Markov Decision Processes

Prof. Junni Zou

Institute of Media, Information and Network
Dept. of Computer Science and Engineering
Shanghai Jiao Tong University
http://min.sjtu.edu.cn

Spring, 2021

Outline

1 Markov Processes

2 Markov Reward Processes

3 Markov Decision Processes

4 Optimal Value Function

Table of Contents

Institute of Media,
Information, and Network

Markov Property

"The future is independent of the past given the present"

Definition

A state S_t is *Markov* if and only of

$$\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_1, \cdots, S_t]$$

- The state captures all relevant information from the history
- Once the state is known, the history may be thrown away
- i.e. The state is a sufficient statistic of the future

Institute of Media,
Information, and Network

State Transition Matrix

For a Markov state s and successor state s', the *state transition probability* is defined by

$$\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s]$$

State transition matrix $\mathcal{P}$ defines transition probabilities from all states s to all successor states s',

$$\mathcal{P} = \begin{bmatrix} \mathcal{P}_{11} & \cdots & \mathcal{P}_{1n} \\ \vdots & & \vdots \\ \mathcal{P}_{n1} & \cdots & \mathcal{P}_{nn} \end{bmatrix}$$

where each row of the matrix sums to 1.

Institute of Media,
Information, and Network

Markov Process

A Markov process is a memoryless random process, i.e. a sequence of random states $S_1, S_2, \cdots$ with the Markov property.

Definition

A *Markov Process (or Markov Chain)* is a tuple $\langle \mathcal{S}, \mathcal{P} \rangle$

- $\mathcal{S}$ is a (finite) set of states
- $\mathcal{P}$ is a state transition probability matrix,

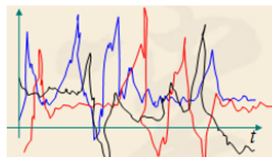$$\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s]$$

Institute of Media,
Information, and Network

Random Process

随机过程的概念

📖随机变量------在每次试验结果中，以一定的概率取
某个事先未知，但为确定的数值。

📖随机信号(过程)------ 随时间参数 **t** 变化的随机变量。
记为，**X(t)**。

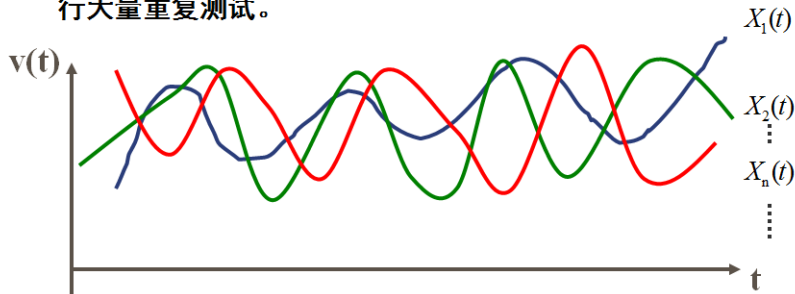**随机信号是与确定信号相对应的。**

Random Process

例：在相同条件下，对同一雷达接收机的噪声电压(电流)进行大量重复测试。
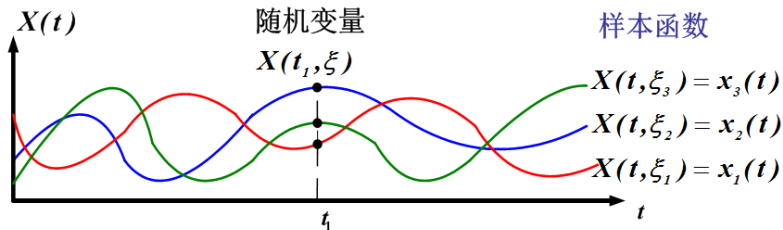

1. 具体的波形事先不能预知；

2. 但必为所有可能波形之一；

3. 所有可能波形 $x_1(t), x_2(t)..., x_n(t),...$ 的集合(总体)构成了随机过程 $X(t)$。

Random Process

1. $x_i(t)$是随机试验第i次的实验结果ξ_i，即随机过程的第i次实现，也称为随机过程的样本(函数)；一次试验结果，随机过程必取一个样本函数 ξ_i；

2. 所有$\{x_i(t) = \xi_i\}$的集合构成随机过程。随机过程既是时间t，又是可能试验结果ξ的函数，记为$X(t,\xi)$；

3. $X(t_1,\xi)$ 表示随机过程 $X(t,\xi)$ 在t_1时刻的各种可能结果的取值，称为随机过程 $X(t,\xi)$在t_1时刻的随机变量。
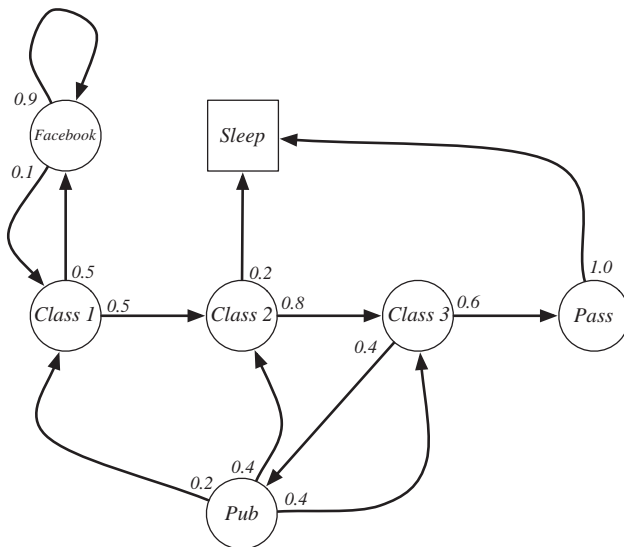
Random Process

随机过程的严格定义：

📖 **定义1：** 设随机试验E的样本空间为 $S=\{\xi\}$，若对于每个元素 $\xi \in \{S\}$，总有一个确定的时间函数 $X(t, \xi), t \in T$ 与之对应，则对于所有的 $\xi \in \{S\}$ 得到**一族时间 t 的函数**，称为随机过程。族中的每一个函数称为该随机过程的样本函数。

📖 **定义2：** 对于每个特定的时刻ti, $X(t_i, \xi)$ 都是一个随机变量，依赖于时间t的一族随机变量 $X(t_1, \xi), X(t_2, \xi), ..., X(t_n, \xi)$ 就组成了随机过程 $X(t, \xi)$。

根据定义，列出随机过程在四种情况下的含义：

- t, ξ 皆为变量时，$X(t, \xi)$ 为时间函数族 $X(t)$
- t 为变量, ξ 确定时，$X(t, \xi)$ 为确定的时间函数 $x(t)$
- ξ 为变量, t 确定时，$X(t, \xi)$ 为一随机变量 $X(t_i)$
- ξ, t 皆确定时，$X(t, \xi)$ 为一确定值 $x(t_i)$
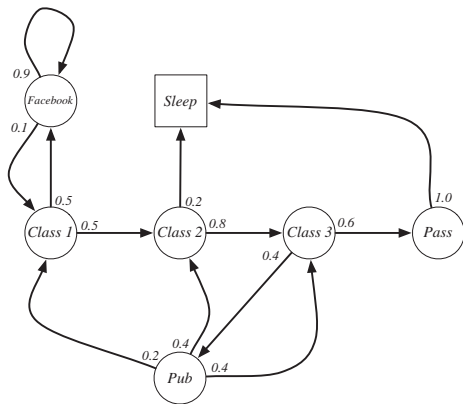
tute of Media,
rmation, and Network

Example: Student Markov Chain
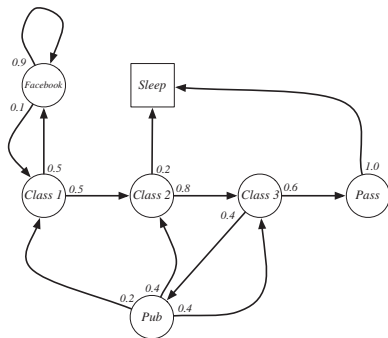
Example: Student Markov Chain Episodes

Sample episodes for Student Markov Chain starting from $S_1 = C_1$

$$S_1, S_2, \cdots, S_T$$

- C1 C2 C3 Pass Sleep
- C1 FB FB C1 C2 Sleep
- C1 C2 C3 Pub C2 C3 Pass Sleep
- C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep

Example: Student Markov Chain Transition Matrix

$$\mathcal{P} = \begin{array}{c} \\ C1 \\ C2 \\ C3 \\ Pass \\ Pub \\ FB \\ Sleep \end{array} \begin{array}{ccccccc} C1 & C2 & C3 & Pass & Pub & FB & Sleep \\ & 0.5 & & & & 0.5 & \\ & & 0.8 & & & & 0.2 \\ & & & 0.6 & 0.4 & & \\ & & & & & & 1.0 \\ 0.2 & 0.4 & 0.4 & & & & \\ 0.1 & & & & & 0.9 & \\ & & & & & & 1 \end{array}$$

Table of Contents

Institute of Media,
Information, and Network

Goal and Rewards
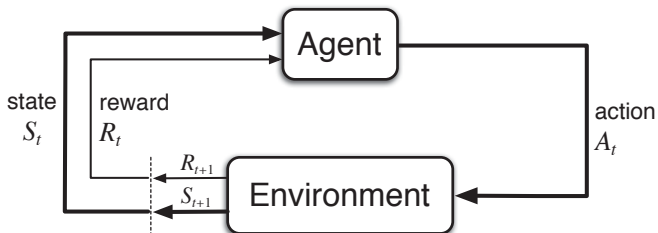
- Informally, the agent's goal is to *maximize the total amount of reward it received*.
- This means maximizing not immediate reward, but cumulative reward in the long run.
- The use of a reward signal to formalize the idea of a goal is one of the most distinctive features of reinforcement learning.

Markov Reward Process

A Markov reward process is a Markov chain with values.

Definition

A *Markov Reward Process* is a tuple $\langle \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma \rangle$
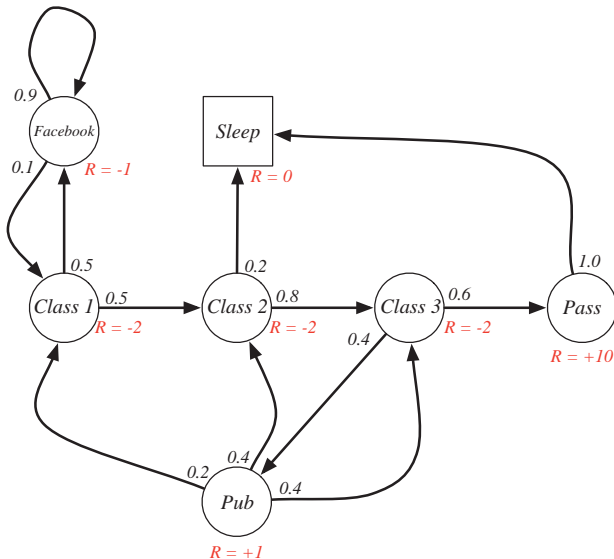
- $\mathcal{S}$ is a finite set of states
- $\mathcal{P}$ is a state transition probability matrix,

$$\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s]$$

- $\mathcal{R}$ is a reward function, $\mathcal{R}_s = \mathbb{E}[R_{t+1} | S_t = s]$
- γ is a discount factor, $\gamma \in [0, 1]$

Institute of Media,
Information, and Network

Example: Student MRP

Return

Definition

The *return* G_t is the total discounted reward from time-step t

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

- The discount $\gamma \in [0, 1]$ is the present value of future rewards
- The value of receiving reward R after $k + 1$ time-steps is $\gamma^k R$
- This values immediate reward above delayed reward.
  - γ close to 0 leads to "myopic" evaluation
  - γ close to 1 leads to "far-sighted" evaluation

Institute of Media,
Information, and Network

Why discount?

Most Markov reward and decision processes are discounted. Why?

- Mathematically convenient to discount rewards
- Avoids infinite returns in cyclic Markov processes
  - If the task is continuous, the final time step would be $T = \infty$
- Uncertainty about the future may not be fully represented
- If the reward is financial, immediate rewards may earn more interest than delayed rewards
- Animal/human behaviour shows preference for immediate reward
- It is sometimes possible to use *undiscounted* Markov reward processes (i.e. $\gamma = 1$), e.g. if all sequences terminate, which we call *episodic task*.

Institute of Media,
Information, and Network

Example: Student MRP Returns

Sample returns for Student MRP: Starting from $S_1 = C1$ with $\gamma = \frac{1}{2}$

$$G_1 = R_2 + \gamma R_3 + \cdots + \gamma^{T-2} R_T$$

C1 C2 C3 Pass Sleep	$v_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4} + 10 * \frac{1}{8}$	$= \quad -2.25$
C1 FB FB C1 C2 Sleep	$v_1 = -2 - 1 * \frac{1}{2} - 1 * \frac{1}{4} - 2 * \frac{1}{8} - 2 * \frac{1}{16}$	$= \quad -3.125$
C1 C2 C3 Pub C2 C3 Pass Sleep	$v_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4} + 1 * \frac{1}{8} - 2 * \frac{1}{16} \cdots$	$= \quad -3.41$
C1 FB FB C1 C2 C3 Pub C1 ...	$v_1 = -2 - 1 * \frac{1}{2} - 1 * \frac{1}{4} - 2 * \frac{1}{8} - 2 * \frac{1}{16} \cdots$	$= \quad -3.20$
FB FB FB C1 C2 C3 Pub C2 Sleep		

Institute of Media,
Information, and Network
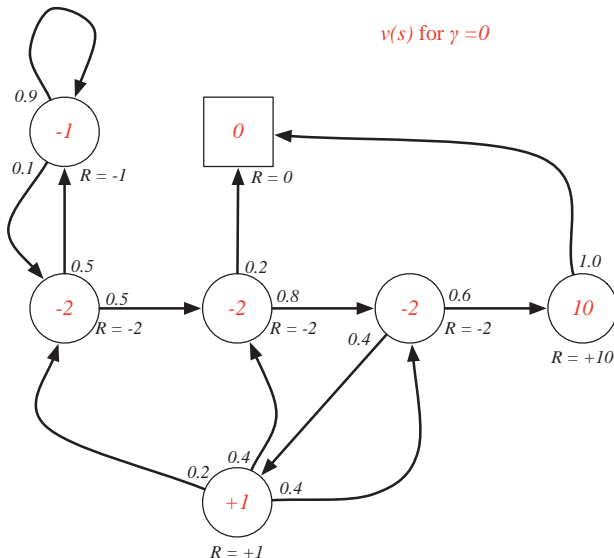
Value Function

The value function $v(s)$ gives the long-term value of state s, i.e. estimate *how good* it is for the agent to be in a given state
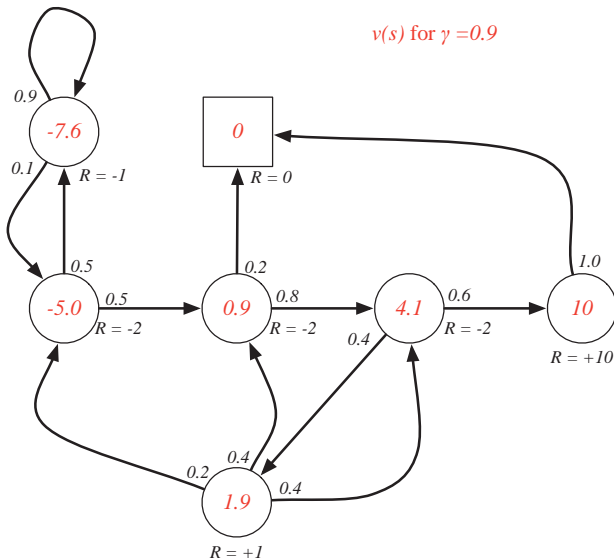
Definition

The *state value function* $v(s)$ of an MRP is the expected return starting from state s
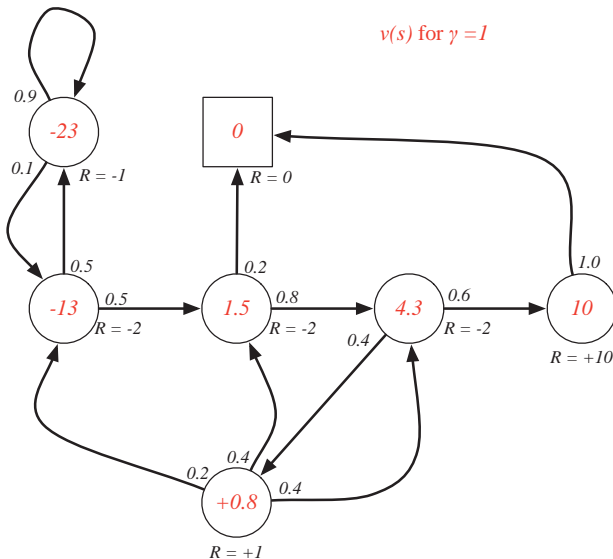
$$v(s) = \mathbb{E}[G_t | S_t = s]$$

Institute of Media,
Information, and Network

Example: State-Value Function for Student MRP (1)

$v(s)$ for $\gamma = 0$

Institute of Media,
Information, and Network

Example: State-Value Function for Student MRP (2)

$v(s)$ for $\gamma = 0.9$

Example: State-Value Function for Student MRP (3)

$v(s)$ for $\gamma = 1$
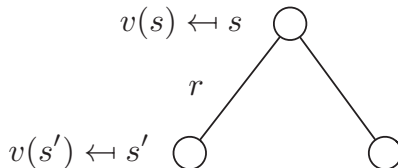
Bellman Equation for MRPs

The value function can be decomposed into two parts:

- immediate reward R_{t+1}
- discounted value of successor state $\gamma v(S_{t+1})$

$$
\begin{aligned}
v(s) &= \mathbb{E}[G_t | S_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots | S_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \cdots) | S_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma G_{t+1} | S_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma v(S_{t+1}) | S_t = s]
\end{aligned}
$$

Institute of Media,
Information, and Network

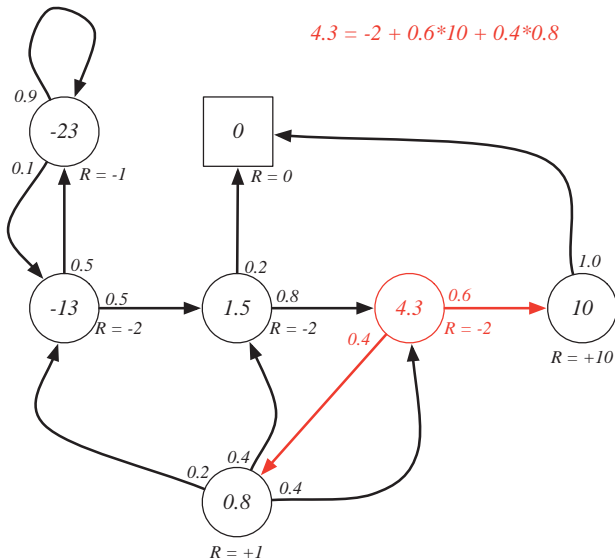# Bellman Equation for MRPs (2)

$$v(s) = \mathbb{E}[R_{t+1} + \gamma v(S_{t+1})|S_t = s]$$

$$v(s) = \mathcal{R}_s + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'} v(s')$$

Example: Bellman Equation for Student MRP

4.3 = -2 + 0.6*10 + 0.4*0.8

Bellman Equation in Matrix Form

The Bellman equation can be expressed concisely using matrices,

$$v = \mathcal{R} + \gamma \mathcal{P} v$$

where v is a column vector with one entry per state

$$\begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix} = \begin{bmatrix} \mathcal{R}_1 \\ \vdots \\ \mathcal{R}_n \end{bmatrix} + \gamma \begin{bmatrix} \mathcal{P}_{11} & \cdots & \mathcal{P}_{1n} \\ \vdots & & \vdots \\ \mathcal{P}_{11} & \cdots & \mathcal{P}_{nn} \end{bmatrix} \begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix}$$

Institute of Media,
Information, and Network

Solving the Bellman Equation

- The Bellman equation is a linear equation
- It can be solved directly:

$$v = \mathcal{R} + \gamma \mathcal{P} v$$
$$(I - \gamma \mathcal{P}) v = \mathcal{R}$$
$$v = (I - \gamma \mathcal{P})^{-1} \mathcal{R}$$

- Computational complexity is $O(n^3)$ for n states
- Direct solution only possible for small MRPs
- There are many iterative methods for large MRPs, e.g.
  - Dynamic programming
  - Monte-Carlo evaluation
  - Temporal-Difference learning

Institute of Media,
Information, and Network

Table of Contents

Institute of Media,
Information, and Network

Markov Decison Process

A Markov decision process (MDP) is a Markov reward process with decisions. It is an *environment* in which all states are Markov.

Definition

A *Markov Decision Process* is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$

- $\mathcal{S}$ is a finite set of states
- $\mathcal{A}$ is a finite set of actions
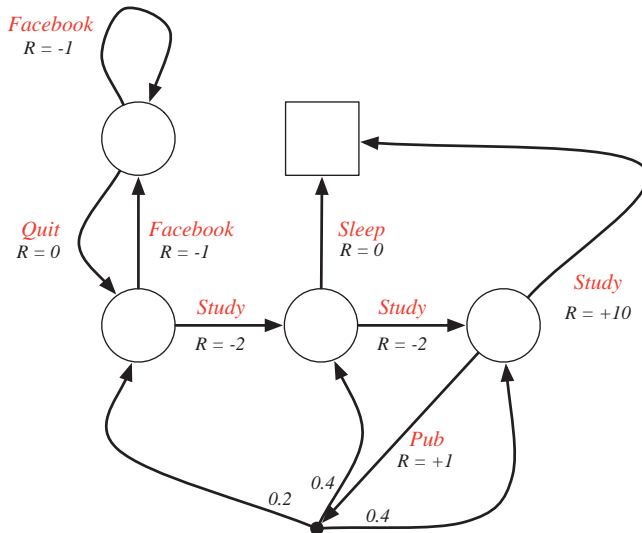- $\mathcal{P}$ is a state transition probability matrix,

$$\mathcal{P}_{ss'}^{a} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$$

- $\mathcal{R}$ is a reward function,

$$\mathcal{R}_{s}^{a} = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$$

- γ is a discount factor $\gamma \in [0, 1]$

Example: Student MDP

Policies (1)

Definition

A *policy* π is a distribution over actions given states,

$$\pi(a|s) = \mathbb{P}[A_t = a | S_t = s]$$

- A policy fully defined the behaviour of an agent
- MDP policies depend on the current state (not the history)
- i.e. Policies are stationary (time-independent),
  $A_t \sim \pi(\cdot|S_t), \forall t > 0$

Institute of Media,
Information, and Network

Policies (2)

- Given an MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ and a policy π
- The state sequence $S_1, S_2, \cdots$ is a Markov process $\langle \mathcal{S}, \mathcal{P}^\pi \rangle$
- The state and reward sequence $S_1, R_2, S_2, \cdots$ is a Markov reward process $\langle \mathcal{S}, \mathcal{P}^\pi, \mathcal{R}^\pi, \gamma \rangle$
- where

$$\mathcal{P}^\pi_{ss'} = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{P}^a_{ss'}$$

$$\mathcal{R}^\pi_s = \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{R}^a_s$$

Institute of Media,
Information, and Network

Value Function Following Policy

Definition

The state-value function $v_\pi(s)$ of an MDP is the expected return starting from state s, and then *following policy π*
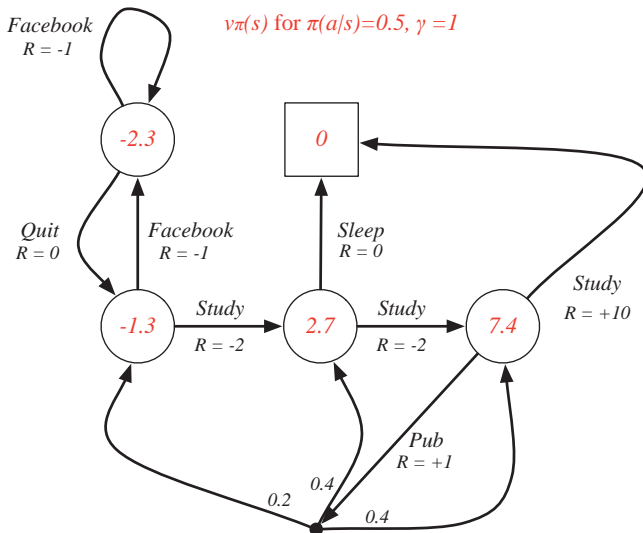
$$v_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s]$$

Definition

The action-value function $q_\pi(s, a)$ is the expected return starting from state s, taking action a, and then *following policy π*

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$$

Institute of Media,
Information, and Network

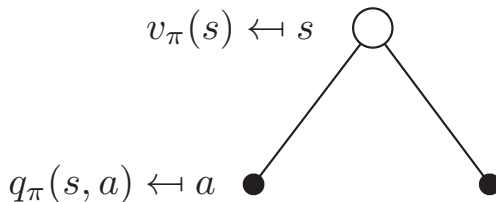# Example: State-Value Function for Student MDP

Bellman Expectation Equation

The state-value function can again be decomposed into immediate reward plus discounted value of successor state,

$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s]$$
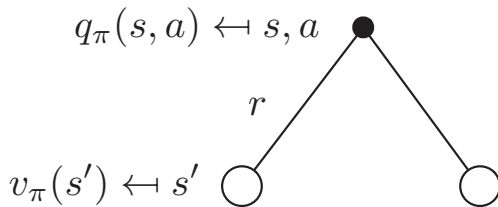
The action-value function can similarly be decomposed,

$$q_\pi(s, a) = \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a]$$

Institute of Media,
Information, and Network

Bellman Expectation Equation for V^{π}

$$v_{\pi}(s) \longleftarrow s \quad \bigcirc$$

$$q_{\pi}(s, a) \longleftarrow a \quad \bullet \qquad\qquad \bullet$$

$$v_{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_{\pi}(s, a)$$

Institute of Media,
Information, and Network

Bellman Expectation Equation for Q^π

$$q_\pi(s,a) \hookleftarrow s, a$$

$$r$$

$$v_\pi(s') \hookleftarrow s'$$

$$q_\pi(s, a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_\pi(s')$$

Institute of Media,
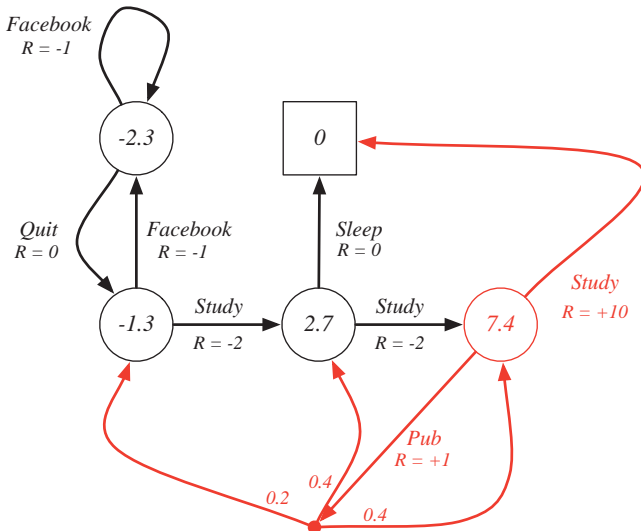Information, and Network

Bellman Expectation Equation for v_π (2)

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s)\left(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_\pi(s')\right)$$

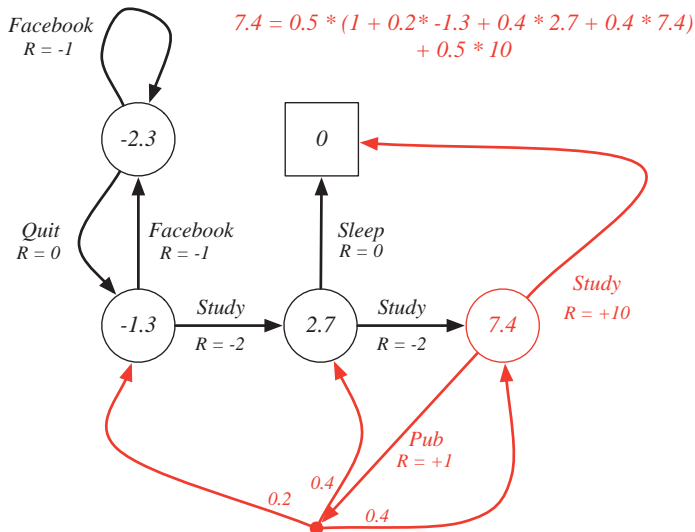# Bellman Expectation Equation for q_π (2)
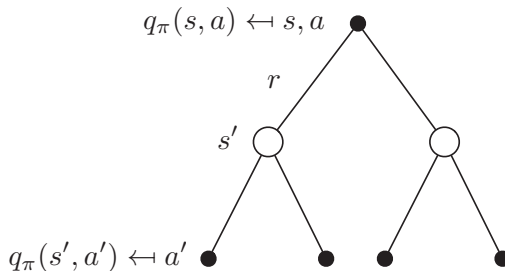
$$q_\pi(s, a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a \sum_{a' \in \mathcal{A}} \pi(a'|s') q_\pi(s', a')$$

Example: Bellman Expectation Equation in Student MDP

Example: Bellman Expectation Equation in Student MDP

$$7.4 = 0.5 * (1 + 0.2 * -1.3 + 0.4 * 2.7 + 0.4 * 7.4) + 0.5 * 10$$

Bellman Expectation Equation (Matrix Form)

The Bellman Expectation equation can be expressed concisely using the induced MRP,

$$v_\pi = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi v_\pi$$

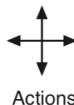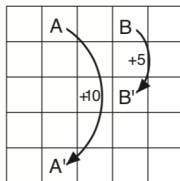with direct solution

$$v_\pi = (I - \gamma \mathcal{P}^\pi)^{-1} \mathcal{R}^\pi$$

Institute of Media,
Information, and Network

Example: Bellman Expectation Equation in Gridworld

Example 3.5: Gridworld Figure 3.2 (left) shows a rectangular gridworld representation of a simple finite MDP. The cells of the grid correspond to the states of the environment. At each cell, four actions are possible: `north`, `south`, `east`, and `west`, which deterministically cause the agent to move one cell in the respective direction on the grid. Actions that would take the agent off the grid leave its location unchanged, but also result in a reward of -1. Other actions result in a reward of 0, except those that move the agent out of the special states A and B. From state A, all four actions yield a reward of $+10$ and take the agent to A$'$. From state B, all actions yield a reward of $+5$ and take the agent to B$'$.

3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

Actions

Table of Contents

Institute of Media,
Information, and Network

Optimal Value Function

Definition

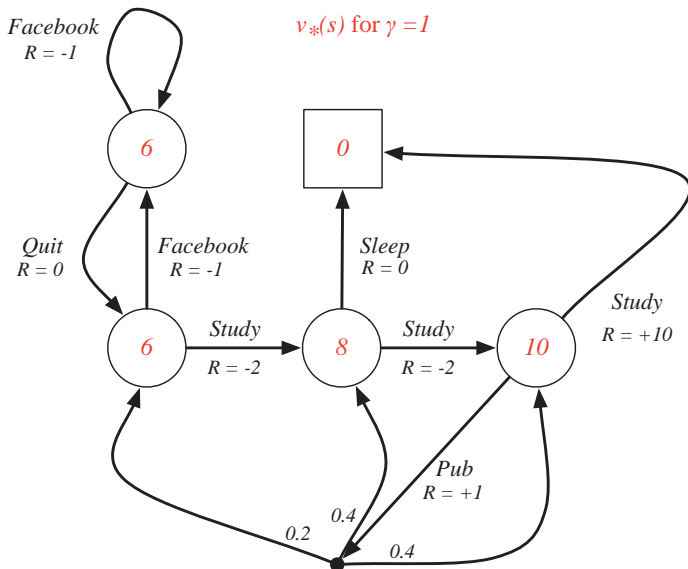The *optimal state-value function* $v_*(s)$ is the maximum value function over all policies

$$v_*(s) = \max_\pi v_\pi(s)$$

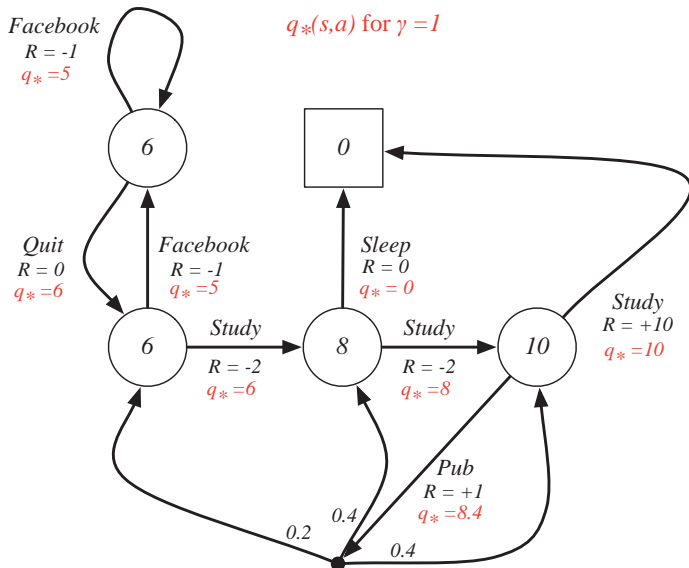The *optimal action-value function* $q_*(s, a)$ is the maximum action-value function over all policies

$$q_*(s, a) = \max_\pi q_\pi(s, a)$$

- The optimal value function specifies the best possible performance in the MDP.
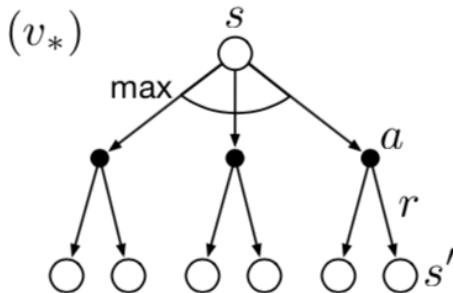- An MDP is "solved" when we know the optimal value.

Institute of Media, Information, and Network

Example: Optimal Value Function for Student MDP

$v_*(s)$ for $\gamma = 1$

Example: Optimal Action-Value Function for Student MDP

$q_*(s,a)$ for $\gamma = 1$

Facebook
R = -1
$q_* = 5$

Quit
R = 0
$q_* = 6$

Facebook
R = -1
$q_* = 5$

Sleep
R = 0
$q_* = 0$

Study
R = +10
$q_* = 10$

Study
R = -2
$q_* = 6$

Study
R = -2
$q_* = 8$

Pub
R = +1
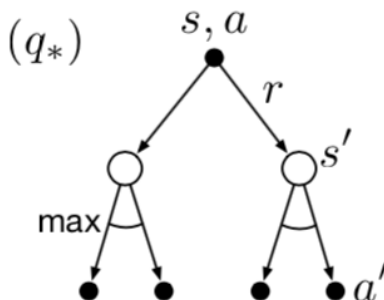$q_* = 8.4$

0.2

0.4

0.4

tute of Media,
rmation, and Network

Bellman Optimality Equation for V^*

$$v_*(s) = \max_a \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_*(s')$$

Institute of Media,
Information, and Network

Bellman Optimality Equation for Q^*

$$q_*(s, a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a \max_{a'} q_*(s', a')$$

Example: Bellman Optimal Equation in Student MDP

$6 = max \{-2 + 8, \ -1 + 6\}$
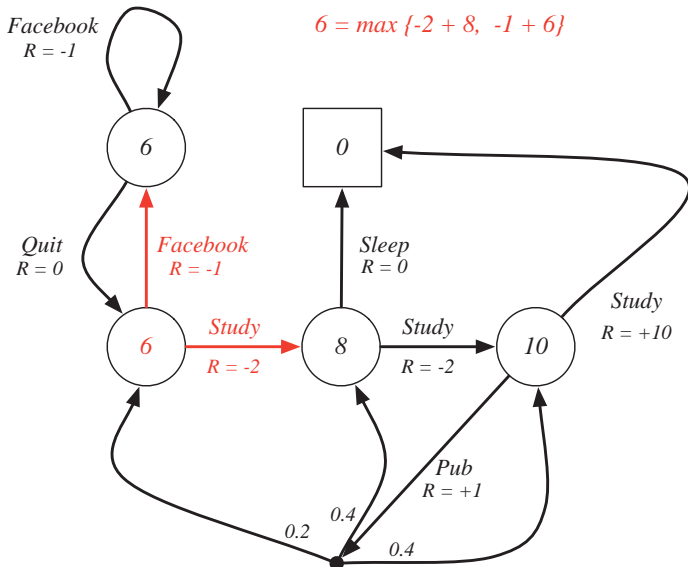
Optimal Policy

Define a partial ordering over policies

$$\pi \geq \pi' \text{ if } v_\pi(s) \geq v_{\pi'}(s), \forall s$$

Theorem

For any Markov Decision Process

- *There exists an optimal policy π_* that is better than or equal to all other policies, $\pi_* \geq \pi, \forall \pi$*
- *All optimal policies achieve the optimal value function, $v_{\pi_*}(s) = v_*(s)$*
- *All optimal policies achieve the optimal action-value function, $q_{\pi_*}(s, a) = q_*(s, a)$*
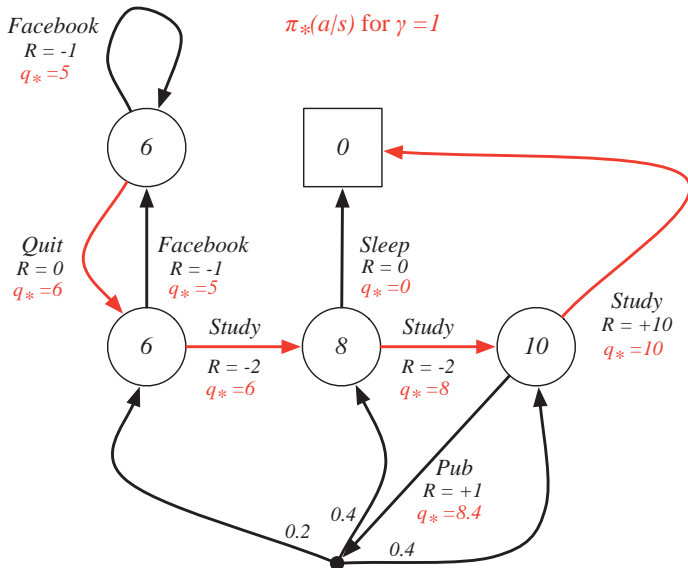
Finding an Optimal Policy

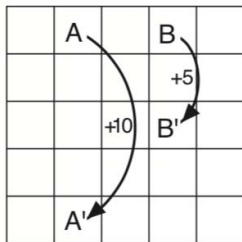An optimal policy can be found by maximizing over $q_*(s, a)$,

$$\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_{a \in \mathcal{A}} q_*(s, a) \\ 0 & \text{otherwise} \end{cases}$$

- There is always a deterministic optimal policy for any MDP
- If we know $q_*(s, a)$, we immediately have the optimal policy

Institute of Media,
Information, and Network

Example: Optimal Policy for Student MDP

Example: Optimal Policy for Gridworld

Gridworld v_* π_*
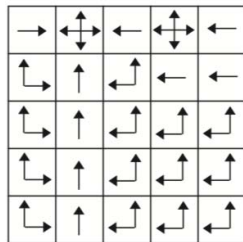
Solving the Bellman Optimality Equation

- Bellman Optimality Equation is non-linear
- No closed form solution (in general)
- Many iterative solution methods
  - Value Iteration
  - Policy Iteration
  - Q-learning
  - Sarsa