CS 747, Autumn 2020: Week 4, Lecture 1

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

Autumn 2020

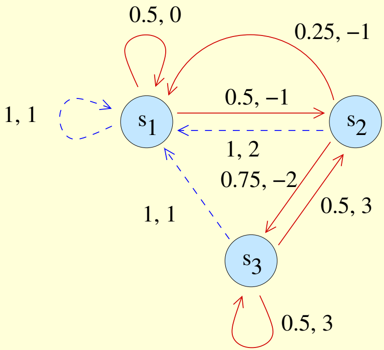
Markov Decision Problems

1. Definitions
   - Markov Decision Problem
   - Policy
   - Value Function
2. MDP planning
3. Alternative formulations
4. Applications
5. Policy Evaluation

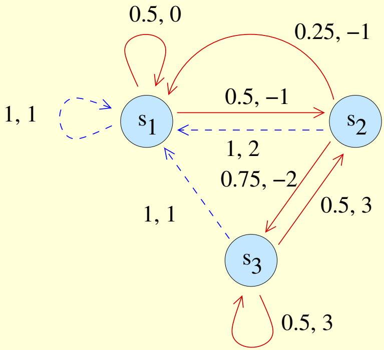
Markov Decision Problems

1. Definitions
   - Markov Decision Problem
   - Policy
   - Value Function
2. MDP planning
3. Alternative formulations
4. Applications
5. Policy Evaluation

Markov Decision Problems (MDPs)



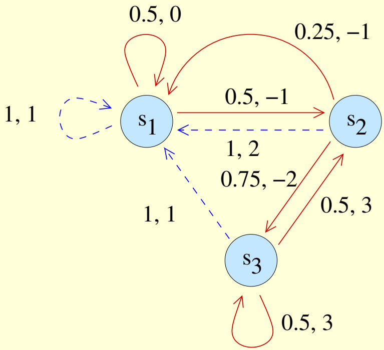
Markov Decision Problems (MDPs)



An MDP $M = (S, A, T, R, \gamma)$ has these elements.

S: a set of states.

Markov Decision Problems (MDPs)

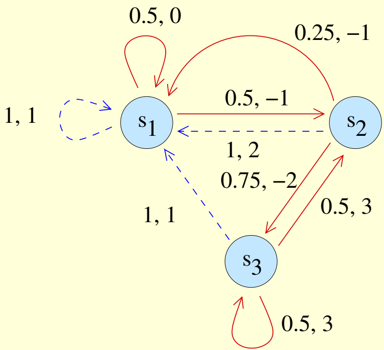


An MDP $M = (S, A, T, R, \gamma)$ has these elements.

S: a set of states.

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

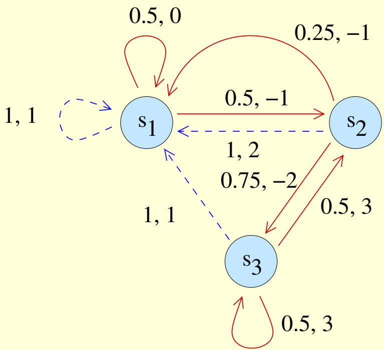
Markov Decision Problems (MDPs)



An MDP $M = (S, A, T, R, \gamma)$ has these elements.

A: a set of actions.

Markov Decision Problems (MDPs)



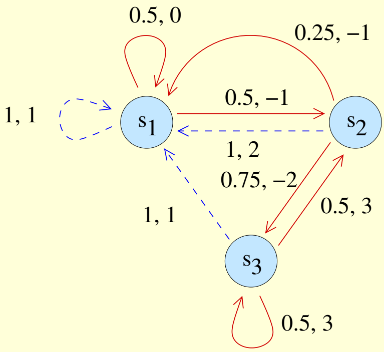
An MDP $M = (S, A, T, R, \gamma)$ has these elements.

A: a set of actions.

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

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

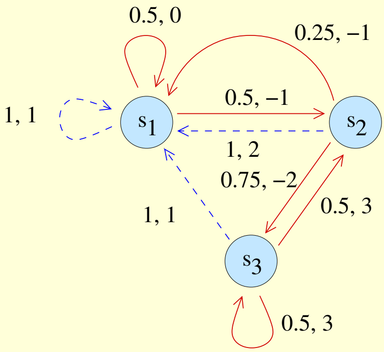
Markov Decision Problems (MDPs)



An MDP $M = (S, A, T, R, \gamma)$ has these elements.

T: a transition function.

Markov Decision Problems (MDPs)

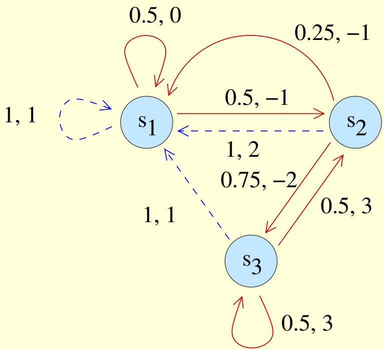


An MDP $M = (S, A, T, R, \gamma)$ has these elements.

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.

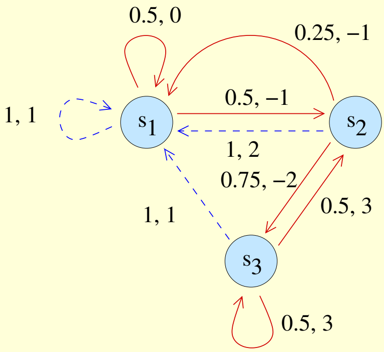
Markov Decision Problems (MDPs)



An MDP $M = (S, A, T, R, \gamma)$ has these elements.

R: a reward function.

Markov Decision Problems (MDPs)

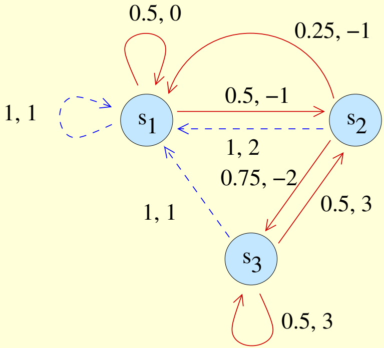


An MDP $M = (S, A, T, R, \gamma)$ has these elements.

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)



An MDP $M = (S, A, T, R, \gamma)$ has these elements.

γ, a discount factor—coming up shortly.

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

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

Describing the Agent's Behaviour

$$\text{Agent} \underset{r^t, s^{t+1}}{\overset{a^t}{\rightleftarrows}} \text{Environment}$$

- How does the agent pick a^t?

Describing the Agent's Behaviour

$$\xrightarrow{\quad a^t \quad}$$

Agent Environment

$$\xleftarrow{\quad r^t, s^{t+1} \quad}$$

- 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

$$\text{Agent} \underset{r^t, s^{t+1}}{\overset{a^t}{\rightleftarrows}} \text{Environment}$$

- 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

$$\text{Agent} \quad \underset{\xleftarrow[\quad\quad\quad]{r^t, s^{t+1}}}{\xrightarrow{\quad a^t \quad}} \quad \text{Environment}$$

- 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

$$\text{Agent} \quad \underset{\xleftarrow[]{r^t, s^{t+1}}}{\xrightarrow{a^t}} \quad \text{Environment}$$

- 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 $\pi : S \rightarrow A$.
  Observe that π is Markovian, deterministic, and stationary.

Describing the Agent's Behaviour

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

- 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 $\pi : S \rightarrow A$.

  Observe that π is Markovian, deterministic, and stationary.

  We will justify this choice in due course!

Illustration: Policy

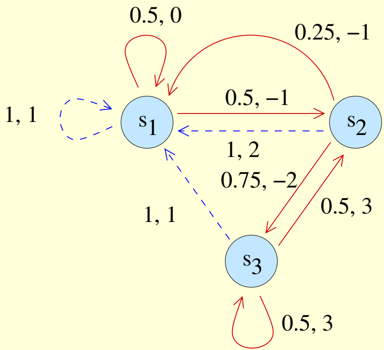


Illustration: Policy

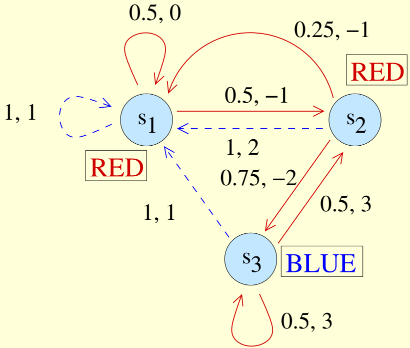
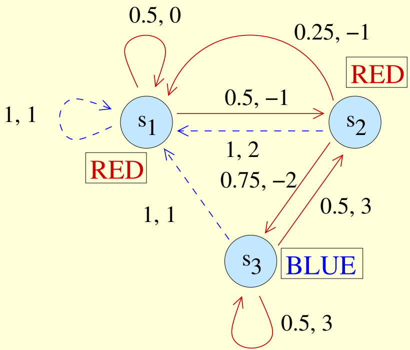


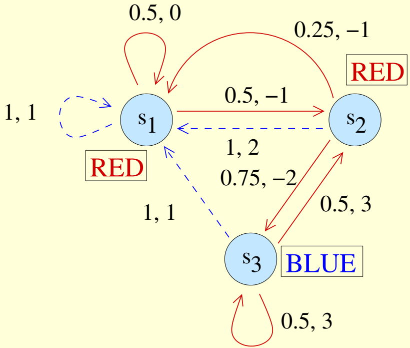
Illustration: Policy



- Illustrated policy π such that

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

Illustration: Policy

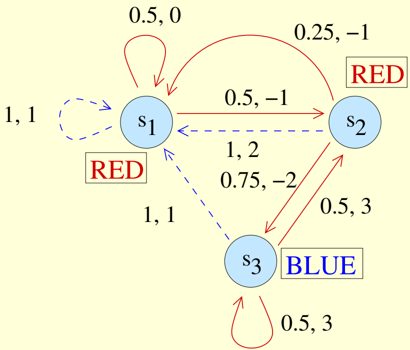


- Illustrated policy π such that

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

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

Illustration: Policy



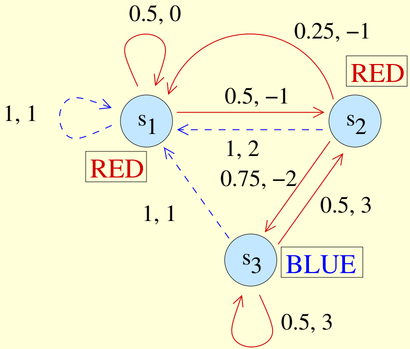
- Illustrated policy π such that

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

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

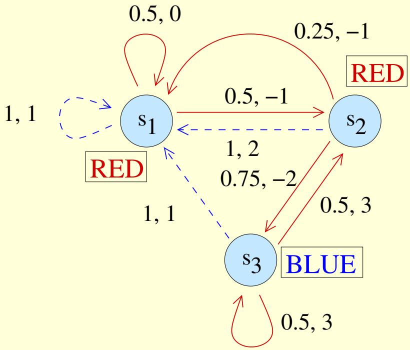
  - s_1, RED, s_1, RED, s_2, RED, s_3, BLUE, s_1, ...
  - s_1, RED, s_2, RED, s_1, RED, s_1, RED, s_1, . . .

Illustration: Policy



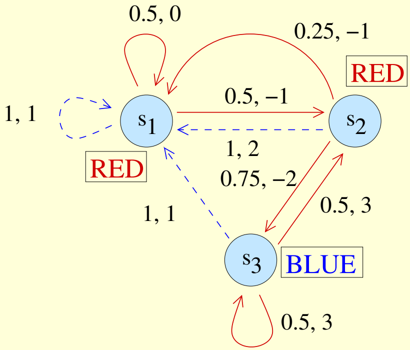
- Let Π denote the set of all policies.

Illustration: Policy



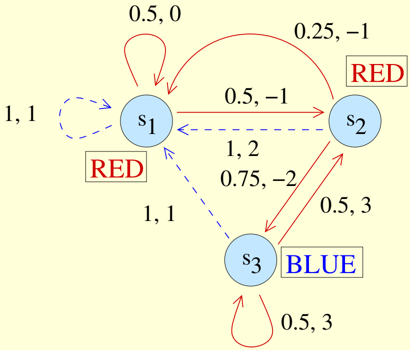
- Let Π denote the set of all policies.
- What is $|\Pi|$?

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

State Values for Policy π

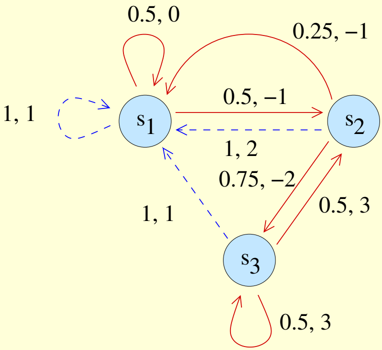
- For $s \in S$, $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}\left[r^0 + \gamma r^1 + \gamma^2 r^2 + \gamma^3 r^3 + \ldots | s^0 = s\right]$ where $\gamma \in [0, 1)$ is a discount factor.

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 "lookahead".

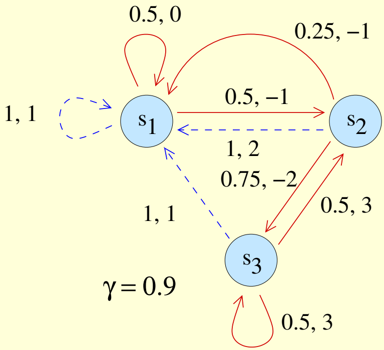
State Values for Policy π

- For $s \in S$, $V^\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi \left[r^0 + \gamma r^1 + \gamma^2 r^2 + \gamma^3 r^3 + \ldots \middle| s^0 = s \right]$ where $\gamma \in [0, 1)$ is a discount factor.
- γ is an element of the MDP. Larger γ, farther "lookahead".



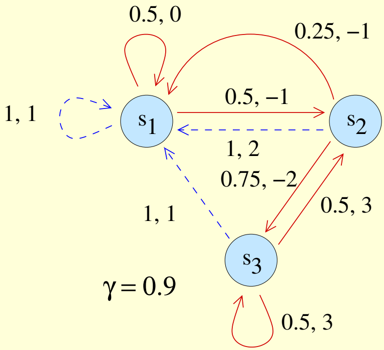
State Values for Policy π

- 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 "lookahead".



State Values for Policy π

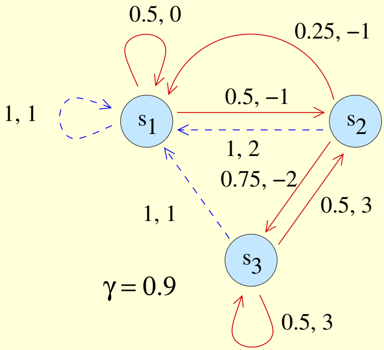
- For $s \in S$, $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}\left[r^0 + \gamma r^1 + \gamma^2 r^2 + \gamma^3 r^3 + \dots | s^0 = s\right]$ where $\gamma \in [0, 1)$ is a discount factor.
- γ is an element of the MDP. Larger γ, farther "lookahead".



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

State Values for Policy π

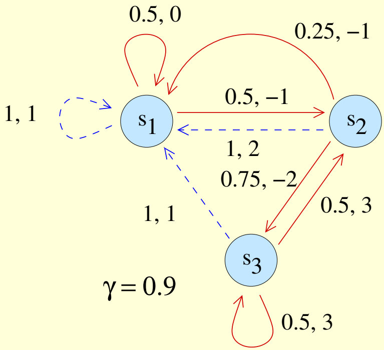
- For $s \in S$, $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}\left[r^0 + \gamma r^1 + \gamma^2 r^2 + \gamma^3 r^3 + \ldots | s^0 = s\right]$ where $\gamma \in [0, 1)$ is a discount factor.
- γ is an element of the MDP. Larger γ, farther "lookahead".



- $V^{\pi}(s)$ 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 $s \in S$, $V^\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi\left[r^0 + \gamma r^1 + \gamma^2 r^2 + \gamma^3 r^3 + \ldots | s^0 = s\right]$ where $\gamma \in [0, 1)$ is a discount factor.
- γ is an element of the MDP. Larger γ, farther "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. Alternative formulations
4. Applications
5. 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**	$\leftarrow$ 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**	$\leftarrow$ 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. Alternative formulations

4. Applications

5. Policy Evaluation

Reward and Transition Functions

- We had assumed

$$T : S \times A \times S \to [0, 1], R : S \times A \times S \to [-R_{\max}, R_{\max}].$$

Reward and Transition Functions

- We had assumed

$$T : S \times A \times S \to [0, 1], R : S \times A \times S \to [-R_{\max}, R_{\max}].$$

- You might encounter alternative definitions of R, T.

Reward and Transition Functions

- We had assumed

$$T : S \times A \times S \to [0, 1], R : S \times A \times S \to [-R_{\max}, R_{\max}].$$

- You might encounter alternative definitions of R, T.
- Sometimes $R(s, a, s')$ is taken as a random variable bounded in $[-R_{\max}, R_{\max}]$.

Reward and Transition Functions

- We had assumed

$$T : S \times A \times S \to [0, 1], R : S \times A \times S \to [-R_{\max}, R_{\max}].$$

- You might encounter alternative definitions of R, T.
- Sometimes $R(s, a, s')$ is taken as a random variable bounded in $[-R_{\max}, R_{\max}]$.
- Sometimes there is a reward $R(s, a)$ given on taking action a from state s, regardless of next state s'.

Reward and Transition Functions

- We had assumed

$$T : S \times A \times S \to [0, 1], R : S \times A \times S \to [-R_{\max}, R_{\max}].$$

- You might encounter alternative definitions of R, T.
- Sometimes $R(s, a, s')$ is taken as a random variable bounded in $[-R_{\max}, R_{\max}]$.
- Sometimes there is a reward $R(s, a)$ given on taking action a from state s, regardless of next state s'.
- Sometimes there is a reward $R(s')$ given on reaching next state s', regardless of start state s and action a.

Reward and Transition Functions

- We had assumed

$$T : S \times A \times S \to [0, 1], R : S \times A \times S \to [-R_{\max}, R_{\max}].$$

- You might encounter alternative definitions of R, T.
- Sometimes $R(s, a, s')$ is taken as a random variable bounded in $[-R_{\max}, R_{\max}]$.
- Sometimes there is a reward $R(s, a)$ given on taking action a from state s, regardless of next state s'.
- Sometimes there is a reward $R(s')$ given on reaching next state s', regardless of start state s and action a.
- Sometimes T and R are combined into a single function $\mathbb{P}\{s', r | s, a\}$ for $s' \in S, r \in [-R_{\max}, R_{\max}]$.

Reward and Transition Functions

- We had assumed

$$T : S \times A \times S \to [0, 1], R : S \times A \times S \to [-R_{\max}, R_{\max}].$$

- You might encounter alternative definitions of R, T.
- Sometimes $R(s, a, s')$ is taken as a random variable bounded in $[-R_{\max}, R_{\max}]$.
- Sometimes there is a reward $R(s, a)$ given on taking action a from state s, regardless of next state s'.
- Sometimes there is a reward $R(s')$ given on reaching next state s', regardless of start state s and action a.
- Sometimes T and R are combined into a single function $\mathbb{P}\{s', r | s, a\}$ for $s' \in S, r \in [-R_{\max}, R_{\max}]$.
- Some authors minimise cost rather than maximise reward.

Reward and Transition Functions

- We had assumed

$$T : S \times A \times S \to [0, 1], R : S \times A \times S \to [-R_{\max}, R_{\max}].$$

- You might encounter alternative definitions of R, T.
- Sometimes $R(s, a, s')$ is taken as a random variable bounded in $[-R_{\max}, R_{\max}]$.
- Sometimes there is a reward $R(s, a)$ given on taking action a from state s, regardless of next state s'.
- Sometimes there is a reward $R(s')$ given on reaching next state s', regardless of start state s and action a.
- Sometimes T and R are combined into a single function $\mathbb{P}\{s', r | s, a\}$ for $s' \in S, r \in [-R_{\max}, R_{\max}]$.
- Some authors minimise cost rather than maximise reward.

- It is relatively straightforward to handle all these variations.

Episodic Tasks

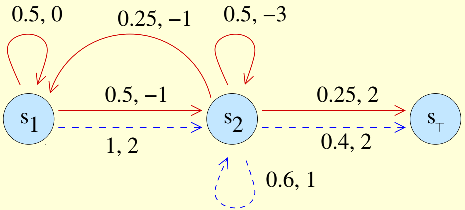
- We considered continuing tasks, in which trajectories are infinitely long.

Episodic Tasks

- We considered continuing tasks, in which trajectories are infinitely long.
- Episodic tasks have a special sink/terminal state $s_\top$ from which there are no outgoing transitions on rewards.

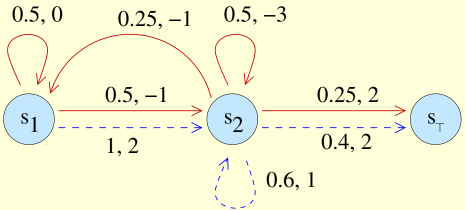
Episodic Tasks

- We considered continuing tasks, in which trajectories are infinitely long.
- Episodic tasks have a special sink/terminal state $s_\top$ from which there are no outgoing transitions on rewards.



Episodic Tasks

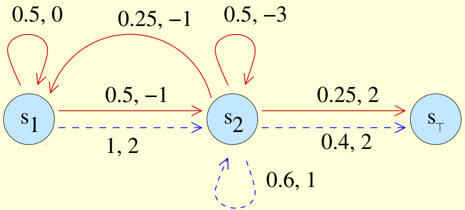
- We considered continuing tasks, in which trajectories are infinitely long.
- Episodic tasks have a special sink/terminal state $s_\top$ from which there are no outgoing transitions on rewards.



- Additionally, from every non-terminal state and for every policy, there is a non-zero probability of reaching the terminal state in a finite number of steps.

Episodic Tasks

- We considered continuing tasks, in which trajectories are infinitely long.
- Episodic tasks have a special sink/terminal state $s_\top$ from which there are no outgoing transitions on rewards.



- Additionally, from every non-terminal state and for every policy, there is a non-zero probability of reaching the terminal state in a finite number of steps.
- Hence, trajectories or episodes almost surely terminate after a finite number of steps.

Definition of Values

- We defined $V^{\pi}(s)$ as an **Infinite discounted reward**:
  $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s]$.

Definition of Values

- We defined $V^{\pi}(s)$ as an **Infinite discounted reward**:
  $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s]$.
  There are other choices.
- **Total reward**:
  $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}[r^0 + r^1 + r^2 + \ldots | s^0 = s]$.
  Can only be used on episodic tasks.

Definition of Values

- We defined $V^{\pi}(s)$ as an **Infinite discounted reward**:
  $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s]$.
  There are other choices.
- **Total reward**:
  $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}[r^0 + r^1 + r^2 + \ldots | s^0 = s]$.
  Can only be used on episodic tasks.
- **Finite horizon reward**:
  $V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}[r^0 + r^1 + r^2 + \cdots + r^{T-1} | s^0 = s]$.
  Horizon $T \geq 1$ specified, rather than γ.
  Optimal policies for this setting need not be stationary.

Definition of Values

- We defined $V^\pi(s)$ as an **Infinite discounted reward**:
  $V^\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s]$.
  There are other choices.
- **Total reward**:
  $V^\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi[r^0 + r^1 + r^2 + \ldots | s^0 = s]$.
  Can only be used on episodic tasks.
- **Finite horizon reward**:
  $V^\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi[r^0 + r^1 + r^2 + \cdots + r^{T-1} | s^0 = s]$.
  Horizon $T \geq 1$ specified, rather than γ.
  Optimal policies for this setting need not be stationary.
- **Average reward** (withholding some technical details):
  $V^\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi[\lim_{m \to \infty} \frac{r^0 + r^1 + \cdots + r^{m-1}}{m} | s^0 = s]$.

Markov Decision Problems

1. Definitions
   - Markov Decision Problem
   - Policy
   - Value Function
2. MDP planning
3. Alternative formulations
4. Applications
5. Policy Evaluation

Controlling a Helicopter (Ng *et al.*, 2003)

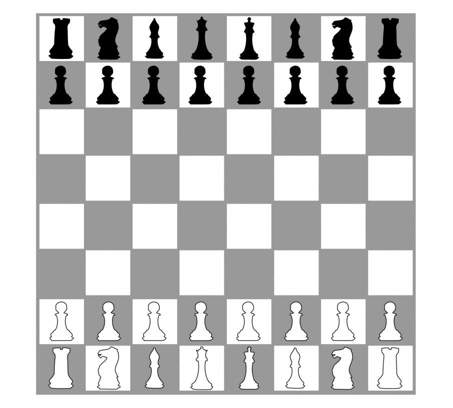
- Episodic or continuing task? What are S, A, T, R, γ?

[1]

1. `https://www.publicdomainpictures.net/pictures/20000/velka/police-helicopter-8712919948643Mk.jpg`.

Succeeding at Chess

- Episodic or continuing task? What are S, A, T, R, γ?



[1]

1. https://www.publicdomainpictures.net/pictures/80000/velka/chess-board-and-pieces.jpg.

Preventing Forest Fires (Lauer *et al.*, 2017)

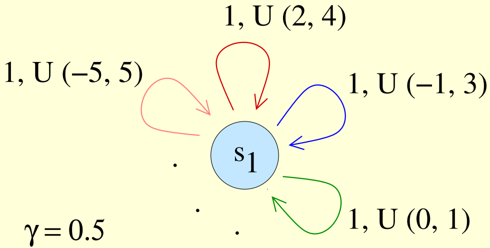
- Episodic or continuing task? What are S, A, T, R, γ?

[1]

1. https://www.publicdomainpictures.net/pictures/270000/velka/firemen-1533752293Zsu.jpg.

A Familiar MDP?

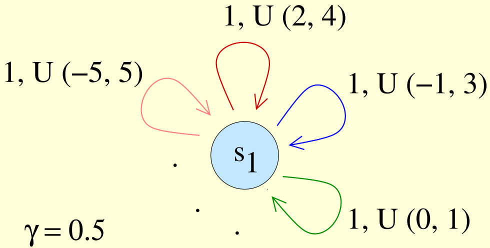
- Single state. k actions.
- For $a \in A$, treat $R(s, a, s')$ as a random variable.



Annotation: "probability, reward distribution".

A Familiar MDP?

- Single state. k actions.
- For $a \in A$, treat $R(s, a, s')$ as a random variable.

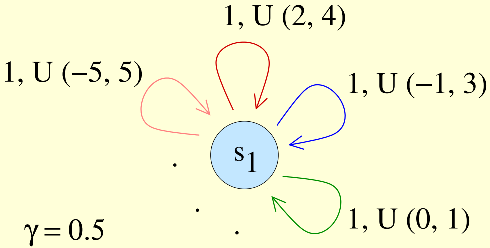


Annotation: "probability, reward distribution".

- Such an MDP is called a

A Familiar MDP?

- Single state. k actions.
- For $a \in A$, treat $R(s, a, s')$ as a random variable.



Annotation: "probability, reward distribution".

- Such an MDP is called a multi-armed bandit!

Markov Decision Problems

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

2. MDP planning

3. Alternative formulations

4. Applications

5. Policy Evaluation

Structure of State Values

Let us investigate state values. For $\pi \in \Pi, s \in S$:

$$V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s]$$

Structure of State Values

Let us investigate state values. 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

Let us investigate state values. For $\pi \in \Pi, s \in S$:

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

Structure of State Values

Let us investigate state values. For $\pi \in \Pi, s \in S$:

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

Structure of State Values

Let us investigate state values. For $\pi \in \Pi, s \in S$:

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

Bellman's Equations

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

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

Bellman's 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 $S = \{s_1, s_2, \ldots, s_n\}$.

Bellman's 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 $S = \{s_1, s_2, \ldots, s_n\}$.
- n equations, n unknowns—$V^{\pi}(s_1)$, $V^{\pi}(s_2)$, $\ldots$ $V^{\pi}(s_2)$.

Bellman's 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 $S = \{s_1, s_2, \ldots, s_n\}$.
- n equations, n unknowns—$V^{\pi}(s_1)$, $V^{\pi}(s_2), \ldots V^{\pi}(s_2)$.
- Linear!

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

Bellman's 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—$V^{\pi}(s_1)$, $V^{\pi}(s_2), \ldots V^{\pi}(s_2)$.
- Linear!
- Guaranteed to have a unique solution if $\gamma < 1$.
- If task is episodic, guaranteed to have a unique solution even if $\gamma = 1$, after we fix $V^{\pi}(s^{\top}) = 0$.

Bellman's 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 $S = \{s_1, s_2, \ldots, s_n\}$.
- n equations, n unknowns—$V^{\pi}(s_1), V^{\pi}(s_2), \ldots V^{\pi}(s_2)$.
- Linear!
- Guaranteed to have a unique solution if $\gamma < 1$.
- If task is episodic, guaranteed to have a unique solution even if $\gamma = 1$, after we fix $V^{\pi}(s^{\top}) = 0$.
- Policy evaluation: 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).

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

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

$Q^{\pi}(s, a)$ is the expected long-term reward from starting at s, taking a at $t = 0$, and following 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 s, taking a at $t = 0$, and following 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 $\pi \in \Pi, s \in S, a \in A$:

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

$Q^{\pi}(s,a)$ is the expected long-term reward from starting at s, taking a at $t = 0$, and following 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 s, taking a at $t = 0$, and following 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 s, taking a at $t = 0$, and following 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 $\pi \in \Pi, s \in S, a \in A$:

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

  $Q^{\pi}(s,a)$ is the expected long-term reward from starting at s, taking a at $t = 0$, and following 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 action value function $Q^{\star}$.

Action Value Function

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

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

  $Q^{\pi}(s,a)$ is the expected long-term reward from starting at s, taking a at $t = 0$, and following 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 action value function $Q^{\star}$.
- We will find use for Q^{π} and $Q^{\star}$ next week.

Markov Decision Problems

1. Definitions
   - Markov Decision Problem
   - Policy
   - Value Function
2. MDP planning
3. Alternative formulations
4. Applications
5. Policy Evaluation