Lecture 8: Markov Decision Process

Ziyu Shao

School of Information Science and Technology
ShanghaiTech University

April 27 & 29, 2020

Outline

Outline

Markov Decision Process

- Markov decision processes formally describe an environment for reinforcement learning
- Where the environment is fully observable
- i.e. The current state completely characterizes the process
- Almost all RL problems can be formalized as MDPs, e.g
  - Optimal control primarily deals with continuous MDPs
  - Partially observable problems can be converted into MDPs
  - Bandits are MDPs with one state

POMDP

Markov Property

Definition

A state S_t is Markovian if and only if

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

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

Markov Chain

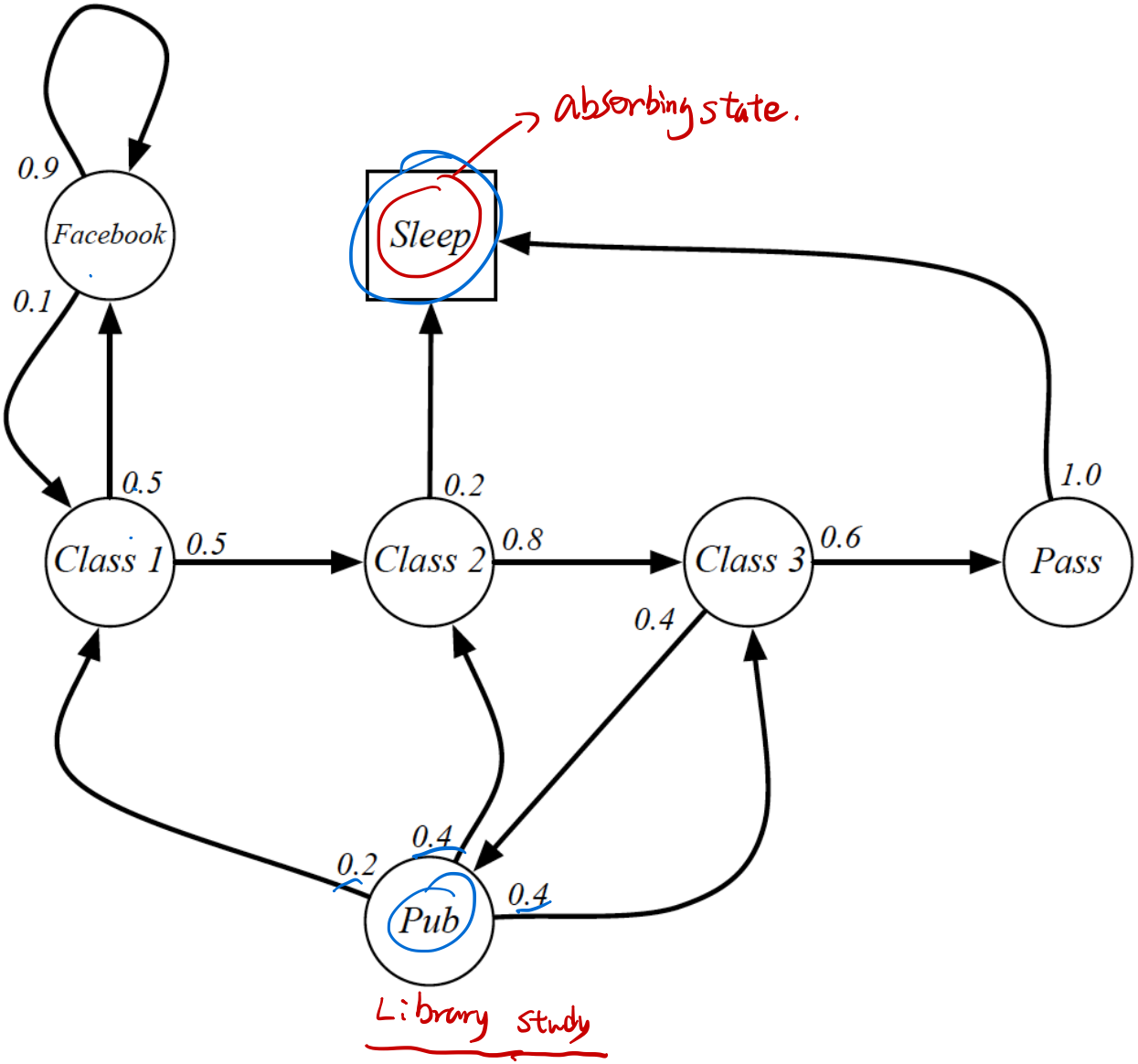
Definition

A discrete-time Markov chain is a tuple $< \mathcal{S}, \mathcal{P} >$

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

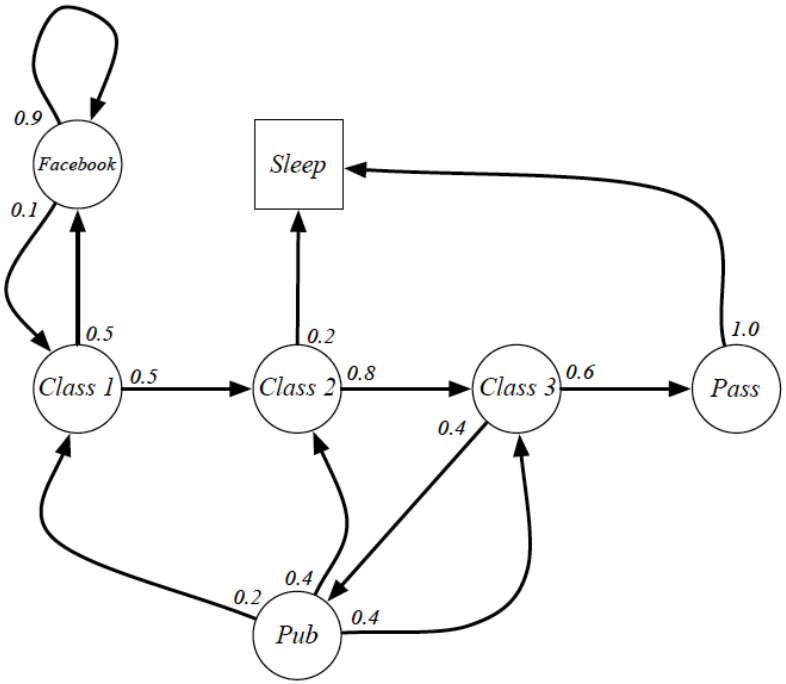
$$\mathcal{P}_{s,s'} = \mathbb{P}[S_{t+1} = s' | S_t = s]$$

Example: Student Markov Chain



Example: Student Markov Chain Episodes

Sample path

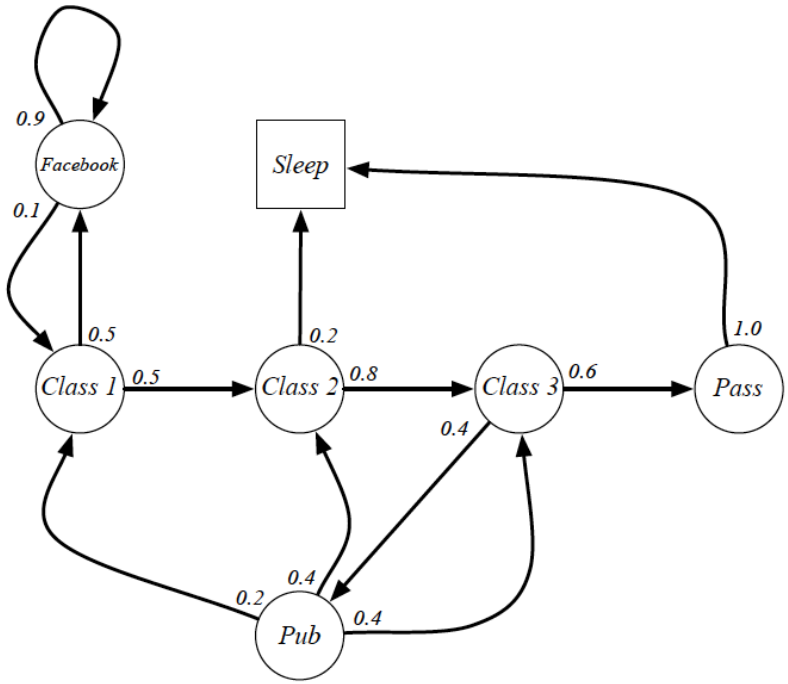


Sample episodes for Student Markov Chain starting from $S_1 = C1$

$$S_1, S_2, ..., S_T$$

- 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



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

Sum of each Row = 1

Stochastic matrix.

Outline

Markov Reward Process

A Markov reward process is a Markov chain with values.

Definition

A Markov Reward Process is a tuple $< \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma >$

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

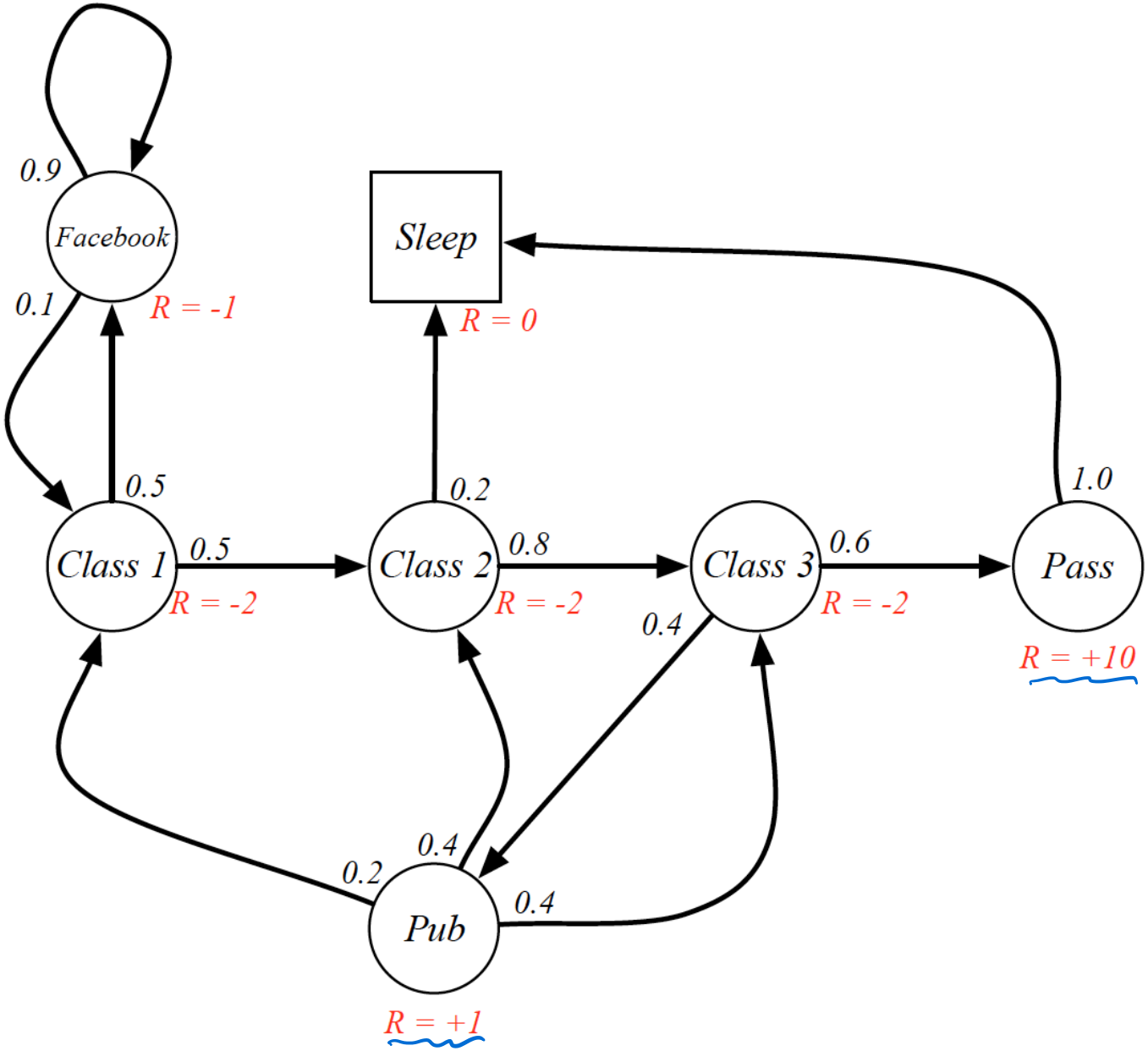
$$\mathcal{P}_{s,s'} = \mathbb{P}[S_{t+1} = s' | S_t = s]$$

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

$$\mathcal{R}_s = \mathbb{E}[R_{t+1} | S_t = s]$$

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

Example: Student MRP



Return

Definition

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

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

- 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

Why Discount?

Most Markov reward and decision processes are discounted. Why?

- Mathematically convenient to discount rewards
- Avoids infinite returns in cyclic Markov processes
- 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 behavior 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.

Value Function

The value function $v(s)$ gives the long-term value of state s

Definition

The state value function $v(s)$ of an MRP is the expected return starting from state s

$$v(s) = \mathbb{E}[G_t \mid S_t = s]$$

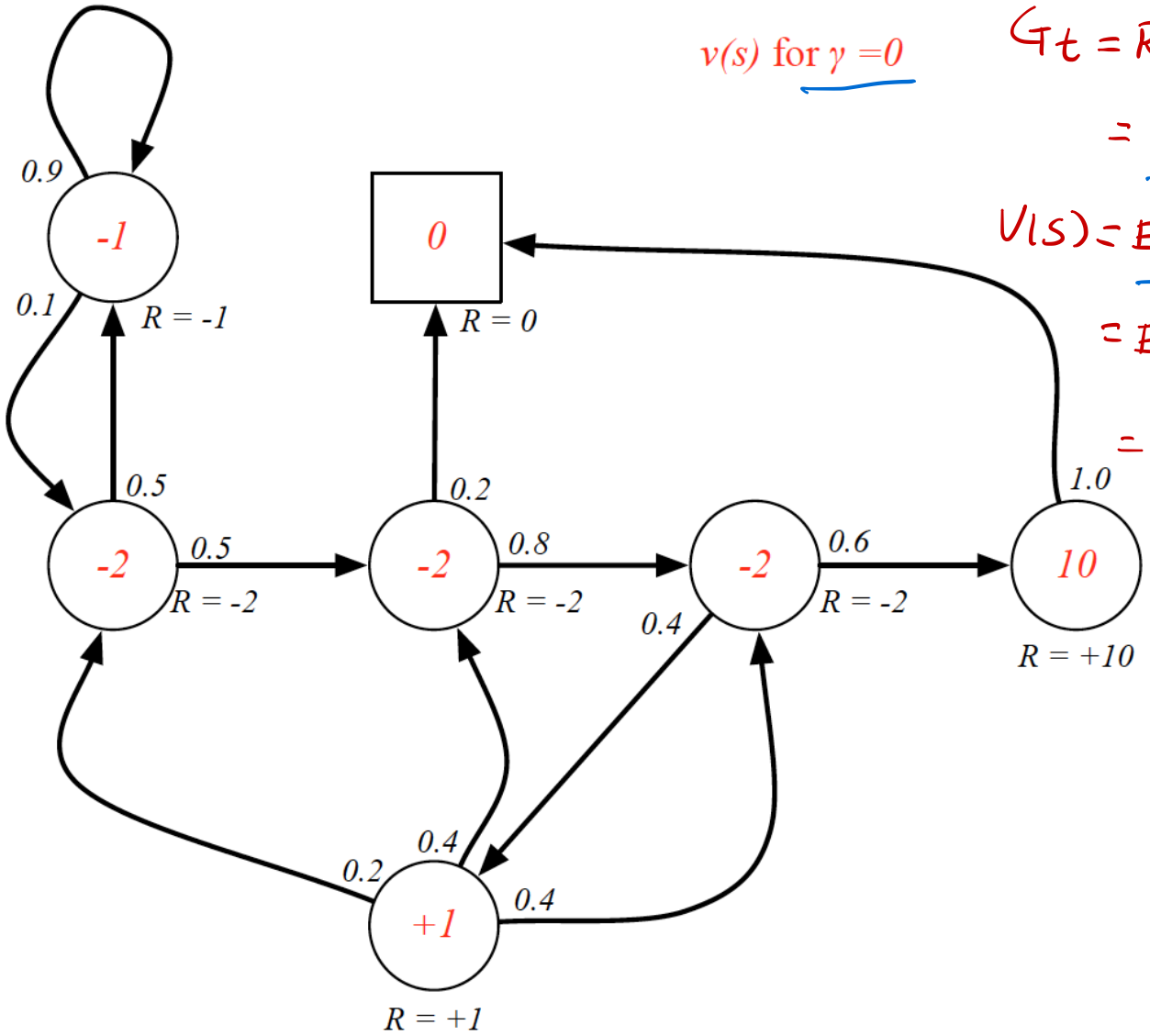
Example: Student MRP Returns

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

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

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

Example: State-Value Function for Student MRP



$G_t = R_{t+1} + \gamma R_{t+2} + \cdots$

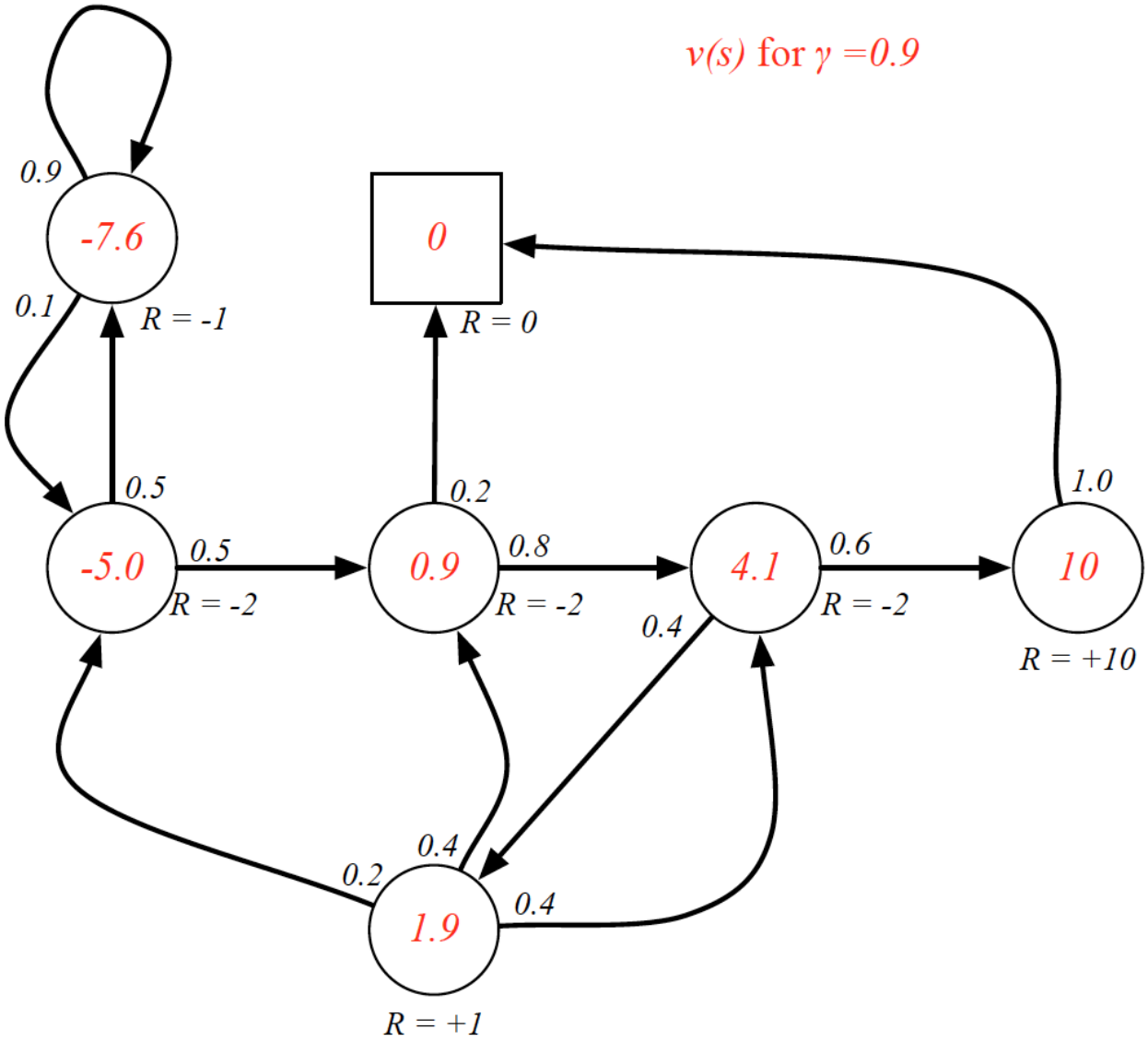
$= R_{t+1}$

$V(s) = E[G_t | S_t = s]$

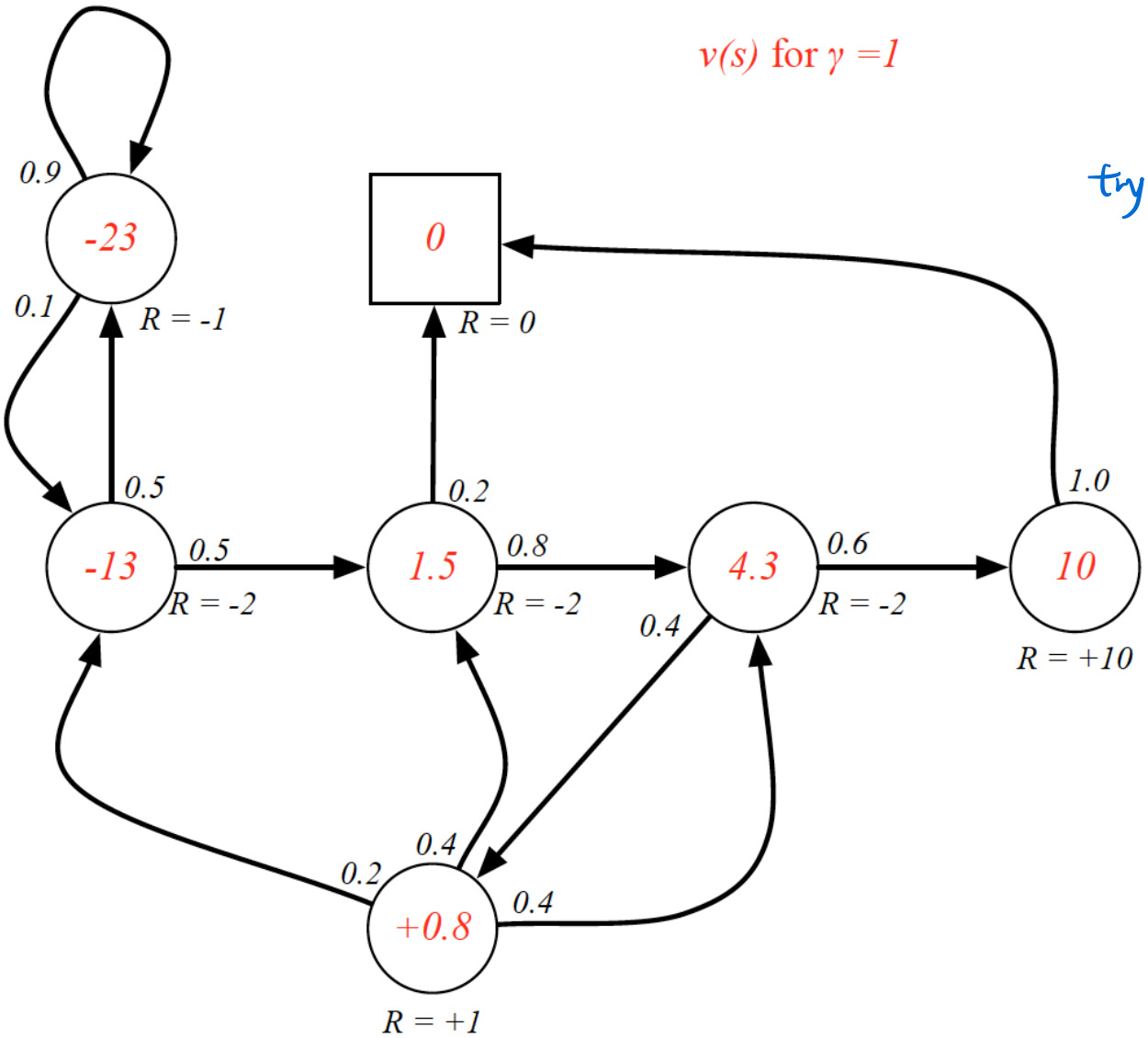
$= E[R_{t+1} | S_t = s]$

$= R_s$

Example: State-Value Function for Student MRP



Example: State-Value Function for Student MRP



Bellman Equation for MRPs

The value function can be decomposed into two parts:

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

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \dots) = R_{t+1} + \gamma G_{t+1}$$

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

$$V(S_t) = \mathbb{E}[R_{t+1} + \gamma V(S_{t+1}) \mid S_t]$$

Bellman Equation for MRPs

Adam's theorem + Markov.

1°. $V(s) = E[G_t | S_t = s]$, $V(S_t) = E[G_t | S_t]$, $V(S_{t+1}) = E[G_{t+1} | S_{t+1}]$

2°. Adam's Law. $E[E(Y|X)] = E(Y)$

Adam's Law with extra conditioning. $\hat{E}(\cdot) = E(\cdot|Z)$ $\Rightarrow$ $\hat{E}[\hat{E}(Y|X)] = \hat{E}(Y)$

$\Leftrightarrow$ $E[E(Y|X,Z)|Z] = E(Y|Z)$

Let $Y = G_{t+1}$, $X = S_{t+1}$, $Z = S_t$;

Then we have $E[E(G_{t+1}|S_{t+1}, S_t)|S_t] = E(G_{t+1}|S_t)$.

Markov property

$E[E[G_{t+1}|S_{t+1}]|S_t] = E[V(S_{t+1})|S_t]$

thus $E[G_{t+1}|S_t] = E[V(S_{t+1})|S_t]$

$\Rightarrow$ $E[G_{t+1}|S_t = s] = E[V(S_{t+1})|S_t = s]$.

3°. $V(s) = E[G_t|S_t = s] = E[R_{t+1} + \gamma G_{t+1}|S_t = s] = E[R_{t+1}|S_t = s] + \gamma E[G_{t+1}|S_t = s]$

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

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

Q.E.D.

Bellman Equation for MRPs

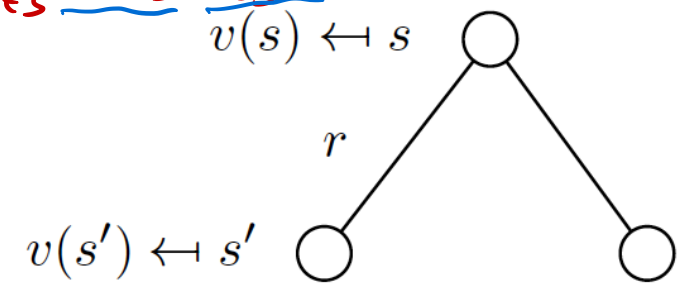
Bellman Equation for MRPs

$$v(s) = E[R_{t+1} \mid S_t = s] + \gamma E[v(S_{t+1}) \mid S_t = s] \quad \text{LOTE.}$$

$$= R_s + \gamma \sum_{s' \in S} E[v(S_{t+1}) \mid S_t = s, S_{t+1} = s'] \cdot P(S_{t+1} = s' \mid S_t = s)$$

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

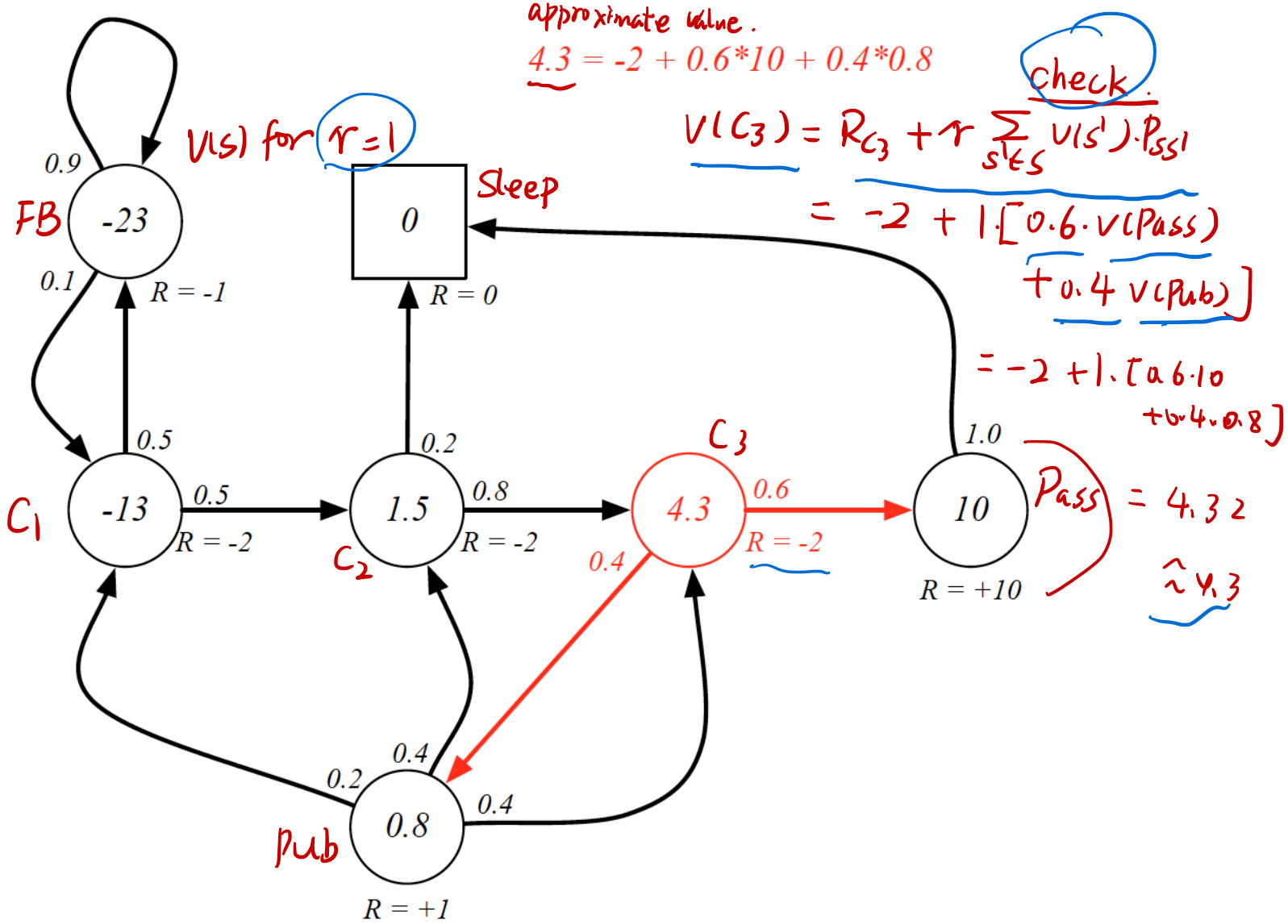
$$= R_s + \gamma \sum_{s' \in S} v(s') \cdot P_{ss'}$$



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

$$V(S_t) = E[R_{t+1} + \gamma V(S_{t+1}) \mid S_t]$$

Example: Bellman Equation for Student MRP



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

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

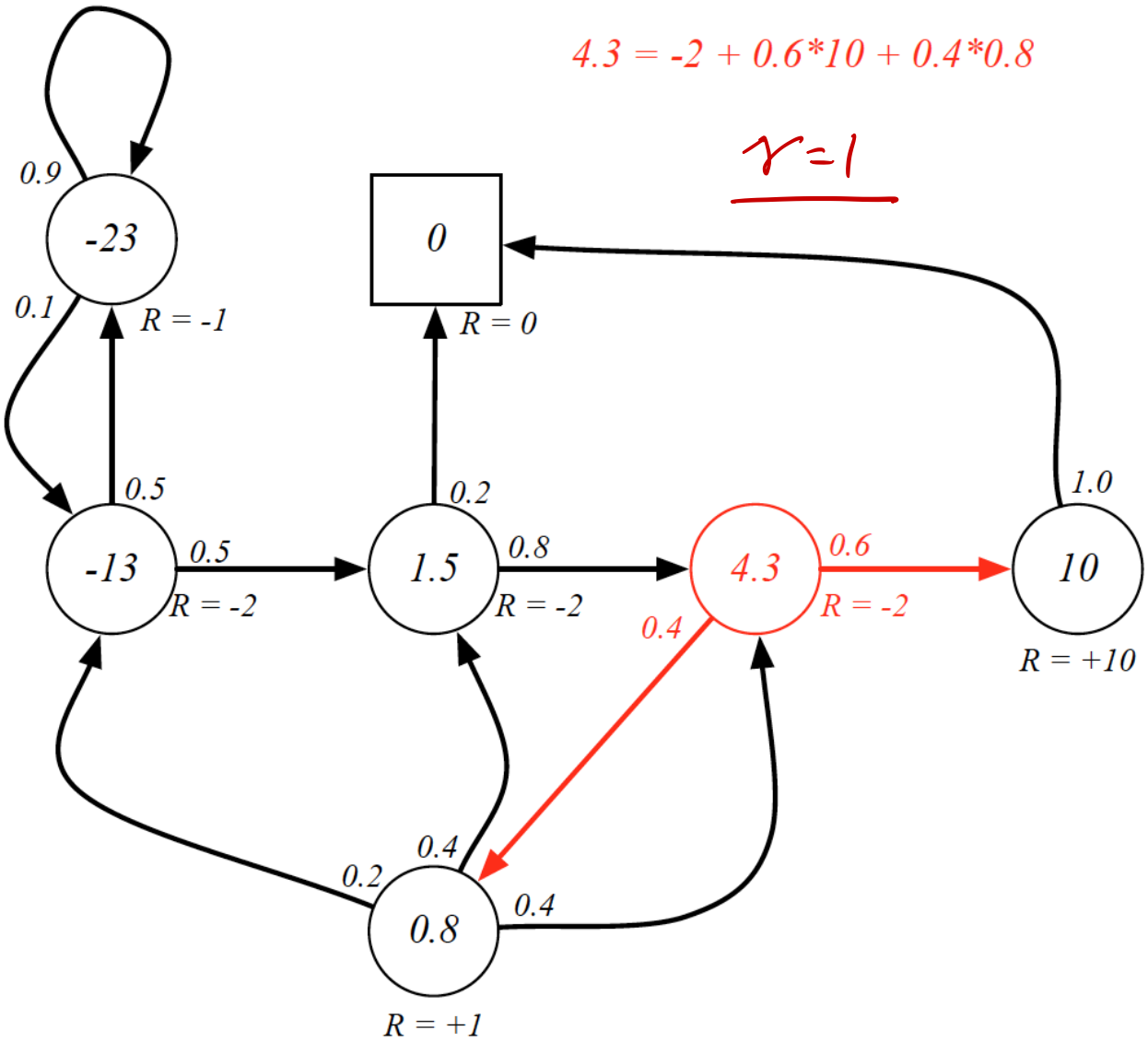
Solving the Bellman Equation

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

$$
\begin{aligned}
v &= \mathcal{R} + \gamma \mathcal{P} v \\
(I - \gamma \mathcal{P})\, v &= \mathcal{R} \\
v &= (I - \gamma \mathcal{P})^{-1} \mathcal{R}
\end{aligned}
$$

- 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

Example: Matrix Solution for Student MRP



Example: Matrix Solution for Student MRP

1°. state transition probability matrix P (Refer to page 9)

2°. $R = \begin{bmatrix} -2 \\ -2 \\ -2 \\ 10 \\ 1 \\ -1 \\ 0 \end{bmatrix}$ $\begin{matrix} C_1 \\ C_2 \\ C_3 \\ \text{Pass} \\ \text{Pub} \\ \text{FB} \\ \text{Sleep} \end{matrix}$

3°. $V = (I - \gamma P)^{-1} R$

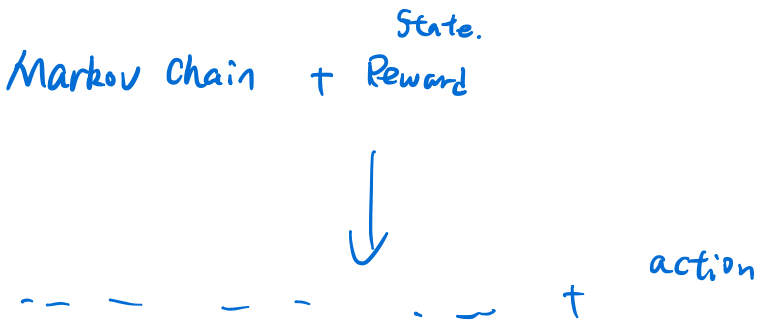
1e-6

$\gamma = 1$, $V = \begin{bmatrix} -12.529 \\ 1.4568 \\ 4.3210 \\ 10.0000 \\ 0.8045 \\ -22.5427 \\ 0 \end{bmatrix}$

$\gamma = 0.9$, $V = \begin{bmatrix} -5.0127 \\ 0.9427 \\ 4.0870 \\ 10.0000 \\ 1.9084 \\ -7.6376 \\ 0 \end{bmatrix}$

Example: Matrix Solution for Student MRP

Outline



Markov Decision Process

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

Definition

A Markov Decision Process is a tuple $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$

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

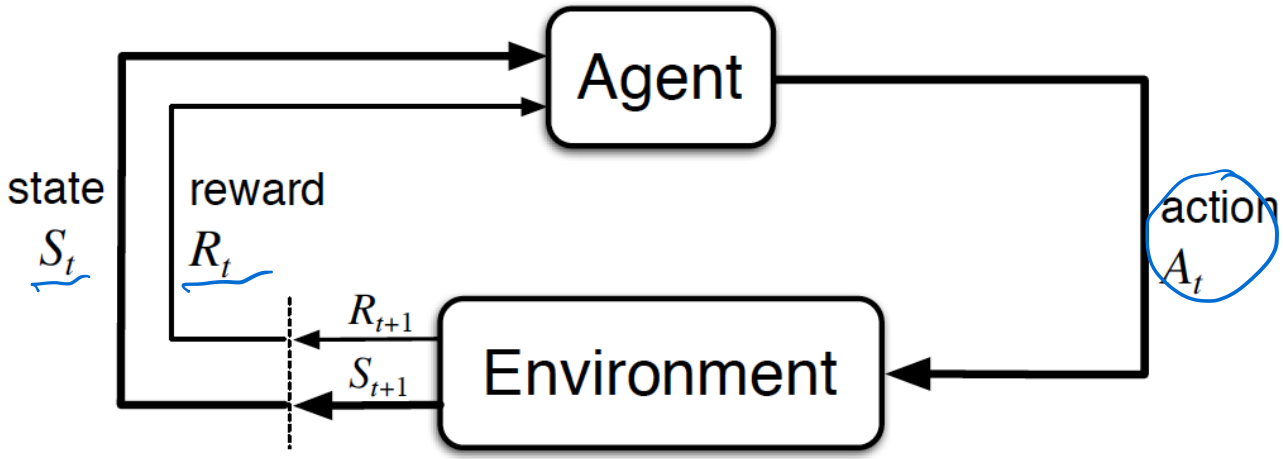
$$\mathcal{P}^a_{s,s'} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$$

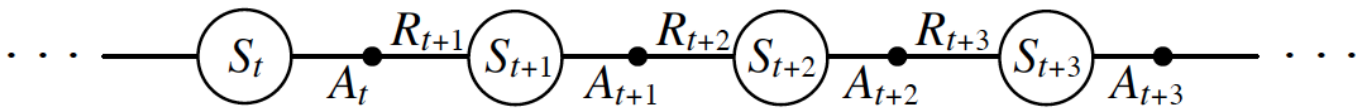
- $\mathcal{R}$ is a reward function,

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

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

The Agent-Environment Interface

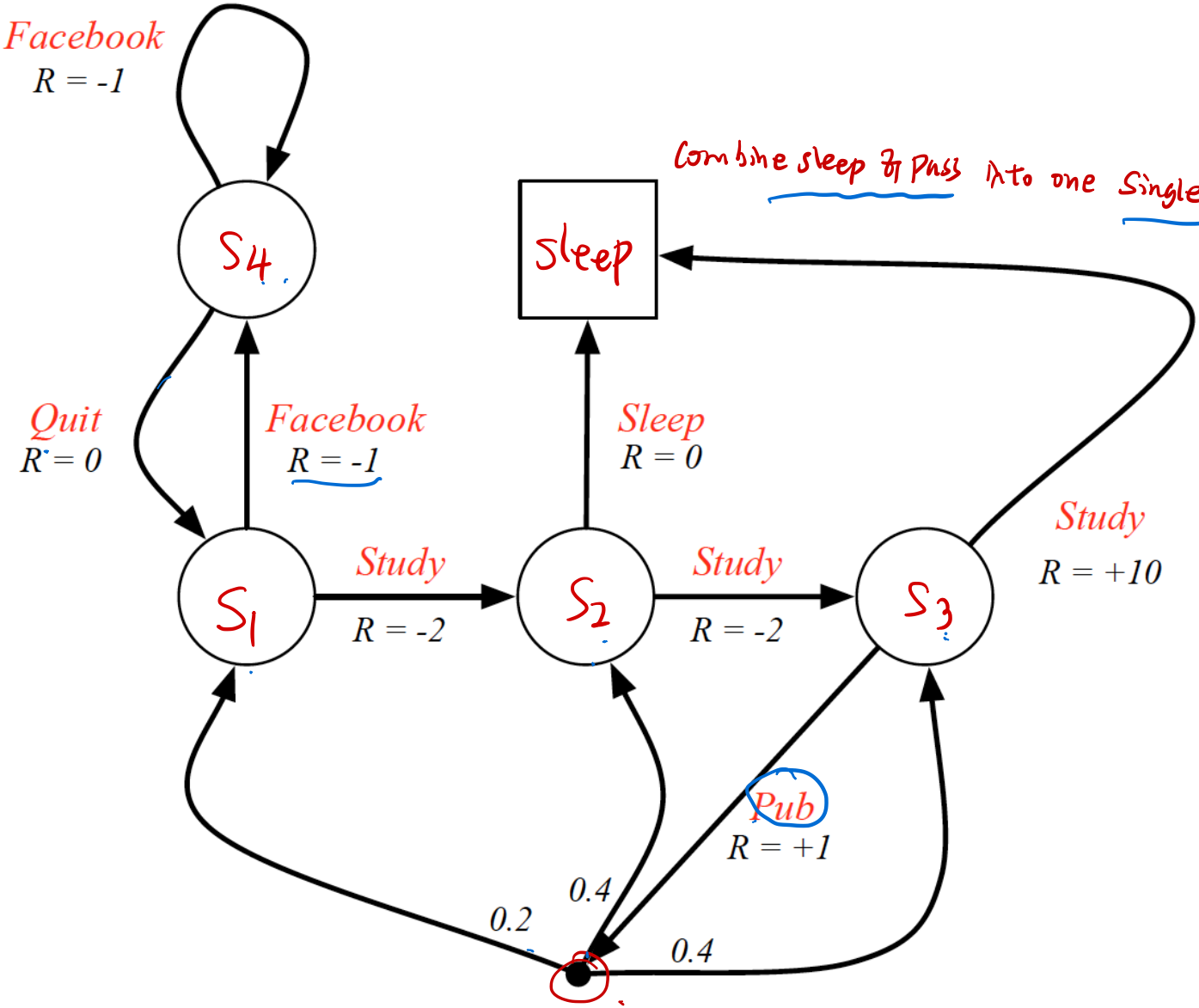




Example: Student MDP

MRP

Reward now depends on action & state.



Policy

Definition

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

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

$$\sum_{a \in \mathcal{A}} \pi(a|s) = 1$$

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

Policy

$1^0.\ P^{\pi}_{s,s'} = P^{\pi}[S_{t+1}=s' \mid S_t=s] \overset{\text{LoTP}}{=} \sum_{a\in A} P[S_{t+1}=s' \mid A_t=a, S_t=s]\, P(A_t=a \mid S_t=s)$

$= \sum_{a\in A} P^{a}_{ss'} \cdot \pi(a|s)$

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

$\pi(a|s) \triangleq P(A_t=a \mid S_t=s)$

$P^{a}_{s,s'} \triangleq P(S_{t+1}=s' \mid A_t=a, S_t=s)$

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

$$1^0.\quad \mathcal{P}^{\pi}_{s,s'} = \sum_{a\in\mathcal{A}} \pi(a|s)\mathcal{P}^{a}_{ss'}$$

$$2^0.\quad \mathcal{R}^{\pi}_{s} = \sum_{a\in\mathcal{A}} \pi(a|s)\mathcal{R}^{a}_{s}$$

$R^{\pi}_{s} = E_{\pi}[R_{t+1} \mid S_t=s] \overset{\text{LoTE}}{=} \sum_{a\in A} E[R_{t+1} \mid S_t=s, A_t=a]\, P(A_t=a \mid S_t=s)$

$= \sum_{a\in A} R^{a}_{s} \cdot \pi(a|s)$

Value Function

Adam's Law + Markov property

Definition

The *state-value function* $v_\pi(s)$ of an MDP is the expected return starting from state s, and then following policy π

$$v_\pi(s) = \mathbb{E}_\pi [G_t \mid S_t = s]$$

$$v_\pi(S_t) = \mathbb{E}_\pi [G_t \mid S_t]$$

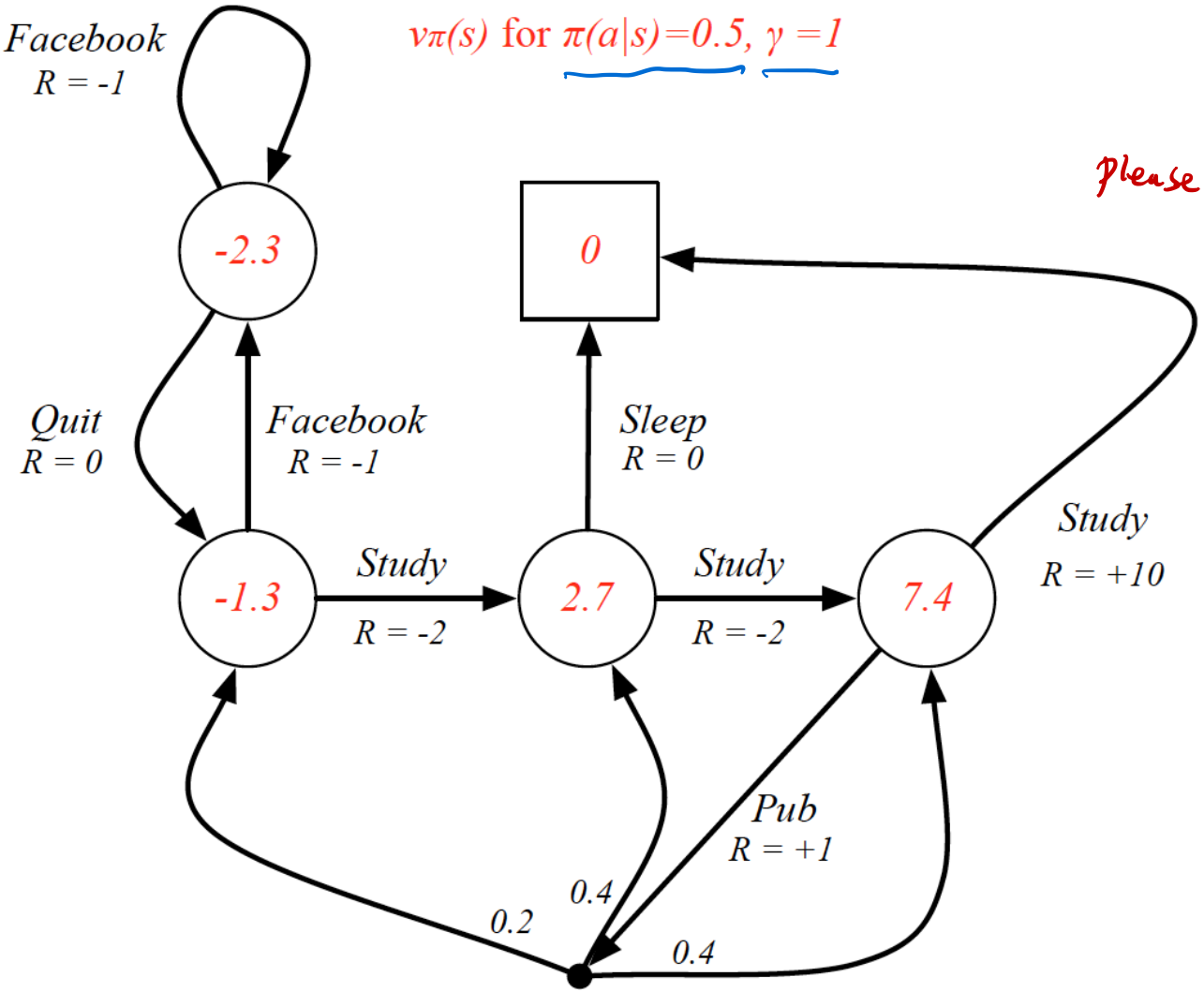
Definition

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

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

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

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,

$$v_\pi(s) = \mathbb{E}_\pi \left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s \right]$$

$$V_\pi(S_t) = E_\pi [R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t]$$

Same proof as MRP

You can try it!

The action-value function can similarly be decomposed,

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

proof next

$$q_\pi(S_t, A_t) = E_\pi [R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t, A_t]$$

Bellman Expectation Equation

1°. $q_\pi(s,a) = E_\pi[G_t \mid S_t=s, A_t=a]$

$q_\pi(S_t, A_t) = E_\pi[G_t \mid S_t, A_t]$, $q_\pi(S_{t+1}, A_{t+1}) = E_\pi[G_{t+1} \mid S_{t+1}, A_{t+1}]$

2°. Adam's law with extra conditioning

$$E[Y|Z] = E[E[Y|X,Z]|Z]$$

Let $Y = G_{t+1}$, $Z = (S_t, A_t)$, $X = (S_{t+1}, A_{t+1})$,

then we have

$$E[G_{t+1} \mid S_t, A_t] = E[E[G_{t+1} \mid S_{t+1}, A_{t+1}, S_t, A_t] \mid S_t, A_t]$$

Markov property

$$= E[E[G_{t+1} \mid S_{t+1}, A_{t+1}] \mid S_t, A_t]$$

$$= E[q_\pi(S_{t+1}, A_{t+1}) \mid S_t, A_t]$$

$$\Rightarrow E_\pi[G_{t+1} \mid S_t=s, A_t=a] = E_\pi[q_\pi(S_{t+1}, A_{t+1}) \mid S_t=s, A_t=a]$$

Bellman Expectation Equation

3°. Then we have

$$q_\pi(s,a) = E_\pi[G_t \mid S_t=s, A_t=a]$$

$G_t = R_{t+1} + \gamma G_{t+1}$

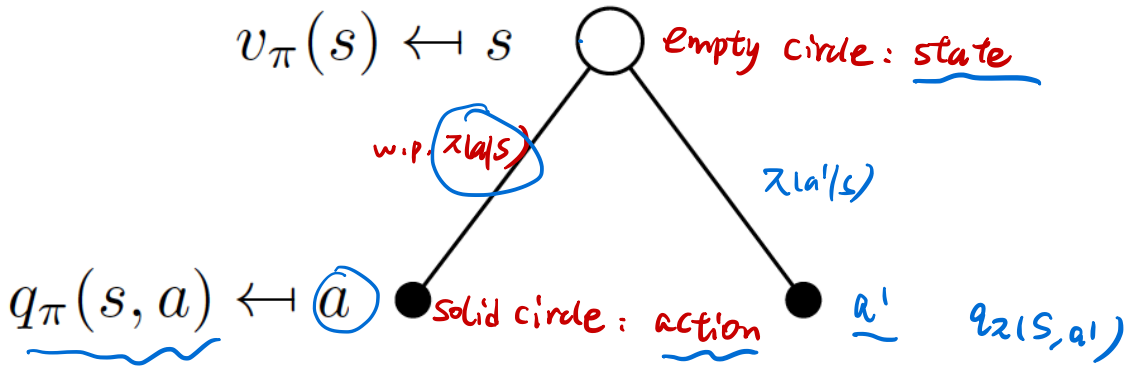
$$= E_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t=s, A_t=a]$$

$$= E_\pi[R_{t+1} \mid S_t=s, A_t=a] + \gamma E_\pi[G_{t+1} \mid S_t=s, A_t=a]$$

$$= E_\pi[R_{t+1} \mid S_t=s, A_t=a] + \gamma E_\pi[q_\pi(S_{t+1}, A_{t+1}) \mid S_t=s, A_t=a]$$

$$= E_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t=s, A_t=a]$$

Bellman Expectation Equation for V^π

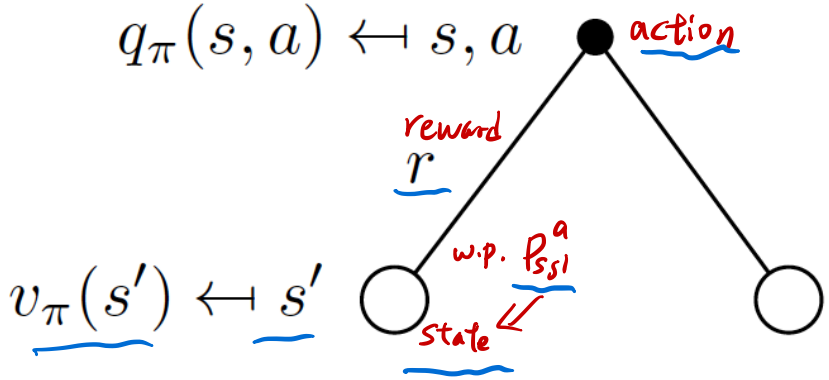


$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s,a)$$

$$V_\pi(s) = E_\pi[G_t | S_t = s] \overset{\text{LOTE}}{=} \sum_{a \in A} E_\pi[G_t | S_t = s, A_t = a] P(A_t = a | S_t = s)$$

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

Bellman Expectation Equation for Q^π



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

Proof next

Bellman Expectation Equation for Q^π

1^0. $\mathbb{E}[q_\pi(S_{t+1}, A_{t+1}) \mid S_{t+1}=s', S_t=s, A_t=a]$

$\overset{\text{Markov Property}}{=} \mathbb{E}[q_\pi(S_{t+1}, A_{t+1}) \mid S_{t+1}=s']$

$\overset{\text{LoTE}}{=} \sum_{a\in A} \mathbb{E}[q_\pi(S_{t+1}, A_{t+1}) \mid S_{t+1}=s', A_{t+1}=a] \cdot P(A_{t+1}=a \mid S_{t+1}=s')$

$= \sum_{a\in A} Q_\pi(s', a)\, \pi(a \mid s')$

$= V_\pi(s')$

2^0. Then we have $\mathbb{E}[q_\pi(S_{t+1}, A_{t+1}) \mid S_t=s, A_t=a]$

$\overset{\text{LoTE}}{=} \sum_{s'\in S} \mathbb{E}[q_\pi(S_{t+1}, A_{t+1}) \mid S_{t+1}=s', S_t=s, A_t=a] \cdot P(S_{t+1}=s' \mid S_t=s, A_t=a)$

$\overset{1^0}{=} \sum_{s'\in S} V_\pi(s') \cdot P^a_{ss'}$

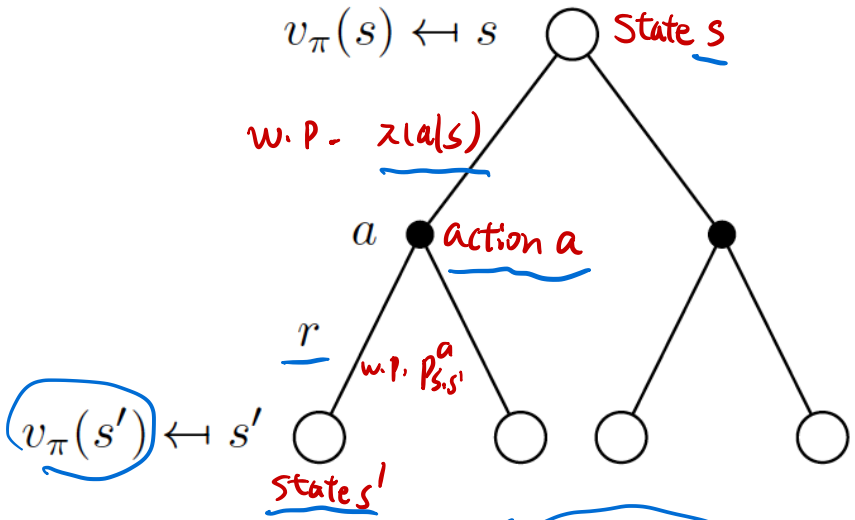
Bellman Expectation Equation for Q^π

3^0. Thus $q_\pi(s,a) = E_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t=s, A_t=a]$

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

2^0. $= R_s^a + \gamma \cdot \sum_{s' \in S} P_{s,s'}^a \cdot v_\pi(s')$

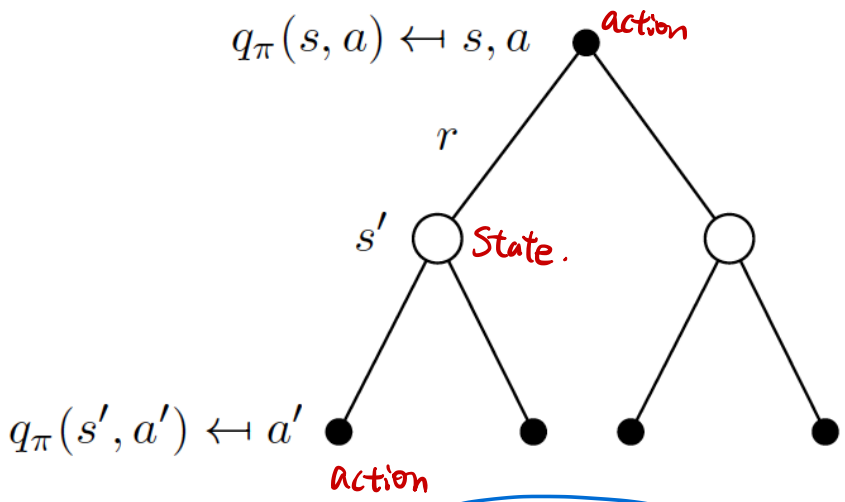
Bellman Expectation Equation for v_π



$$V_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s)\, q_\pi(s,a) \quad \text{and} \quad q_\pi(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{s,s'}^a V_\pi(s')$$

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

Bellman Expectation Equation for q_π



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

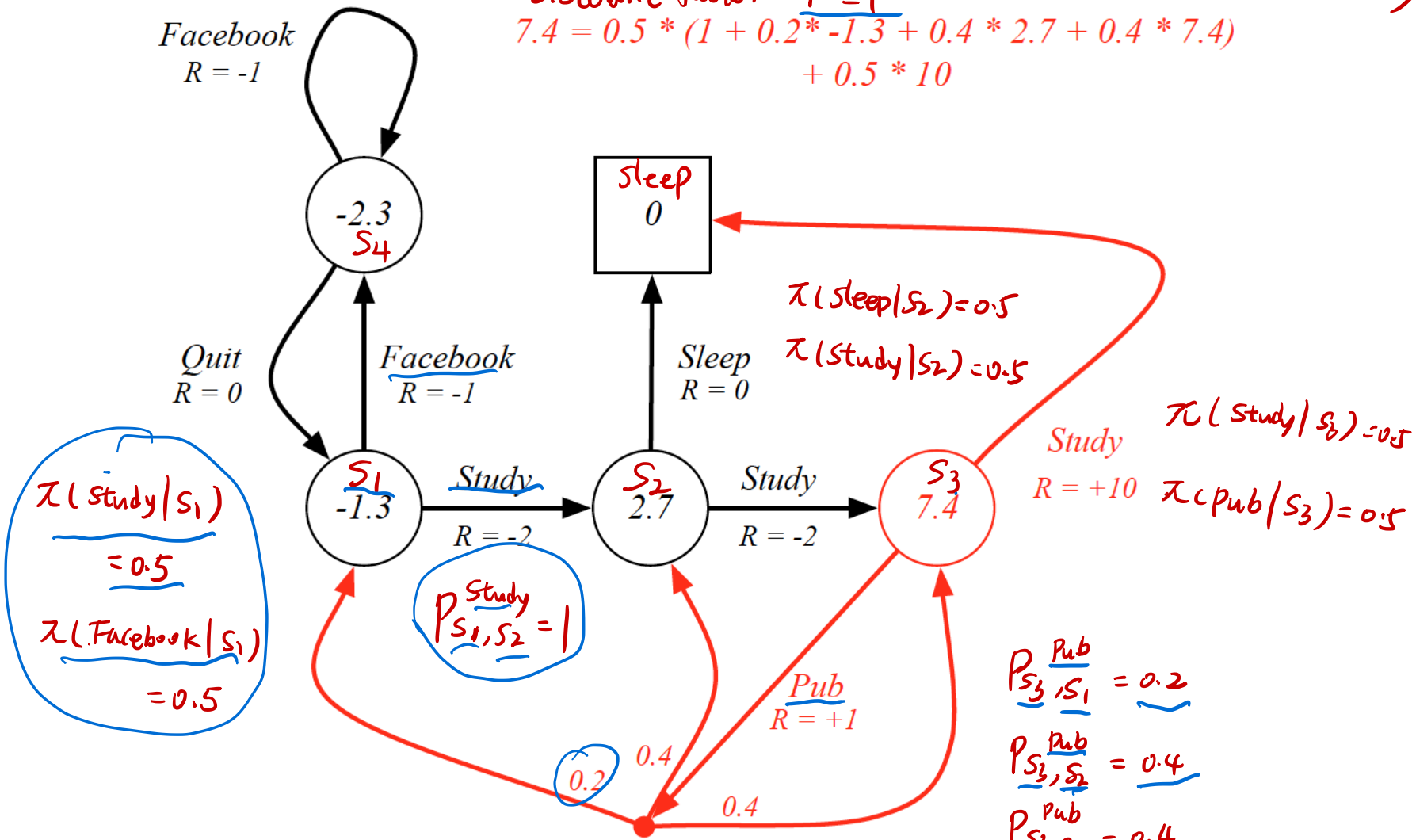
$q_\pi(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in S} P_{s,s'}^a V_\pi(s')$ and $V_\pi(s') = \sum_{a' \in A} \pi(a'|s') q_\pi(s',a')$

Example: Bellman Expectation Equation in Student MDP

Policy π : uniformly random (equal probability for all options)

discount factor $\gamma = 1$

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



Example: Bellman Expectation Equation in Student MDP

1°. $V_\pi(s) = \sum_{a \in A} \pi(a|s) \left[R_s^a + \gamma \sum_{s' \in S} P_{s,s'}^a V_\pi(s') \right]$

$V_j \triangleq V_\pi(s_j)$

2°. $V_1 = V_\pi(s_1) = \underset{\uparrow}{0.5}\ (R_{s_1}^{Study} + 1 \cdot 1 \cdot V_\pi(s_2)) + \underset{\uparrow}{0.5}\ (R_{s_1}^{Facebook} + 1 \cdot 1 \cdot V_\pi(s_4))$

$\pi(Study|s_1)$ $\quad$ $\pi(Facebook|s_1)$

$= 0.5\ (-2 + V_2) + 0.5\ (-1 + V_4)$

$V_2 = V_\pi(s_2) = 0.5\ (-2 + V_3) + 0.5\ (0 + 0)$

$V_3 = V_\pi(s_3) = 0.5\ (1 + 0.2V_1 + 0.4V_2 + 0.4V_3) + 0.5\ (0 + 0)$

$V_4 = V_\pi(s_4) = 0.5\ (0 + V_1) + 0.5\ (-1 + V_4)$

Example: Bellman Expectation Equation in Student MDP

3°. then We have

$V_1 = -1.3$
$V_2 = 2.7$
$V_3 = 7.4$
$V_4 = -2.3$

4°. we have obtained the state value. Now we turn to state-action value. (action.)

$$q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{s,s'}^a V_\pi(s')$$

$\Rightarrow$ $q_\pi(s_1, \text{study}) = -2 + 1 \cdot V_2 = -2 + 2.7 = 0.7$

$q_\pi(s_1, \text{Facebook}) = -1 + 1 \cdot V_4 = -1 - 2.3 = -3.3$

$q_\pi(s_2, \text{sleep}) = 0 + 0 = 0.$

$q_\pi(s_2, \text{Study}) = -2 + 1 \cdot V_3 = -2 + 7.4 = 5.4$

$q_\pi(s_3, \text{Study}) = 10 + 0 = 10$

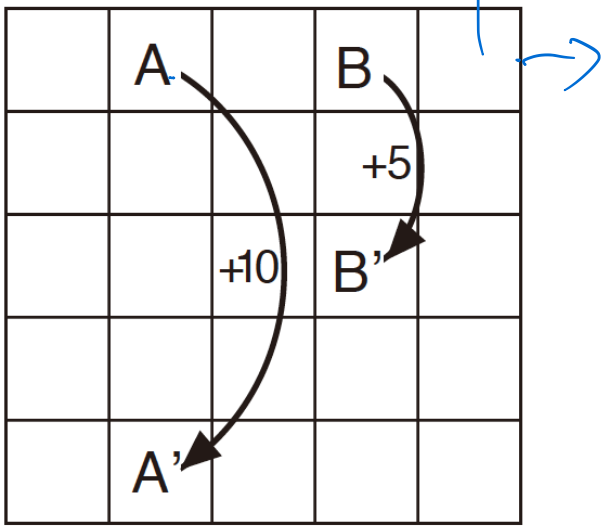
$q_\pi(s_3, \text{Pub}) = 1 + 0.2 V_1 + 0.4 V_2 + 0.4 V_3 = 1 + 0.2(-1.3) + 0.4(2.7) + 0.4(7.4) = 4.78$

$q_\pi(s_4, \text{Facebook}) = -1 + V_4 = -1 - 2.3 = -3.3$

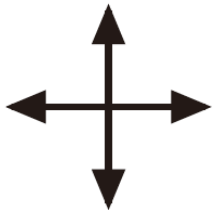
$q_\pi(s_4, \text{Quit}) = 0 + V_1 = 0 - 1.3 = -1.3$

Example: Bellman Expectation Equation in Student MDP

Example: Bellman Expectation Equation in Gridworld



(a)



Actions

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

(b)

try it !

What is the value function for the uniform random policy?

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

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

Example: Bellman Expectation Equation (Matrix Form)

$\gamma = 1$;

$$R_s^\pi = \sum_{a \in A} \pi(a|s) R_s^a \quad , \underline{R^\pi} = [R_s^\pi]$$

$$P_{s,s'}^\pi = \sum_{a \in A} \pi(a|s) P_{s,s'}^a \quad , \underline{P^\pi} = [P_{s,s'}^\pi]$$

$$\underline{P^\pi} = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ \text{Sleep} \end{array} \begin{array}{c} \begin{array}{cccccc} S_1 & S_2 & S_3 & S_4 & \text{Sleep} \end{array} \\ \begin{bmatrix} & 0.5 & & 0.5 & \\ & & 0.5 & & 0.5 \\ 0.5\times 0.2 = 0.1 & 0.5\times 0.4 = 0.2 & 0.5\times 0.4 = 0.2 & & 0.5 \\ 0.5 & & & 0.5 & \\ & & & & 1 \end{bmatrix} \end{array}$$

$$\underline{R^\pi} = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ \text{Sleep} \end{array} \begin{bmatrix} 0.5(-2)+0.5(-1) \\ 0.5(-2)+0.5(0) \\ 0.5(0)+0.5(1) \\ 0.5(-1)+0.5(0) \\ 0 \end{bmatrix} = \begin{bmatrix} -1.5 \\ -1 \\ 5.5 \\ -0.5 \\ 0 \end{bmatrix}$$

$$, \underline{V_\pi} = (I - \gamma P^\pi)^{-1} R^\pi$$

Example: Bellman Expectation Equation (Matrix Form)

Optimal Value Function

Definition

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

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

$$\pi^* \in \arg\max_\pi v_\pi(s)$$

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

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

$$q_*(s,a) = \max_\pi q_\pi(s,a) = \max_\pi \left[R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \cdot v_\pi(s') \right]$$

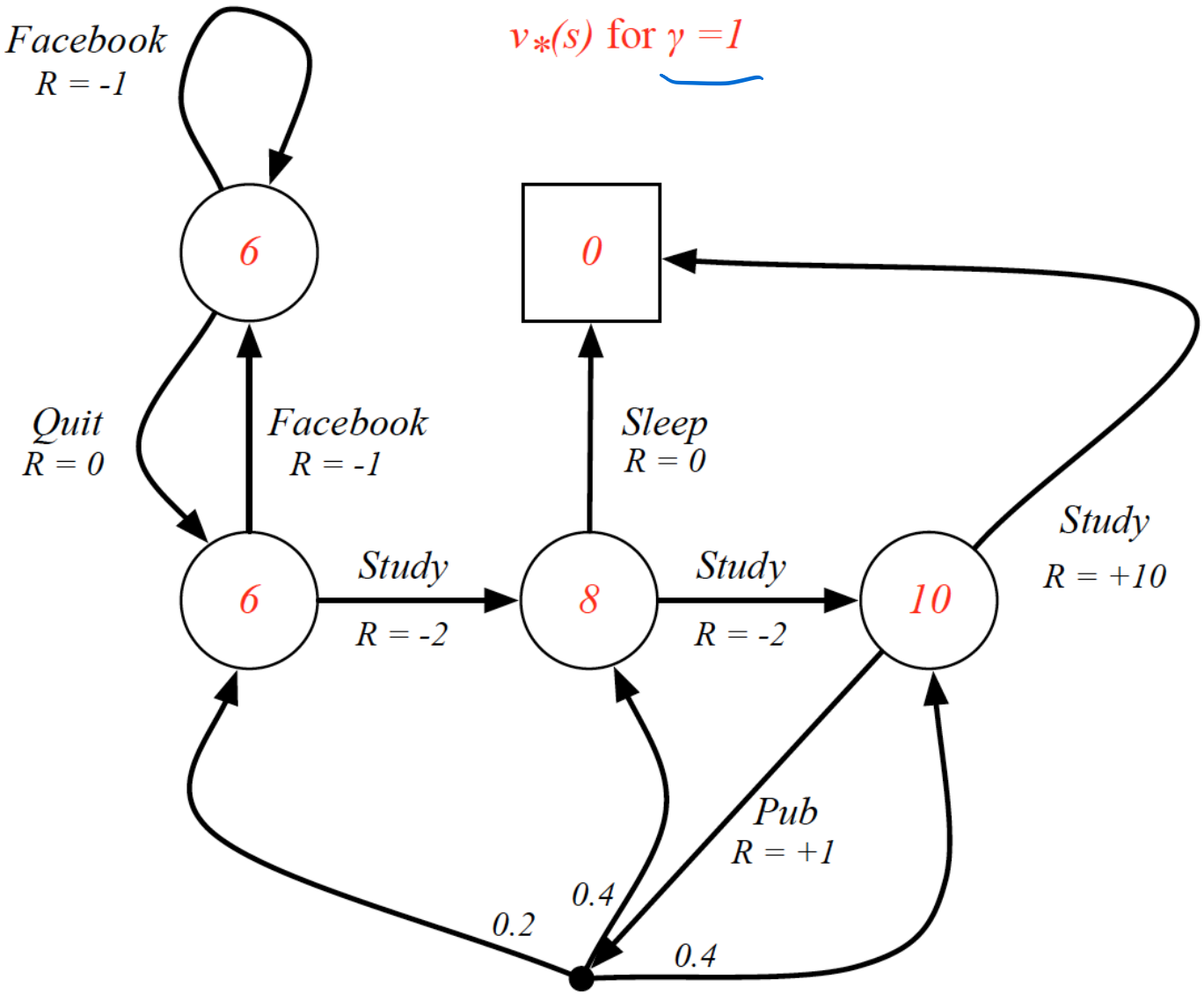
$$= R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \cdot \max_\pi v_\pi(s')$$

$$(\pi^*)' \in \arg\max_\pi q_\pi(s,a)$$

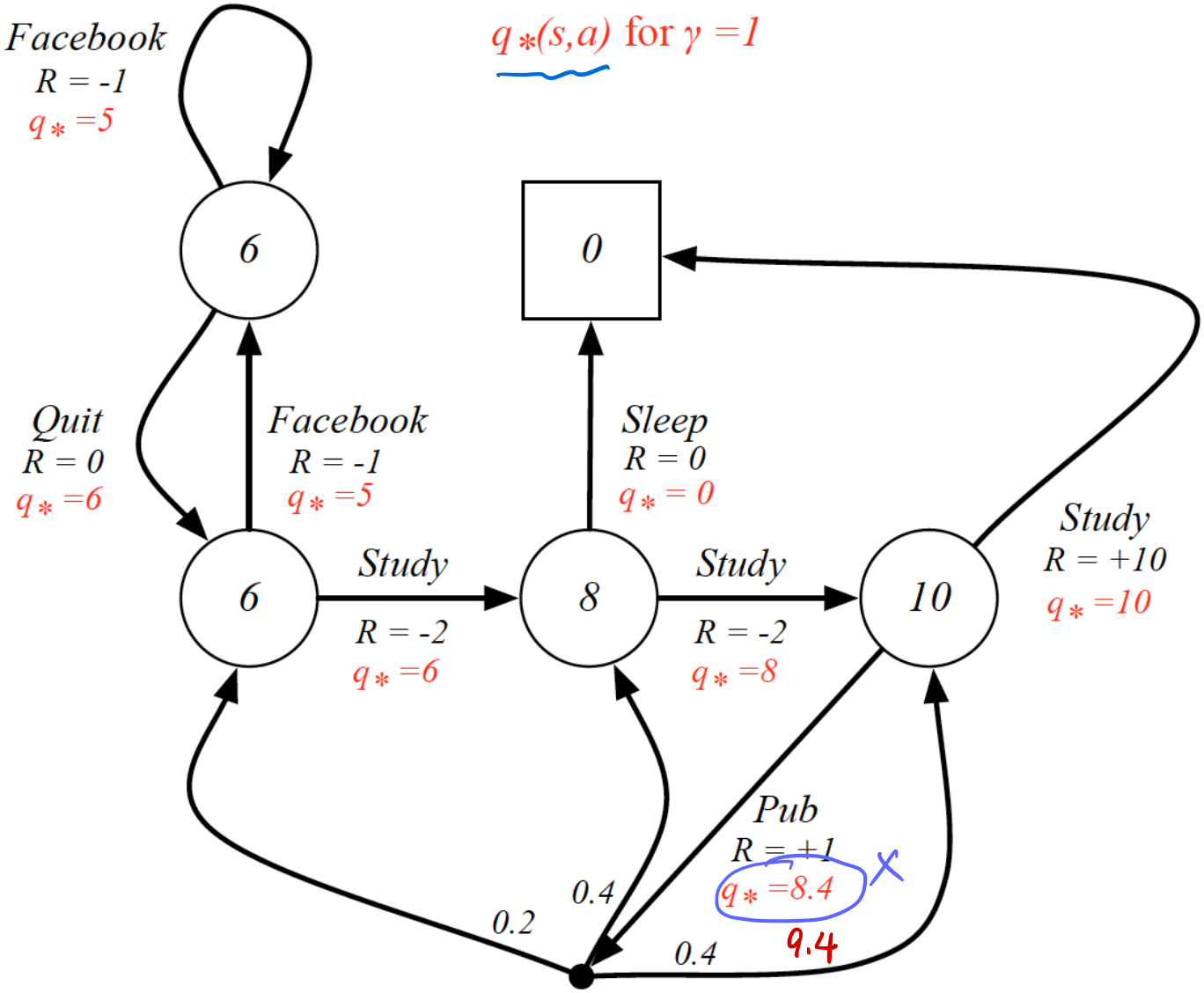
Intuitively $(\pi^*)' = \pi^*$.

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

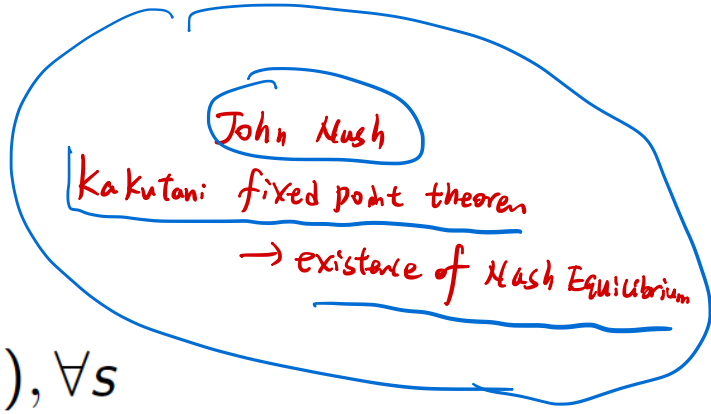
Example: Optimal Value Function for Student MDP



Example: Optimal Action-Value Function for Student MDP



Optimal Policy



Define a partial ordering over policies

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

Banach fixed point theorem.

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

Linear programming. ($c_f > 0$ given constants)

$$\max_{f \in F} p_f c_f \quad \text{s.t.} \sum_{f \in F} p_f = 1, \quad 0 \le p_f \le 1$$

Solution $\Rightarrow$ $f^0 = \arg\max_f c_f$, $p_f^* = \begin{cases} 1 & \text{if } f = f^0 \\ 0 & \text{otherwise} \end{cases}$

assume f^0 is unique.

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

If π' is a random optimal policy. $0 < \pi'(a|s) < 1$, $a \in A$.

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

$$\le \sum_{a \in A} \pi'(a|s) q_*(s,a)$$

$$\le \sum_{a \in A} \pi'(a|s) \cdot q_*(s,a^*)$$

$$= q_*(s,a^*) \quad (1)$$

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

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

$\max_{f \in F} c_f = c_{\max}$

$v_*(s) \ge v_{\pi^*}(s)$

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

$$= q_*(s,a^*) \quad (2)$$

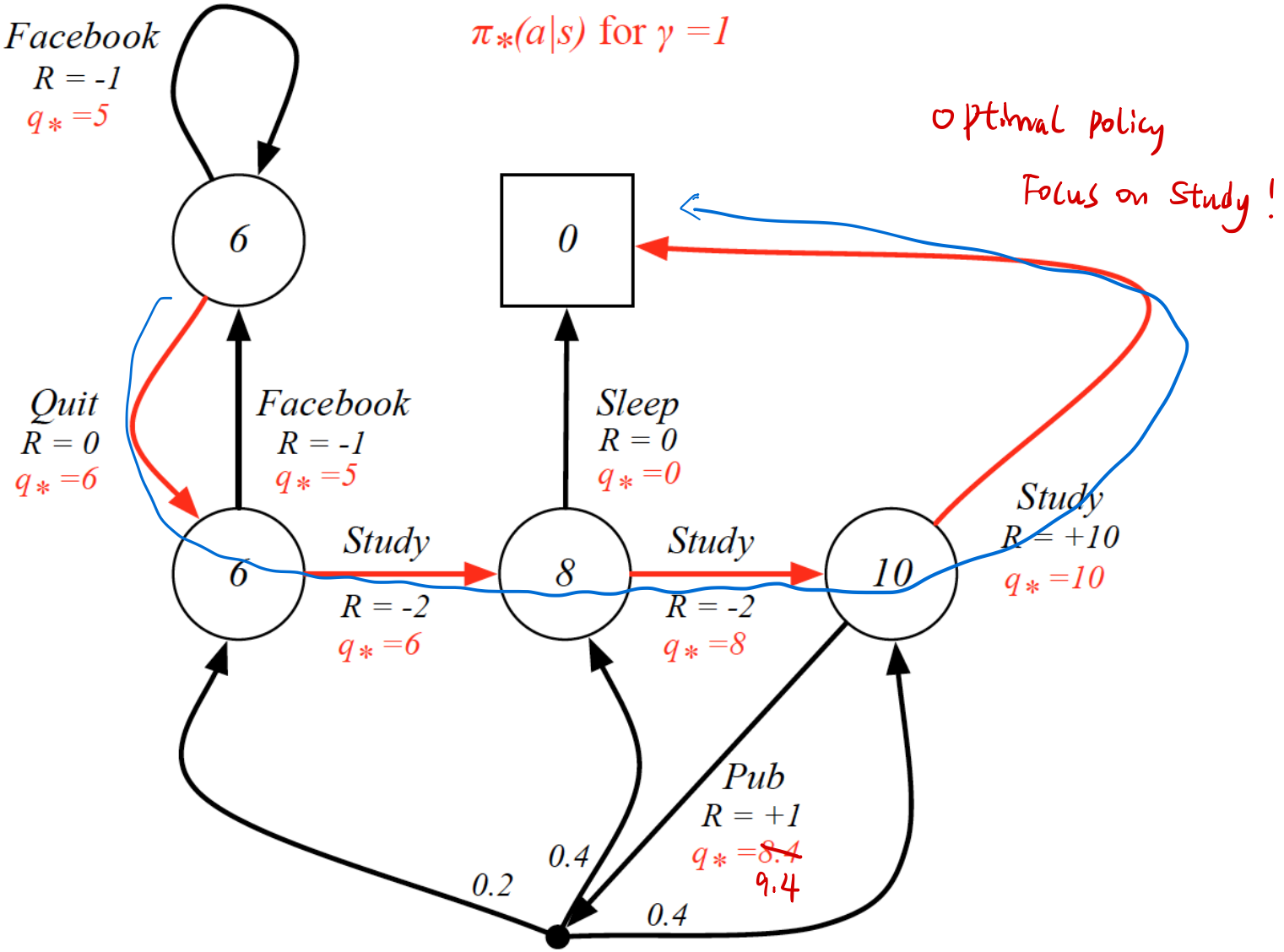
$1 > \pi(a|s) > 0$, $\forall a \in A$. random policy.

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

From (1) we know if you have a choice between picking exactly the action you want ($c_{\pi^*(a|s)}$) versus picking a probability distribution over potentially optimal and non-optimal actions, you would always prefer to pick exactly the best action.

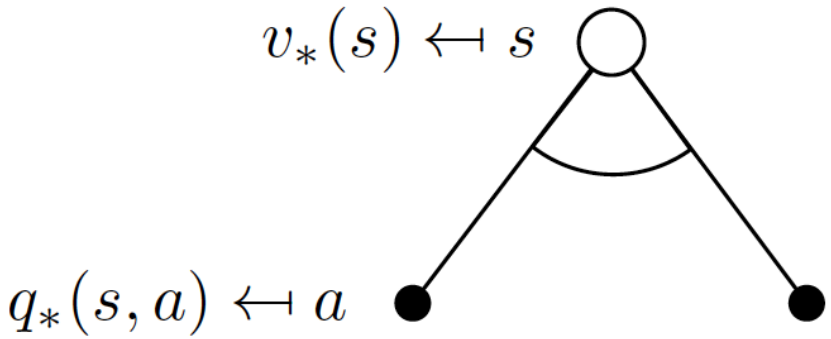
From (1)+(2) $v_*(s) = q_*(s,a^*) = \max_{a \in A} q_*(s,a)$

Example: Optimal Policy for Student MDP



Bellman Optimality Equation for v_*

The optimal value functions are recursively related by the Bellman optimality equations:



$$v_*(s) = \max_a q_*(s,a)$$

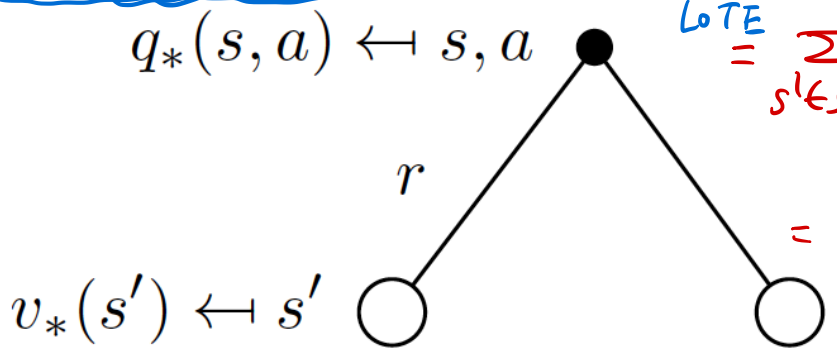
Bellman Optimality Equation for Q^*

1° $q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')$

2° $q_*(s,a) = \max_\pi q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \cdot \max_\pi v_\pi(s') = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \cdot v_*(s')$

$$q_*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a\right]$$

3° $E[R_{t+1} | S_t = s, A_t = a] = R_s^a$, $E[v_*(S_{t+1}) | S_t = s, A_t = a]$



LoTE $= \sum_{s' \in S} E[v_*(S_{t+1}) | S_{t+1} = s', S_t = s, A_t = a] \cdot P(S_{t+1} = s' | S_t = s, A_t = a)$

$= \sum_{s' \in S} \cdot v_*(s') \cdot P_{ss'}^a$

4° $q_*(s,a) = E[R_{t+1} + \gamma V_*(S_{t+1}) | S_t = s, A_t = a]$

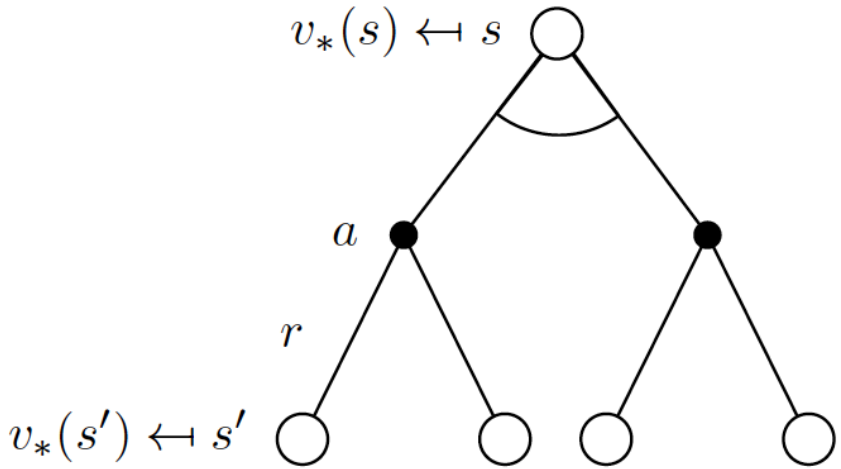
$$q_*(s,a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_*(s')$$

$q_*(S_t, A_t) = E[R_{t+1} + \gamma v_*(S_{t+1}) | S_t, A_t]$

Bellman Optimality Equation for V^*

$V^*(s) = \max_a q^*(s,a)$, and $q^*(s,a) = \mathbb{E}[R_{t+1} + \gamma V^*(S_{t+1}) | S_t = s, A_t = a]$

$$v_*(s) = \max_a \mathbb{E}\left[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a\right]$$

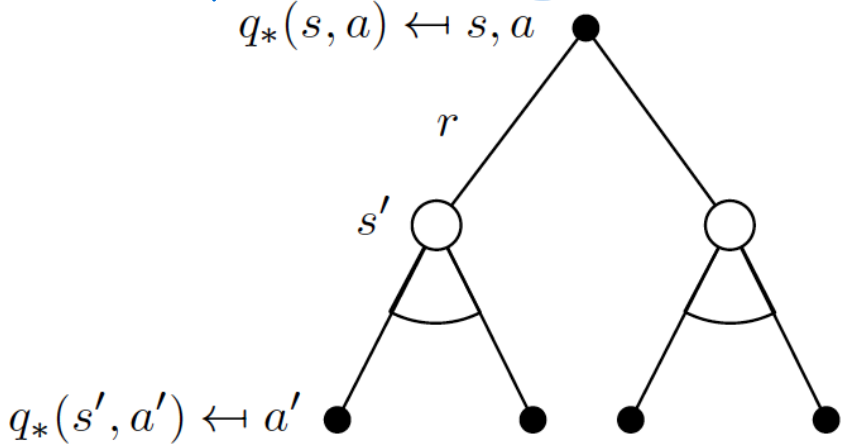


$q^*(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V^*(s')$

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

Bellman Optimality Equation for Q^*

$$q_*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \middle| S_t = s, A_t = a\right]$$



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

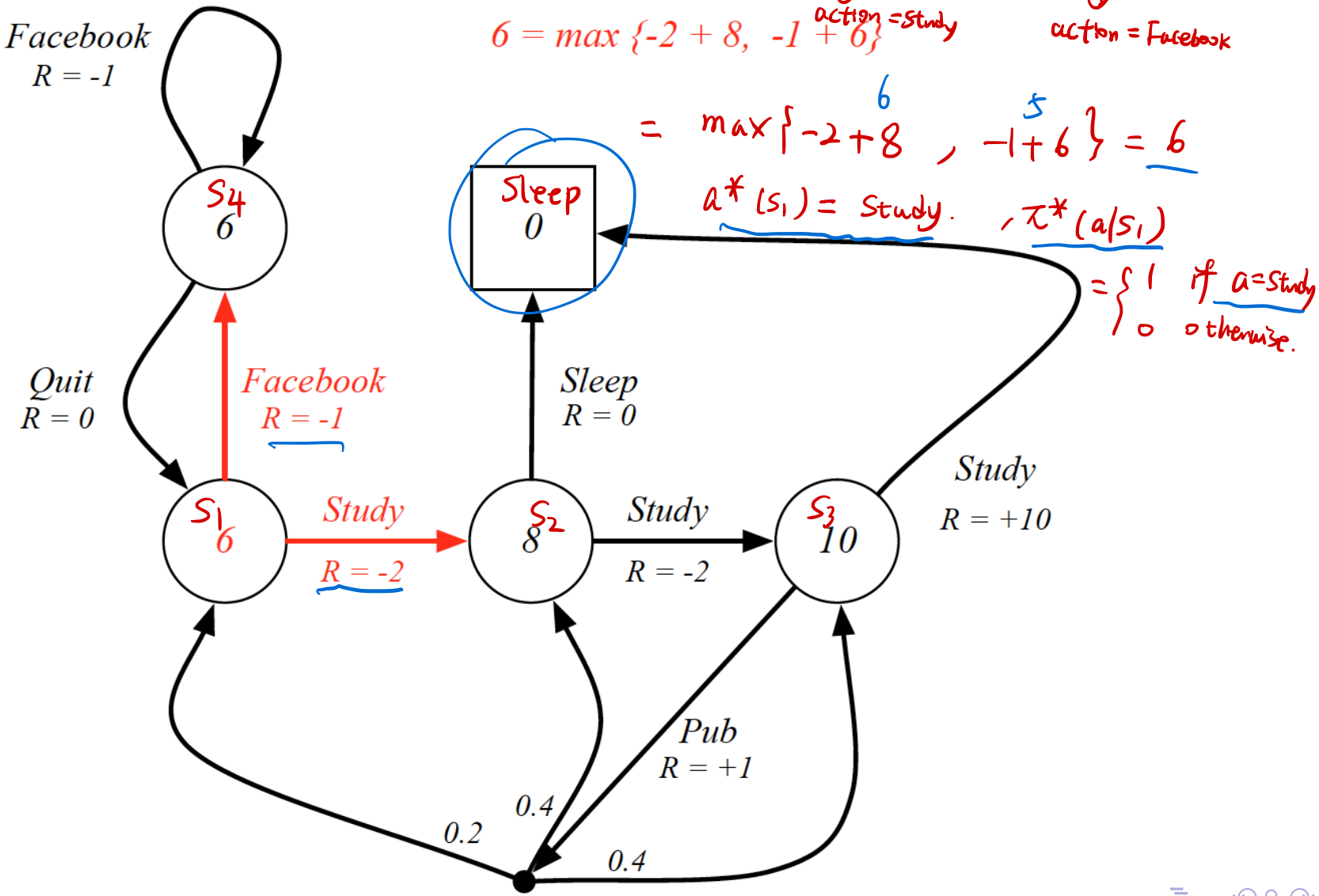
Example: Bellman Optimality Equation in Student MDP

$V^*(s_1) = \max\{-2 + V^*(s_2),\ -1 + V^*(s_4)\}$

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

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

$a^*(s_1) = \text{Study}$, $\pi^*(a|s_1) = \begin{cases} 1 & \text{if } a = \text{Study} \\ 0 & \text{otherwise.} \end{cases}$



Solution

Overall, we obtain $q_*(s,a)$ first, then obtain $V_*(s)$

1°. $q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{s,s'}^a V_\pi(s')$

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

Since $V_\pi(\text{sleep}) \equiv 0$, $\forall \pi$, then we have

$q_\pi(s_2, \text{sleep}) = R_{s_2}^{\text{sleep}} + 1 \cdot V_\pi(\text{sleep}) = 0 + 0$, $\forall \pi$

$\Rightarrow q_*(s_2, \text{sleep}) = 0$ ✓

border condition. terminal states

on the other hand,

$q_\pi(s_3, \text{study}) = R_{s_3}^{\text{study}} + 1 \cdot V_\pi(\text{sleep}) = 10 + 0 = 10$, $\forall \pi$

$\Rightarrow q_*(s_3, \text{study}) = 10$ ✓

Solution

$$q_*(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{s,s'}^a \cdot \max_{a'} q_*(s',a')$$

2°. $q_*(s_1, \text{Study}) = R_{s_1}^{\text{Study}} + \gamma P_{s_1,s_2}^{\text{Study}} \max_{a'} q_*(s_2,a')$

$= R_{s_1}^{\text{Study}} + \gamma P_{s_1,s_2}^{\text{Study}} \max\left(q_*(s_2,\text{Sleep}), q_*(s_2,\text{Study})\right)$

$= -2 + \max[0, q_*(s_2,\text{Study})]$

$q_*(s_1,\text{Facebook}) = R_{s_1}^{\text{Facebook}} + \gamma P_{s_1,s_4}^{\text{Facebook}} \cdot \max_{a'} q_*(s_4,a')$

$= R_{s_1}^{\text{Facebook}} + \gamma P_{s_1,s_4}^{\text{Facebook}} \cdot \max[q_*(s_4,\text{Facebook}), q_*(s_4,\text{Quit})]$

$= -1 + \max[q_*(s_4,\text{Facebook}), q_*(s_4,\text{Quit})]$

$q_*(s_2,\text{sleep}) = 0$

$q_*(s_2,\text{Study}) = -2 + \max[10, q_*(s_3,\text{Pub})]$

$q_*(s_3,\text{Study}) = 10$

Solution

$$q_*(s_3, \text{Pub}) = 1 + 0.2 \cdot \max[q_*(s_1, \text{Study}), q_*(s_1, \text{Facebook})] + 0.4 \max[q_*(s_2, \text{Study}), 0] + 0.4 \max[q_*(s_3, \text{Pub}), 10]$$

$$q_*(s_4, \text{Facebook}) = -1 + \max[q_*(s_4, \text{Facebook}), q_*(s_4, \text{Quit})]$$

$$q_*(s_4, \text{Quit}) = 0 + \max[q_*(s_1, \text{Study}), q_*(s_1, \text{Facebook})]$$

Solution

3°. $q_*(S_2, \text{Study}) = -2 + \max[10, q_*(S_3, \text{Pub})] \geq 8$

$q_*(S_1, \text{Study}) = -2 + \max[0, q_*(S_2, \text{Study})] \geq 6$

$= -2 + q_*(S_2, \text{Study})$

$q_*(S_4, \text{Facebook}) = -1 + \max[q_*(S_4, \text{Facebook}), q_*(S_4, \text{Quit})]$

$\overset{?}{=} -1 + q_*(S_4, \text{Facebook})$ ✗

$= -1 + q_*(S_4, \text{Quit}) < q_*(S_4, \text{Quit})$

$q_*(S_1, \text{Facebook}) = -1 + \max[q_*(S_4, \text{Facebook}), q_*(S_4, \text{Quit})]$

$= -1 + q_*(S_4, \text{Quit})$

Solution

$$q_*(s_4, \text{Quit}) = \max\left[q_*(s_1, \text{Study}), q_*(s_1, \text{Facebook}) \right]$$

$$= \max\left[q_*(s_1, \text{Study}), q_*(s_4, \text{Quit}) - 1 \right]$$

$$\overset{?}{=} q_*(s_4, \text{Quit}) - 1 \quad \text{(X)}$$

$$= q_*(s_1, \text{Study}) = -2 + q_*(s_2, \text{Study})$$

$$q_*(s_3, \text{Pub}) = 1 + 0.2 \max\left[\underbrace{q_*(s_4, \text{Quit})}_{q_*(s_1, \text{Study})}, \underbrace{-1 + q_*(s_4, \text{Quit})}_{q_*(s_1, \text{Facebook})} \right]$$

$$+ 0.4 \max\left[\underbrace{q_*(s_2, \text{Study})}_{\geq 8}, 0 \right]$$

$$+ 0.4 \max\left[q_*(s_3, \text{Pub}), 10 \right]$$

$$= 1 + 0.2\, q_*(s_4, \text{Quit}) + 0.4\, q_*(s_2, \text{Study}) + 0.4 \max\left[q_*(s_3, \text{Pub}), 10 \right]$$

$$= 1 + 0.2\left[-2 + q_*(s_2, \text{Study}) \right] + 0.4\, q_*(s_2, \text{Study}) + 0.4 \max\left[q_*(s_3, \text{Pub}), 10 \right]$$

$$= 0.6 + 0.6\, q_*(s_2, \text{Study}) + 0.4 \max\left[q_*(s_3, \text{Pub}), 10 \right]$$

Solution

4°. if $\max[q_*(s_3, \text{Pub}), 10] = q_*(s_3, \text{Pub})$ (≥ 10)

$\Rightarrow q_*(s_3, \text{Pub}) = 0.6 + 0.6\, q_*(s_2, \text{Study}) + 0.4\, q_*(s_3, \text{Pub})$

$\Rightarrow q_*(s_3, \text{Pub}) = 1 + q_*(s_2, \text{Study})$

On the other hand, we have

$q_*(s_2, \text{Study}) = -2 + \max[10, q_*(s_3, \text{Pub})] = -2 + q_*(s_3, \text{Pub})$

$\Rightarrow q_*(s_3, \text{Pub}) = -1 + q_*(s_3, \text{Pub})$ ✗

Thus $q_*(s_3, \text{Pub}) < 10$, $\max[q_*(s_3, \text{Pub}), 10] = 10$

Solution

5°. Then we have

$$q_*(s_3, \text{pub}) = 0.6 + 0.6\, q_*(s_2, \text{study}) + 4 = 4.6 + 0.6\, q_*(s_2, \text{study})$$

on the other hand, $q_*(s_2, \text{study}) = -2 + \max[0, q_*(s_3, \text{pub})]$

$$= -2 + 10 = 8$$

$$\Rightarrow \quad q_*(s_1, \text{study}) = -2 + q_*(s_2, \text{study}) = -2 + 8 = 6$$

$$q_*(s_3, \text{pub}) = 4.6 + 0.6\, q_*(s_2, \text{study}) = 4.6 + 0.6 \times 8 = 9.4$$

$$q_*(s_4, \text{quit}) = q_*(s_1, \text{study}) = 6$$

$$q_*(s_4, \text{Facebook}) = -1 + q_*(s_4, \text{quit}) = -1 + 6 = 5$$

$$q_*(s_1, \text{Facebook}) = -1 + q_*(s_4, \text{quit}) = -1 + 6 = 5$$

Solution

6°. $V^*(s) = \max_a q_*(s,a)$

$\Rightarrow V_*(s_1) = \max_a q_*(s_1,a) = \max[q_*(s_1,\text{Study}), q_*(s_1,\text{Facebook})]$

$= \max[6,5] = 6$

$V^*(s_2) = \max[q_*(s_2,\text{Study}), q_*(s_2,\text{Sleep})]$

$= \max[8,0] = 8$

$V^*(s_3) = \max[q_*(s_3,\text{Study}), q_*(s_3,\text{Pub})]$

$= \max[10, 9.4] = 10$

$V^*(s_4) = \max[q_*(s_4,\text{Quit}), q_*(s_4,\text{Facebook})]$

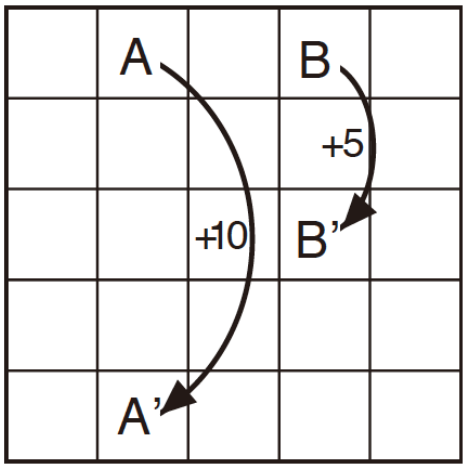
$= \max[6,5] = 6$

$V^*(\text{Sleep}) \equiv 0$

Solution

Solution

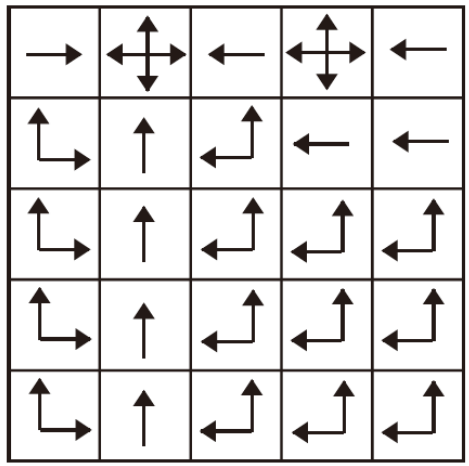
Example: Bellman Optimality Equation in Gridworld



a) gridworld

22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7

b) v_*



c) π_*

What is the optimal value function over all possible policies?
What is the optimal policy?

try it !

Solution

Solving the Bellman Optimality Equation in General

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

Control. LQR

Solving the Bellman Optimality Equation

- Finding an optimal policy by solving the Bellman Optimality Equation requires the following:
  - accurate knowledge of environment dynamics
  - we have enough space and time to do the computation
  - the Markov Property
- How much space and time do we need?
  - polynomial in number of states
  - BUT, number of states is often huge
  - So exhaustive sweeps of the state space are not possible

Approximation and Reinforcement Learning

- RL methods: Approximating Bellman optimality equations
- Balancing reward accumulation and system identification (model learning) in case of unknown dynamics
- The on-line nature of reinforcement learning makes it possible to approximate optimal policies in ways that put more effort into learning to make good decisions for frequently encountered states, at the expense of less effort for infrequently encountered states.

Outline

Main References

- Reinforcement Learning: An Introduction (second edition), R. Sutton & A. Barto, 2018.
- RL course slides from Richard Sutton, University of Alberta.
- RL course slides from David Silver, University College London.