CS 747, Autumn 2022: Lecture 6

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

Autumn 2022

Markov Decision Problems

1. Definitions
    - Markov Decision Problem
    - Policy
    - Value Function

2. MDP planning

3. Policy evaluation

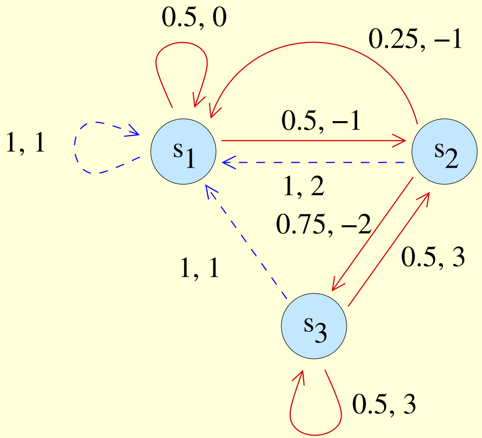
Markov Decision Problems

1. Definitions
    - Markov Decision Problem
    - Policy
    - Value Function

2. MDP planning

3. Policy evaluation

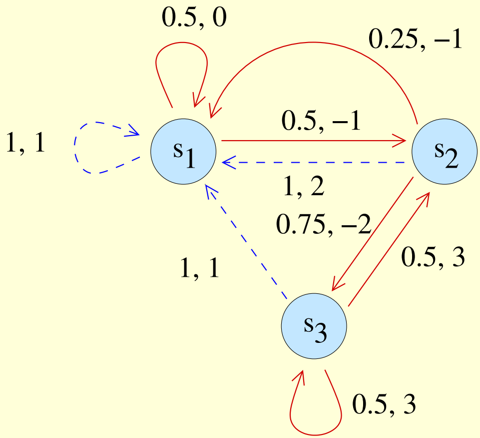
Markov Decision Problems (MDPs)



Markov Decision Problems (MDPs)

Elements of MDP $M = (S, A, T, R, \gamma)$.

S: a set of states.

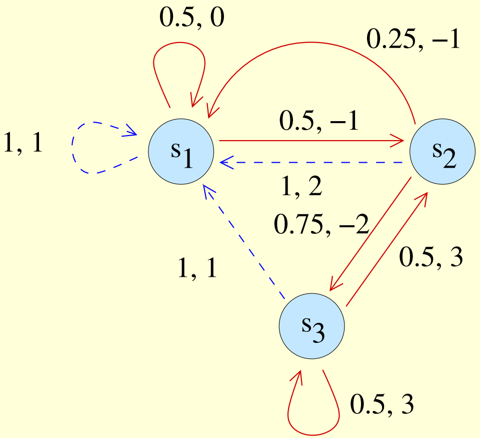


Markov Decision Problems (MDPs)

Elements of MDP $M = (S, A, T, R, \gamma)$.

S: a set of states.

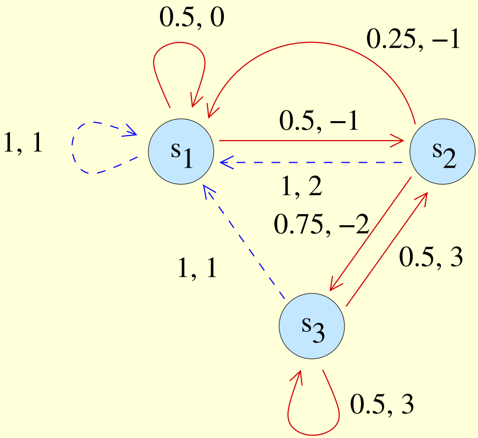
Let us assume $S = \{s_1, s_2, \dots, s_n\}$, and hence $|S| = n$.



Markov Decision Problems (MDPs)

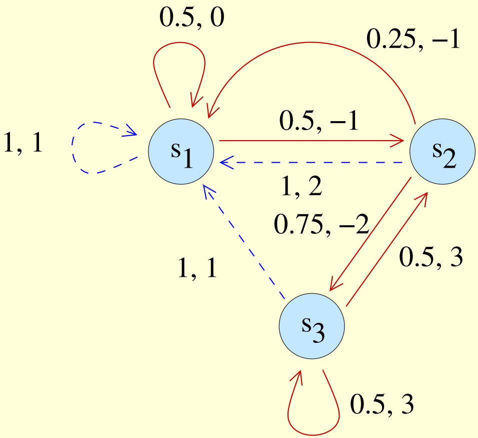
Elements of MDP $M = (S, A, T, R, \gamma)$.

A: a set of actions.



Markov Decision Problems (MDPs)

Elements of MDP $M = (S, A, T, R, \gamma)$.



A: a set of actions.

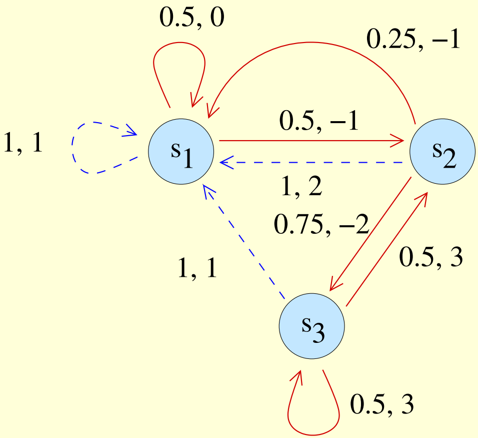
Let us assume $A = \{a_1, a_2, \ldots, a_k\}$, and hence $|A| = k$.

Here $A = \{\text{RED}, \text{BLUE}\}$.

Markov Decision Problems (MDPs)

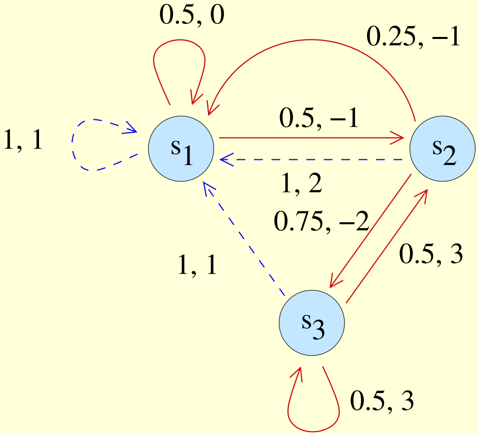
Elements of MDP $M = (S, A, T, R, \gamma)$.

T: a transition function.



Markov Decision Problems (MDPs)

Elements of MDP $M = (S, A, T, R, \gamma)$.



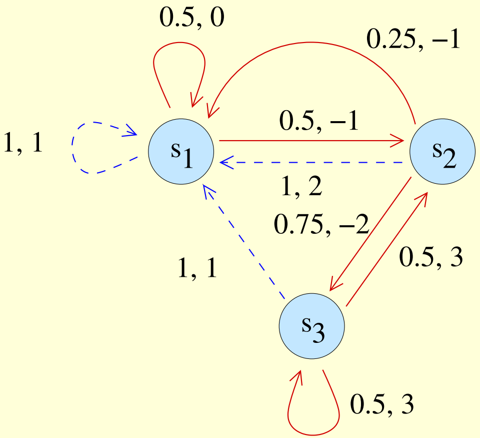
T: a transition function.

- For $s, s' \in S, a \in A$: $T(s, a, s')$ is the probability of reaching s' by starting at s and taking action a.
- Thus, $T(s, a, \cdot)$ is a probability distribution over S.

Markov Decision Problems (MDPs)

Elements of MDP $M = (S, A, T, R, \gamma)$.

R: a reward function.

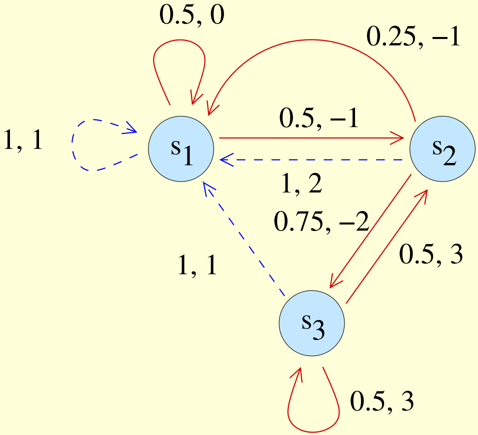


Markov Decision Problems (MDPs)

Elements of MDP $M = (S, A, T, R, \gamma)$.

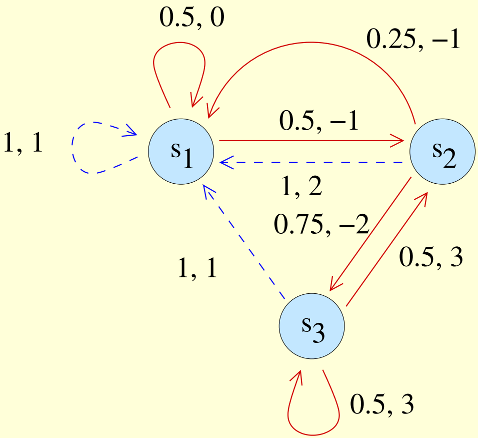
R: a reward function.

- For $s, s' \in S, a \in A$: $R(s, a, s')$ is the (numeric) reward for reaching s' by starting at s and taking action a.
- Assume rewards are from $[-R_{\max}, R_{\max}]$ for some $R_{\max} \geq 0$.



Markov Decision Problems (MDPs)

Elements of MDP $M = (S, A, T, R, \gamma)$.



γ, a discount factor—coming up.

Agent-Environment Interaction

$t = 0$

Agent is born in some state s^0, takes action a^0.

Environment generates and provides the agent

next state $s^1 \sim T(s^0, a^0, \cdot)$ and

reward $r^0 = R(s^0, a^0, s^1)$.

Agent-Environment Interaction

$t = 0$

Agent is born in some state s^0, takes action a^0.
Environment generates and provides the agent
next state $s^1 \sim T(s^0, a^0, \cdot)$ and
reward $r^0 = R(s^0, a^0, s^1)$.

$t = 1$

Agent is in state s^1, takes action a^1.
Environment generates and provides the agent
next state $s^2 \sim T(s^1, a^1, \cdot)$ and
reward $r^1 = R(s^1, a^1, s^2)$.

Agent-Environment Interaction

$t = 0$

Agent is born in some state s^0, takes action a^0.
Environment generates and provides the agent
next state $s^1 \sim T(s^0, a^0, \cdot)$ and
reward $r^0 = R(s^0, a^0, s^1)$.

$t = 1$

Agent is in state s^1, takes action a^1.
Environment generates and provides the agent
next state $s^2 \sim T(s^1, a^1, \cdot)$ and
reward $r^1 = R(s^1, a^1, s^2)$.

$\vdots$

Agent-Environment Interaction

$t = 0$

Agent is born in some state s^0, takes action a^0.
Environment generates and provides the agent
next state $s^1 \sim T(s^0, a^0, \cdot)$ and
reward $r^0 = R(s^0, a^0, s^1)$.

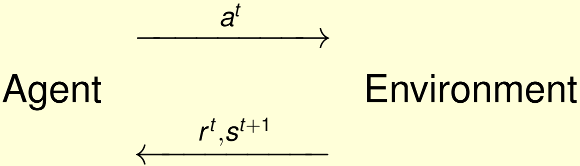
$t = 1$

Agent is in state s^1, takes action a^1.
Environment generates and provides the agent
next state $s^2 \sim T(s^1, a^1, \cdot)$ and
reward $r^1 = R(s^1, a^1, s^2)$.

$\vdots$

Resulting trajectory: $s^0, a^0, r^0, s^1, a^1, r^1, s^2, \ldots$.

Describing the Agent’s Behaviour



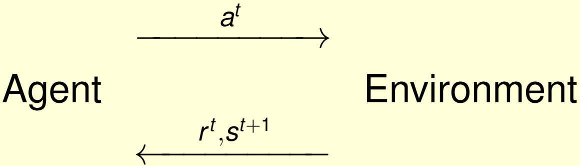
a^t

Agent

Environment

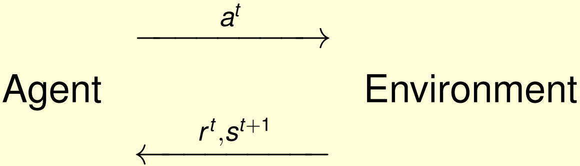
r^t, s^{t+1}

Describing the Agent’s Behaviour



- How does the agent pick a^t?

Describing the Agent's Behaviour



a^t

Agent

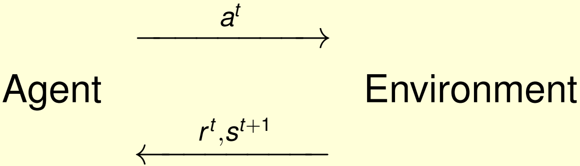
Environment

r^t, s^{t+1}

- How does the agent pick a^t?
  In principle, it can decide by looking at the preceding history

$$s^0, a^0, r^0, s^1, a^1, r^1, s^2, \ldots, s^t.$$

Describing the Agent's Behaviour



a^t

Agent

Environment

r^t, s^{t+1}

- How does the agent pick a^t?

  In principle, it can decide by looking at the preceding history

  $$s^0, a^0, r^0, s^1, a^1, r^1, s^2, \ldots, s^t.$$

  For now let us assume that a^t is picked based on s^t alone.

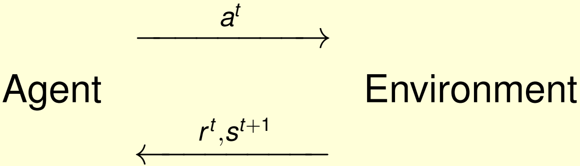
Describing the Agent's Behaviour

$$\begin{tikzcd}
\text{Agent} \arrow[r, shift left=2, "a^t"] & \text{Environment} \arrow[l, shift left=2, "{r^t, s^{t+1}}"]
\end{tikzcd}$$

- How does the agent pick a^t?
  In principle, it can decide by looking at the preceding history
  $$s^0, a^0, r^0, s^1, a^1, r^1, s^2, \ldots, s^t.$$
  For now let us assume that a^t is picked based on s^t alone.

- In other words, the agent follows a policy $\pi : S \to A$.

Describing the Agent's Behaviour



a^t

Agent

Environment

r^t, s^{t+1}

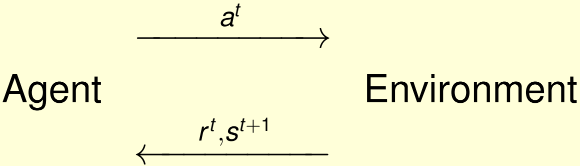
- How does the agent pick a^t?
  In principle, it can decide by looking at the preceding history

  $$s^0, a^0, r^0, s^1, a^1, r^1, s^2, \ldots, s^t.$$

  For now let us assume that a^t is picked based on s^t alone.

- In other words, the agent follows a policy $\pi : S \to A$.
  Observe that π is Markovian, deterministic, and stationary.

Describing the Agent's Behaviour



a^t

Agent

Environment

r^t, s^{t+1}

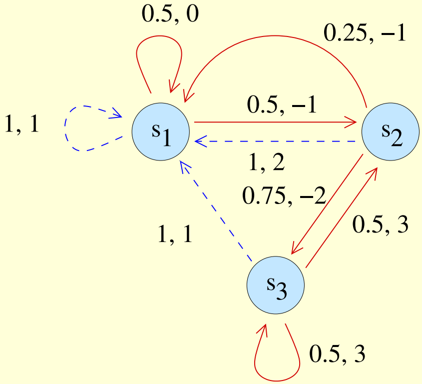
- How does the agent pick a^t?
  In principle, it can decide by looking at the preceding history

  $$s^0, a^0, r^0, s^1, a^1, r^1, s^2, \ldots, s^t.$$

  For now let us assume that a^t is picked based on s^t alone.

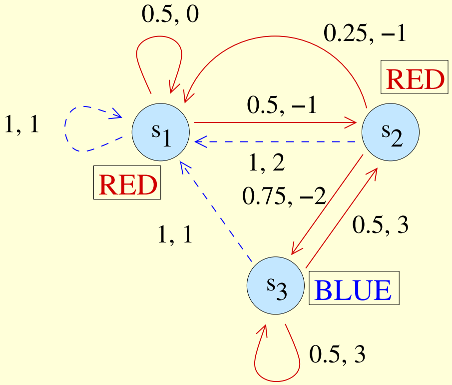
- In other words, the agent follows a policy $\pi : S \to A$.
  Observe that π is Markovian, deterministic, and stationary.
  We will justify this choice in due course!

Illustration: Policy



-

Illustration: Policy



-

Illustration: Policy

- Illustrated policy π such that

$$\pi(s_1) = \text{RED}; \pi(s_2) = \text{RED}; \pi(s_3) = \text{BLUE}.$$

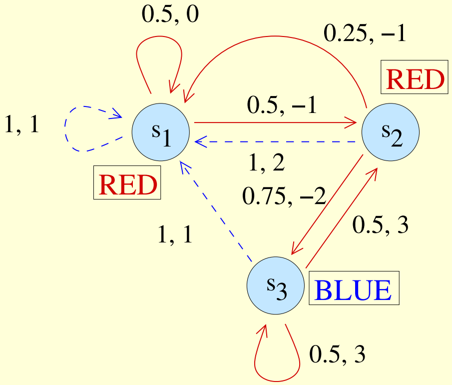


Illustration: Policy

- Illustrated policy π such that

  $\pi(s_1) = \text{RED}; \pi(s_2) = \text{RED}; \pi(s_3) = \text{BLUE}.$

  What happens by "following" π, starting at s_1?

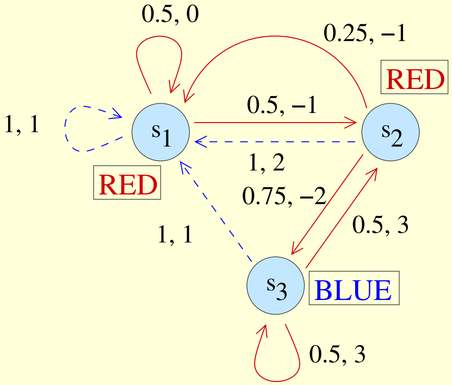


Illustration: Policy

- Illustrated policy π such that

  $\pi(s_1) =$ RED; $\pi(s_2) =$ RED; $\pi(s_3) =$ BLUE.

  What happens by "following" π, starting at s_1?

  - s_1, RED, s_1, RED, s_2, RED, s_3, BLUE, s_1, ...
  - s_1, RED, s_2, RED, s_1, RED, s_1, RED, s_1, ...
  -

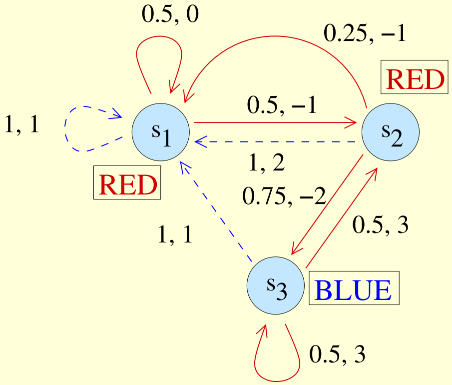


Illustration: Policy

- Let Π denote the set of all policies.

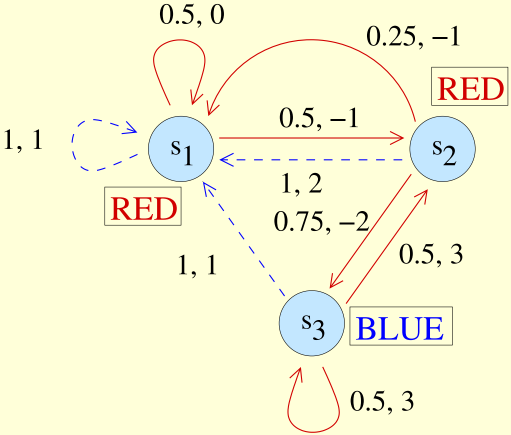


Illustration: Policy

- Let Π denote the set of all policies.
- What is $|\Pi|$?

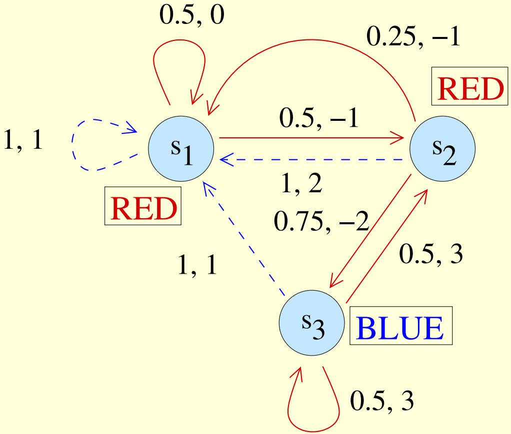


Illustration: Policy

- Let Π denote the set of all policies.
- What is $|\Pi|$? k^n.

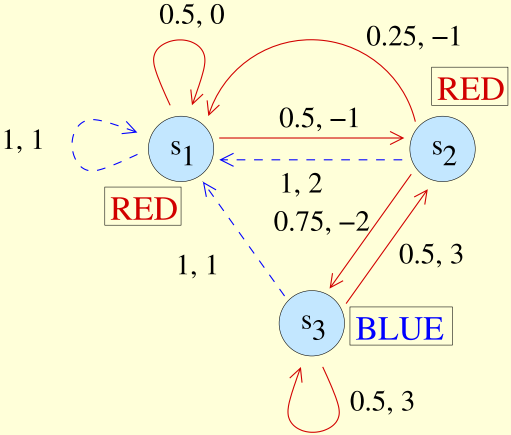
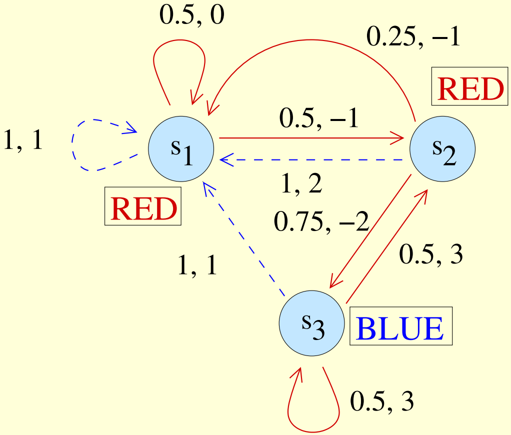


Illustration: Policy

- Let Π denote the set of all policies.
- What is $|\Pi|$? k^n.
- Which $\pi \in \Pi$ is a “good” policy?



State Values for Policy π

- For $s \in S$, $V^{\pi}(s) \overset{\text{def}}{=} \mathbb{E}_{\pi}\left[r^0 + \quad r^1 + \quad\quad r^2 + \quad\quad r^3 + \ldots \middle| s^0 = s\right]$,

State Values for Policy π

- For $s \in S$, $V^{\pi}(s) \overset{\text{def}}{=} \mathbb{E}_{\pi}\left[r^0 + \gamma r^1 + \gamma^2 r^2 + \gamma^3 r^3 + \dots | s^0 = s\right]$, where $\gamma \in [0, 1)$ is a discount factor.

State Values for Policy π

- For $s \in S$, $V^{\pi}(s) \overset{\text{def}}{=} \mathbb{E}_{\pi}\left[r^0 + \gamma r^1 + \gamma^2 r^2 + \gamma^3 r^3 + \ldots | s^0 = s\right]$,
  where $\gamma \in [0, 1)$ is a discount factor.

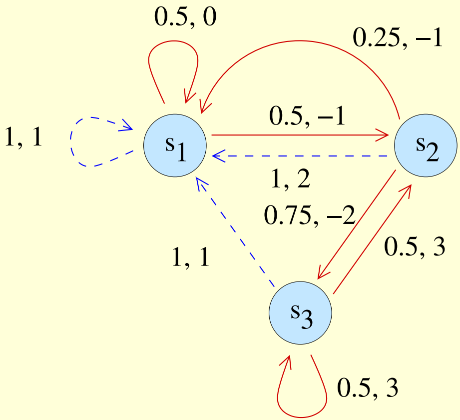
- γ is an element of the MDP.

  Larger γ, farther the "lookahead".

State Values for Policy π

- For $s \in S$, $V^{\pi}(s) \overset{\text{def}}{=} \mathbb{E}_{\pi}\left[r^0 + \gamma r^1 + \gamma^2 r^2 + \gamma^3 r^3 + \ldots | s^0 = s\right]$, where $\gamma \in [0, 1)$ is a discount factor.

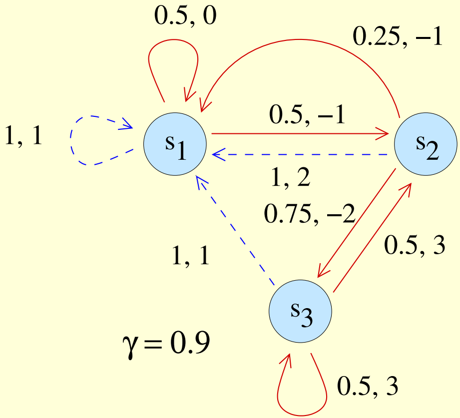
- γ is an element of the MDP.
  Larger γ, farther the "lookahead".



State Values for Policy π

- For $s \in S$, $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}\left[r^0 + \gamma r^1 + \gamma^2 r^2 + \gamma^3 r^3 + \dots | s^0 = s\right]$, where $\gamma \in [0, 1)$ is a discount factor.

- γ is an element of the MDP.
  Larger γ, farther the "lookahead".

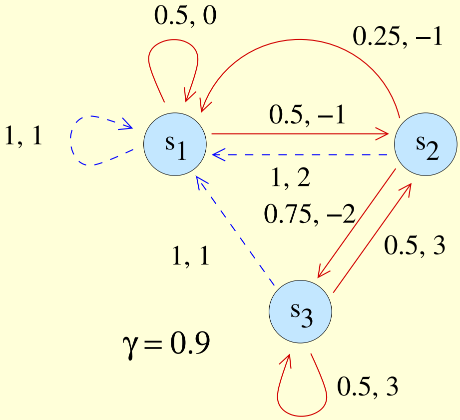


State Values for Policy π

- For $s \in S$, $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}\left[r^0 + \gamma r^1 + \gamma^2 r^2 + \gamma^3 r^3 + \ldots | s^0 = s\right]$, where $\gamma \in [0, 1)$ is a discount factor.

- γ is an element of the MDP. Larger γ, farther the "lookahead".

- $V^{\pi}(s)$ is the value of state s under policy π.



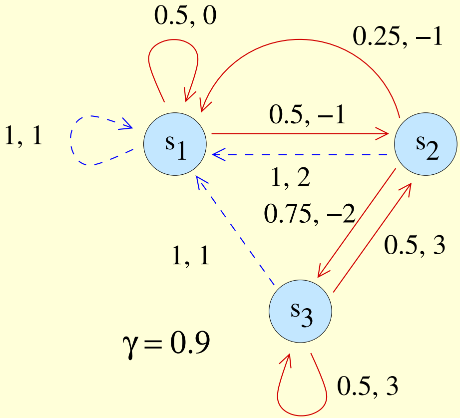
State Values for Policy π

- For $s \in S$, $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}\left[r^0 + \gamma r^1 + \gamma^2 r^2 + \gamma^3 r^3 + \ldots \mid s^0 = s\right]$, where $\gamma \in [0, 1)$ is a discount factor.

- γ is an element of the MDP. Larger γ, farther the "lookahead".

- $V^{\pi}(s)$ is the value of state s under policy π.

- $V^{\pi} : S \to \mathbb{R}$ is the value function of π.



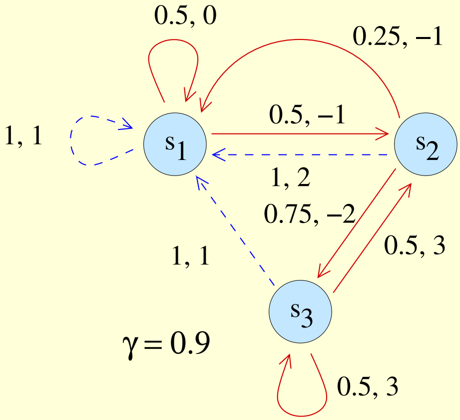
State Values for Policy π

- For $s \in S$, $V^{\pi}(s) \overset{\text{def}}{=} \mathbb{E}_{\pi}\left[r^0 + \gamma r^1 + \gamma^2 r^2 + \gamma^3 r^3 + \ldots | s^0 = s\right]$,
  where $\gamma \in [0, 1)$ is a discount factor.

- γ is an element of the MDP.
  Larger γ, farther the "lookahead".

- $V^{\pi}(s)$ is the value of state s under policy π.

- $V^{\pi} : S \to \mathbb{R}$ is the value function of π.
  "Larger is better".



Markov Decision Problems

1. Definitions
    - Markov Decision Problem
    - Policy
    - Value Function

2. MDP planning

3. Policy evaluation

Optimal Policies

- Here are value functions from our example MDP.

π	$V^\pi(s_1)$	$V^\pi(s_2)$	$V^\pi(s_3)$
RRR	4.45	6.55	10.82
RRB	-5.61	-5.75	-4.05
RBR	2.76	4.48	9.12
RBB	2.76	4.48	3.48
BRR	10.0	9.34	13.10
BRB	10.0	7.25	10.0
BBR	10.0	11 .0	14.45
BBB	10.0	11.0	10.0

Optimal Policies

- Here are value functions from our example MDP.

π	$V^\pi(s_1)$	$V^\pi(s_2)$	$V^\pi(s_3)$
RRR	4.45	6.55	10.82
RRB	-5.61	-5.75	-4.05
RBR	2.76	4.48	9.12
RBB	2.76	4.48	3.48
BRR	10.0	9.34	13.10
BRB	10.0	7.25	10.0
BBR	10.0	11 .0	14.45
BBB	10.0	11.0	10.0

Which policy would you prefer?

Optimal Policies

- Here are value functions from our example MDP.

π	$V^\pi(s_1)$	$V^\pi(s_2)$	$V^\pi(s_3)$	
RRR	4.45	6.55	10.82	
RRB	-5.61	-5.75	-4.05	
RBR	2.76	4.48	9.12	
RBB	2.76	4.48	3.48	
BRR	10.0	9.34	13.10	
BRB	10.0	7.25	10.0	
BBR	**10.0**	**11.0**	**14.45**	← Optimal policy
BBB	10.0	11.0	10.0	

Which policy would you prefer?

Optimal Policies

- Here are value functions from our example MDP.

π	$V^\pi(s_1)$	$V^\pi(s_2)$	$V^\pi(s_3)$	
RRR	4.45	6.55	10.82	
RRB	-5.61	-5.75	-4.05	
RBR	2.76	4.48	9.12	
RBB	2.76	4.48	3.48	
BRR	10.0	9.34	13.10	
BRB	10.0	7.25	10.0	
BBR	**10.0**	**11.0**	**14.45**	← Optimal policy
BBB	10.0	11.0	10.0	

Which policy would you prefer?

Every MDP is guaranteed to have an optimal policy $\pi^\star$ s.t.

$$\forall \pi \in \Pi, \forall s \in S : V^{\pi^\star}(s) \geq V^\pi(s).$$

MDP Planning

MDP Planning problem: Given $M = (S, A, T, R, \gamma)$, find a policy $\pi^\star$ from the set of all policies Π such that $\forall s \in S, \forall \pi \in \Pi$: $V^{\pi^\star}(s) \geq V^{\pi}(s)$.

MDP Planning

MDP Planning problem: Given $M = (S, A, T, R, \gamma)$, find a policy $\pi^\star$ from the set of all policies Π such that $\forall s \in S, \forall \pi \in \Pi$: $V^{\pi^\star}(s) \geq V^{\pi}(s)$.

- Every MDP is guaranteed to have a deterministic, Markovian, stationary optimal policy.

MDP Planning

MDP Planning problem: Given $M = (S, A, T, R, \gamma)$, find a policy $\pi^\star$ from the set of all policies Π such that $\forall s \in S, \forall \pi \in \Pi$: $V^{\pi^\star}(s) \geq V^{\pi}(s)$.

- Every MDP is guaranteed to have a deterministic, Markovian, stationary optimal policy.
- An MDP can have more than one optimal policy.

MDP Planning

MDP Planning problem: Given $M = (S, A, T, R, \gamma)$, find a policy $\pi^\star$ from the set of all policies Π such that $\forall s \in S, \forall \pi \in \Pi$: $V^{\pi^\star}(s) \geq V^{\pi}(s)$.

- Every MDP is guaranteed to have a deterministic, Markovian, stationary optimal policy.
- An MDP can have more than one optimal policy.
- However, the value function of every optimal policy is the same, unique "optimal value function" $V^\star$.

Markov Decision Problems

1. Definitions
   - Markov Decision Problem
   - Policy
   - Value Function
2. MDP planning
3. Policy Evaluation

Structure of State Values

For $\pi \in \Pi, s \in S: V^{\pi}(s) \overset{\text{def}}{=} \mathbb{E}_{\pi}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s]$

Structure of State Values

$$\text{For } \pi \in \Pi, s \in S: \; V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s]$$

$$= \sum_{s' \in S} T(s, \pi(s), s') \mathbb{E}_{\pi}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s, s^1 = s']$$

Structure of State Values

For $\pi \in \Pi, s \in S$:
$$\begin{aligned}
V^\pi(s) &\overset{\text{def}}{=} \mathbb{E}_\pi[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s] \\
&= \sum_{s' \in S} T(s, \pi(s), s') \mathbb{E}_\pi[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s, s^1 = s'] \\
&= \sum_{s' \in S} T(s, \pi(s), s') \mathbb{E}_\pi[r^0 | s^0 = s, s^1 = s'] \\
&\quad + \gamma \sum_{s' \in S} T(s, \pi(s), s') \mathbb{E}_\pi[r^1 + \gamma r^2 + \ldots | s^0 = s, s^1 = s']
\end{aligned}$$

Structure of State Values

$$
\begin{aligned}
\text{For } \pi \in \Pi, s \in S: \; V^{\pi}(s) &\overset{\text{def}}{=} \mathbb{E}_{\pi}[r^0 + \gamma r^1 + \gamma^2 r^2 + \dots | s^0 = s] \\
&= \sum_{s' \in S} T(s, \pi(s), s') \mathbb{E}_{\pi}[r^0 + \gamma r^1 + \gamma^2 r^2 + \dots | s^0 = s, s^1 = s'] \\
&= \sum_{s' \in S} T(s, \pi(s), s') \mathbb{E}_{\pi}[r^0 | s^0 = s, s^1 = s'] \\
&\quad + \gamma \sum_{s' \in S} T(s, \pi(s), s') \mathbb{E}_{\pi}[r^1 + \gamma r^2 + \dots | s^0 = s, s^1 = s'] \\
&= \sum_{s' \in S} T(s, \pi(s), s') R(s, \pi(s), s') \\
&\quad + \gamma \sum_{s' \in S} T(s, \pi(s), s') \mathbb{E}_{\pi}[r^1 + \gamma r^2 + \dots | s^1 = s']
\end{aligned}
$$

Structure of State Values

$$
\begin{aligned}
\text{For } \pi \in \Pi, s \in S: \; V^\pi(s) &\overset{\text{def}}{=} \mathbb{E}_\pi[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s] \\
&= \sum_{s' \in S} T(s, \pi(s), s') \mathbb{E}_\pi[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s, s^1 = s'] \\
&= \sum_{s' \in S} T(s, \pi(s), s') \mathbb{E}_\pi[r^0 | s^0 = s, s^1 = s'] \\
&\quad + \gamma \sum_{s' \in S} T(s, \pi(s), s') \mathbb{E}_\pi[r^1 + \gamma r^2 + \ldots | s^0 = s, s^1 = s'] \\
&= \sum_{s' \in S} T(s, \pi(s), s') R(s, \pi(s), s') \\
&\quad + \gamma \sum_{s' \in S} T(s, \pi(s), s') \mathbb{E}_\pi[r^1 + \gamma r^2 + \ldots | s^1 = s'] \\
&= \sum_{s' \in S} T(s, \pi(s), s') \left\{ R(s, \pi(s), s') + \gamma V^\pi(s') \right\}.
\end{aligned}
$$

Bellman Equations

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

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

Bellman Equations

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

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

- Recall that $S = \{s_1, s_2, \ldots, s_n\}$.

Bellman Equations

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

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

- Recall that $S = \{s_1, s_2, \dots, s_n\}$.
- n equations, n unknowns—$V^{\pi}(s_1), V^{\pi}(s_2), \dots, V^{\pi}(s_n)$.

Bellman Equations

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

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

- Recall that $S = \{s_1, s_2, \dots, s_n\}$.
- n equations, n unknowns—$V^{\pi}(s_1)$, $V^{\pi}(s_2)$, $\dots$, $V^{\pi}(s_n)$.
- Linear!

Bellman Equations

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

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

- Recall that $S = \{s_1, s_2, \dots, s_n\}$.
- n equations, n unknowns—$V^{\pi}(s_1), V^{\pi}(s_2), \dots, V^{\pi}(s_n)$.
- Linear!
- Guaranteed to have a unique solution if $\gamma < 1$.

Bellman Equations

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

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

- Recall that $S = \{s_1, s_2, \ldots, s_n\}$.
- n equations, n unknowns—$V^{\pi}(s_1), V^{\pi}(s_2), \ldots, V^{\pi}(s_n)$.
- Linear!
- Guaranteed to have a unique solution if $\gamma < 1$.
- Policy evaluation: step of computing V^{π} for a given policy π.

Are We Done with this Topic?

Are We Done with this Topic?

- We claimed that among all the policies for a given MDP, there must be an optimal policy $\pi^\star$.

Are We Done with this Topic?

- We claimed that among all the policies for a given MDP, there must be an optimal policy $\pi^\star$.

- Now you know how to compute the value function of any given policy π.

Are We Done with this Topic?

- We claimed that among all the policies for a given MDP, there must be an optimal policy $\pi^\star$.

- Now you know how to compute the value function of any given policy π.

- Can you put the two ideas together and construct an algorithm to find $\pi^\star$?

Are We Done with this Topic?

- We claimed that among all the policies for a given MDP, there must be an optimal policy $\pi^\star$.

- Now you know how to compute the value function of any given policy π.

- Can you put the two ideas together and construct an algorithm to find $\pi^\star$?

- Yes! Evaluate each policy and identify one that has a value function dominating all the others'.

Are We Done with this Topic?

- We claimed that among all the policies for a given MDP, there must be an optimal policy $\pi^\star$.
- Now you know how to compute the value function of any given policy π.
- Can you put the two ideas together and construct an algorithm to find $\pi^\star$?
- Yes! Evaluate each policy and identify one that has a value function dominating all the others'.
- This approach needs poly$(n, k) \cdot k^n$ arithmetic operations. We hope to be more efficient (wait for next week).