in \Pi$ such that for all $\pi \in \Pi$, $\pi^\star \succeq \pi$.

Could there exist $\lambda \in \Lambda \setminus \Pi$ such that $\neg(\pi^\star \succeq \lambda)$? No.

History and Stochasticity

- In MDPs, the agent can sense state, and the consequence of each action depends solely on state.

History and Stochasticity

- In MDPs, the agent can sense state, and the consequence of each action depends solely on state.
- We are maximising an infinite sum of expected discounted rewards—the challenge at each time step is the same: to maximise the expected long-term reward starting from the current state!

History and Stochasticity

- In MDPs, the agent can sense state, and the consequence of each action depends solely on state.
- We are maximising an infinite sum of expected discounted rewards—the challenge at each time step is the same: to maximise the expected long-term reward starting from the current state!

- History and stochasticity can help if the agent is unable to sense state perfectly. Such a situation arises in an abstraction called the Partially Observable MDP (POMDP).

History and Stochasticity

- In MDPs, the agent can sense state, and the consequence of each action depends solely on state.
- We are maximising an infinite sum of expected discounted rewards—the challenge at each time step is the same: to maximise the expected long-term reward starting from the current state!

- History and stochasticity can help if the agent is unable to sense state perfectly. Such a situation arises in an abstraction called the Partially Observable MDP (POMDP).
- Optimal policies for the finite horizon reward setting are in general non-stationary (time-dependent).

History and Stochasticity

- In MDPs, the agent can sense state, and the consequence of each action depends solely on state.
- We are maximising an infinite sum of expected discounted rewards—the challenge at each time step is the same: to maximise the expected long-term reward starting from the current state!

- History and stochasticity can help if the agent is unable to sense state perfectly. Such a situation arises in an abstraction called the Partially Observable MDP (POMDP).
- Optimal policies for the finite horizon reward setting are in general non-stationary (time-dependent).
- Optimal policies (“strategies”) in many types of multi-player games are in general stochastic (“mixed”) because the next state depends on all the players’ actions, but each player chooses only their own.

Markov Decision Problems

1. Action value function

2. Policy iteration
   - Policy improvement
   - Policy improvement theorem and proof
   - Policy iteration algorithm

3. History-dependent and stochastic policies

Markov Decision Problems

1. Action value function

2. Policy iteration
   - Policy improvement
   - Policy improvement theorem and proof
   - Policy iteration algorithm

3. History-dependent and stochastic policies

Next class: Running time of policy iteration, review of MDP planning.