Introduction to MDP

Tuesday, February 11, 2020 11:37 PM

Reinforcement Learning

Formulating RL problem as Markov Decision Process (MDP)

Markov Property

A state S_t is Markov, if and only if:

$$P[S_t|S_{t-1}] = P[S_t|S_1, S_2, \ldots, S_{t-1}]$$

$$p(S_t | S_{t-1}) = p(S_t | S_1, S_2 \ldots S_{t-1})$$

Markov Process

$S_1, S_2 \ldots S_n$

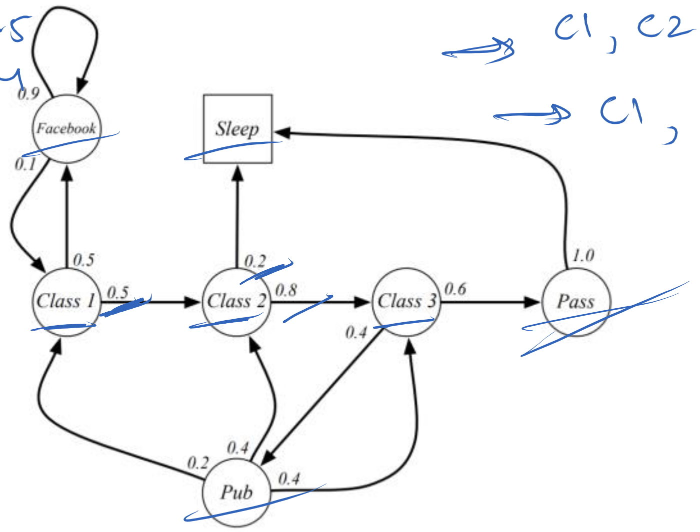
$\langle S, P \rangle$

[C1] $p(C_2|C_1) = 0.5$

$\Rightarrow$ C1, C2, S

C1
C2
C3
Pa
Pu
S
F

$P(C_2|C_1) = 0.5$
$P(Pu|C_3) = 0.4$

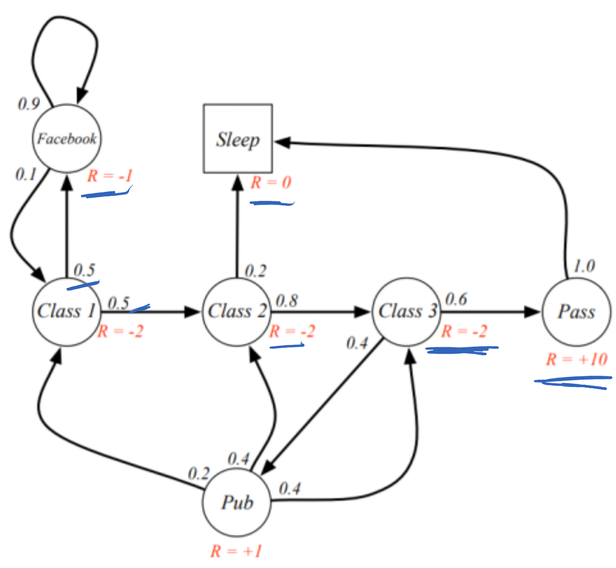


→ C1, C2, S

→ C1, C2, C3, Pu, C2, C3, Pa

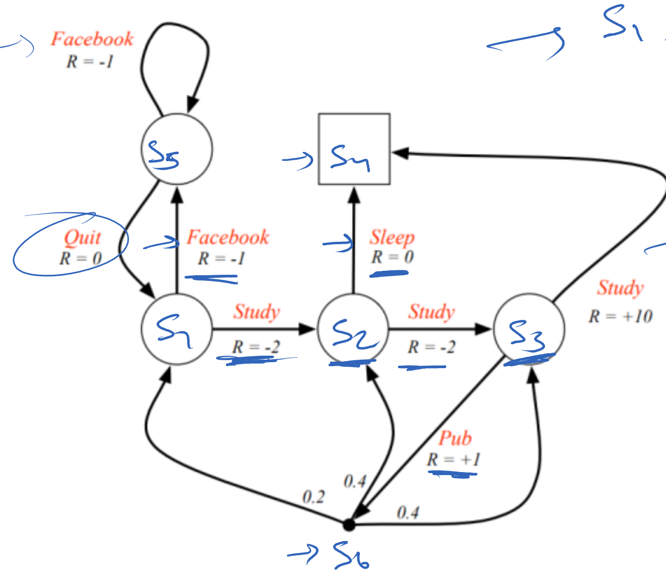
Markov Reward Process

< S, P, R >



Markov Decision Process

$\langle S, P, R, A \rangle$



→ S_1, Study, S_2, Study, S_3, Study, S_4.

→ S_1, Facebook, S_5, Facebook, S_5, Quit, S_1, Study, S_2, Sleep, S_4.

Components of MDP

Tuesday, February 11, 2020 11:37 PM

Reinforcement Learning

Understanding the components of MDPs

Deterministic and Stochastic Processes

$$f(x) \rightarrow y$$

$$g(x) \rightarrow \underbrace{[P(y_1), P(y_2) \ldots P(y_n)]}$$

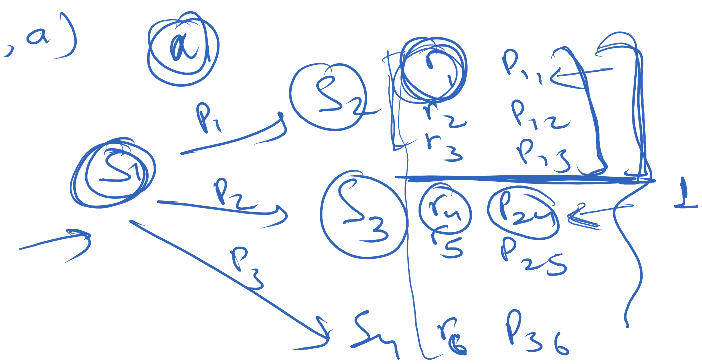
$$g(x) \rightarrow y_i$$

Components of MDP

$< S, A, R, P >$
$S = Finite\ set\ of\ state$
$A = Finite\ set\ of\ actions$
$R = Finite\ set\ of\ all\ rewards$
$P = Environment\ dynamics\ function$

Understanding Environment Dynamics function

$p(s'|s,a)$



$$p(s'|s,a) = \Pr\{S_t = s' | S_{t-1} = s\ , A_{t-1} = a\}$$

$$\sum_{s'} p(s'|s,a) = 1, \qquad \forall\ s \in S, a \in A(s)$$

$$p(s',r|s,a) = \Pr\{S_t = s', R_t = r \,|\, S_{t-1} = s, A_{t-1} = a\}$$

$$\sum_{s'}\sum_{r} p(s',r|s,a) = 1, \qquad \forall\ s \in S, a \in A(s)$$

$$p(s'|s,a) = \sum_{r} p(s',r|s,a)$$

$P_1 + P_2 + P_3 = 1$

$$r(s,a) = \mathbb{E}[R_t | S_{t-1} = s, A_{t-1} = a] = \sum_{r} r \sum_{s'} p(s',r|s,a)$$

$\frac{P_{11}}{P_{11}+P_{12}+P_{13}}$

$$r(s,a,s') = \mathbb{E}[R_t | S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r} r \frac{p(s',r|s,a)}{p(s'|s,a)}$$

$p(S_2 | s, a_1) = P_1$

$p(s', r | s, a)$

$= P_{11} + P_{12} + P_{13}$

$r(S_1, a_1) = r_1 \times P_{11} + r_2 \times P_{12} + \ldots\ r_6 \times P_{36}$

$r(S_1, a_1, S_2) =$

Choices (Actions)

States.

Reward.

MDP Example

Saturday, February 15, 2020 1:43 PM

Reinforcement Learning
MDP example

Recycling Robot example for MDP

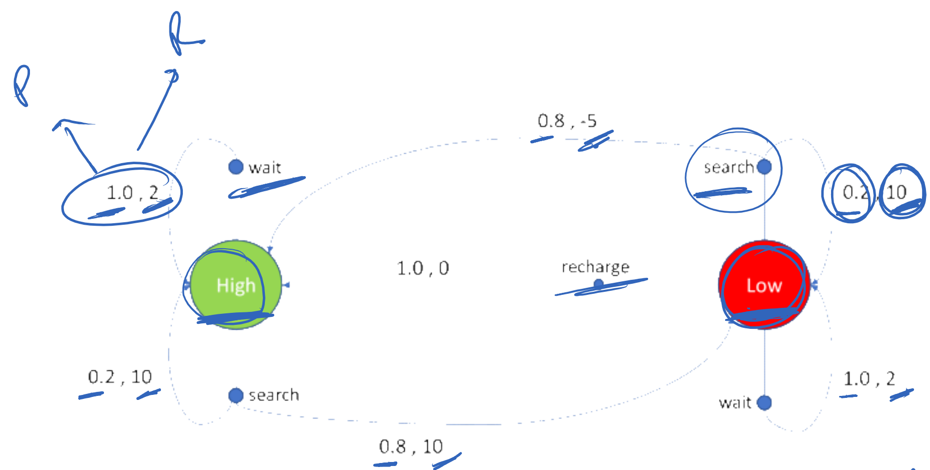
$A(\text{high}) = \{\text{search}, \text{wait}\}$

$A(\text{low}) = \{\text{search}, \text{wait}, \text{recharge}\}.$

$S = \{\text{low}, \text{high}\}$

$A = \{\text{search}, \text{wait}, \text{recharge}\}$

$R = \{R_{\text{search}}, R_{\text{wait}}, R_{\text{rescued}}\}$



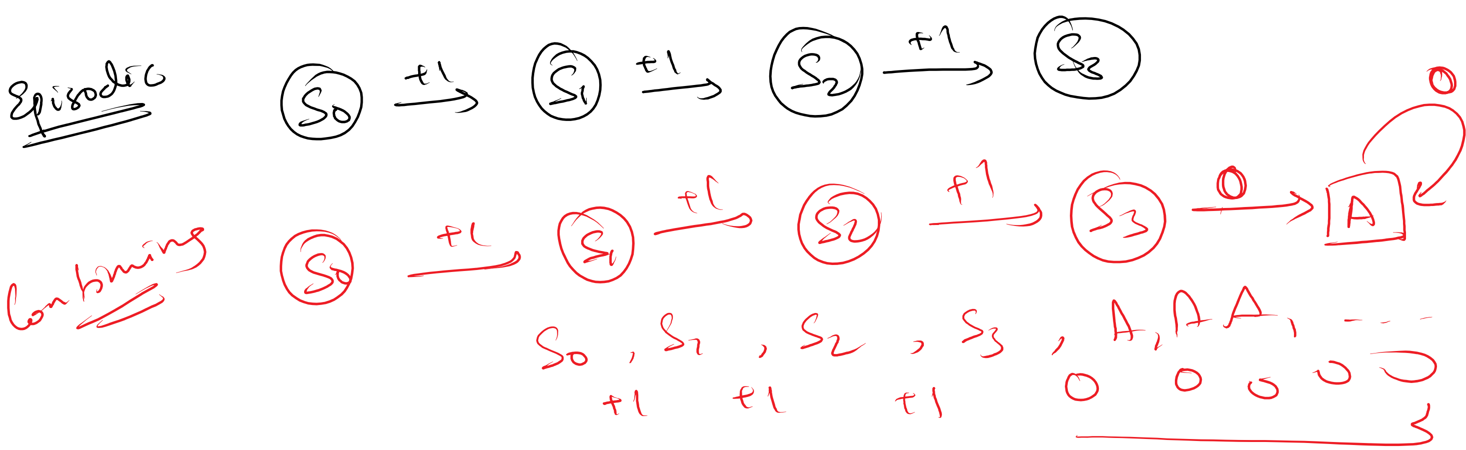
(s, a, s')

s	a	s'	$r(s,a,s')$	$p(s'\|s,a)$	$p(s',r\|s,a)$
low	search	low	10	0.2	0.2
low	search	high	-5	0.8	0.8
low	wait	low	2	1.0	1.0
low	recharge	high	0	1.0	1.0
high	wait	high	2	1.0	1.0
high	search	high	10	0.2	0.2
high	search	low	10	0.8	0.8

Episodic and continuing tasks

Saturday, February 15, 2020 11:50 PM

Reinforcement Learning
Episodic and continuing tasks



Rewards and return

Sunday, February 16, 2020 11:56 AM

Reinforcement Learning
Rewards and Returns

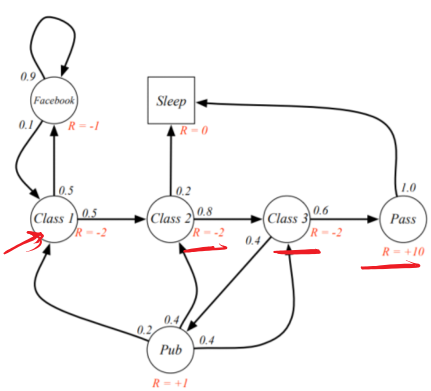
Rewards and returns

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

With discounting,

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

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



Rewards

C1 C2 C3 Passed Sleep.

$G(C_1) = -2 + (-2) + 10 + 0 = +6$

$t \rightarrow R_t$

return, $G_t = \underbrace{R_{t+1} + R_{t+2} + \cdots + R_T}$

Return

$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots\cdots$

$\gamma \in [0, 1]$

$\gamma < 1$

$R_t = 1$ $\quad G_t = 1 + \gamma + \gamma^2 + \gamma^3 + \cdots$

$= \frac{1}{1-\gamma}$

$(\gamma^{k-1}) R_{t+k}$

$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$

$= R_{t+1} + \gamma \left(\underbrace{R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4}}_{G_{t+1}} \right)$

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

$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$

$= \sum_{k=0}^{T} \gamma^k R_{t+k+1}$

Policy

Sunday, February 16, 2020 4:30 PM

Reinforcement Learning
Policy

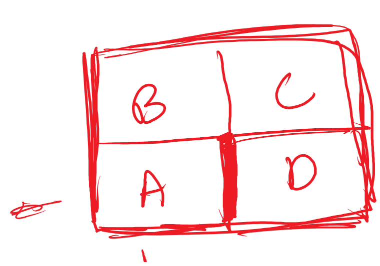
$\pi(s)$

Deterministic $\pi(s_1) \longrightarrow a$

stochastic $\pi(s_1) \longrightarrow p(a_1),\ p(a_2),\ p(a_3) \ldots$

$\pi(a_1|s) \longrightarrow y_1 \longrightarrow p(a_1)$

$\pi(a_2|s) \longrightarrow y_2 \longrightarrow p(a_2)$



Actions

$\uparrow$ 25%

$\leftarrow$ 25%

$\rightarrow$ 25%

$\downarrow$ 25%

$\rightarrow$

$\uparrow \quad \downarrow$

$\pi(\uparrow|A) = 90\%$ 100%

$\pi(\uparrow|A) = 25\%$

$\pi(\leftarrow|A) = 25\%$

State and action values

Sunday, February 16, 2020 4:53 PM

Reinforcement Learning

State and Action Value functions

State Value function

$v_\pi(s)$

$$v_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s] = \mathbb{E}_\pi[\sum_{k=0}^{T} \gamma^k R_{t+k+1} | S_t = s], \quad \forall\ s \in S$$

State value

$G_t \rightarrow R \rightarrow A \rightarrow \pi$

Action Value function

$q_\pi(a, s)$

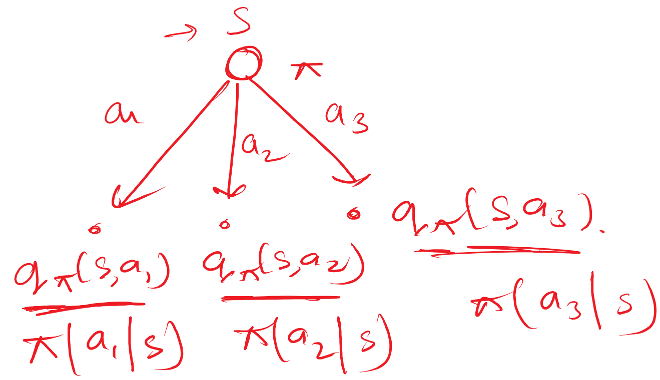
$$q_\pi(s, a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a] = \mathbb{E}_\pi[\sum_{k=0}^{T} \gamma^k R_{t+k+1} | S_t = s, A_t = a], \quad \forall\ s \in S, a \in A(s)$$

$\rightarrow S$

$Q \quad \pi$

v_π using q_π

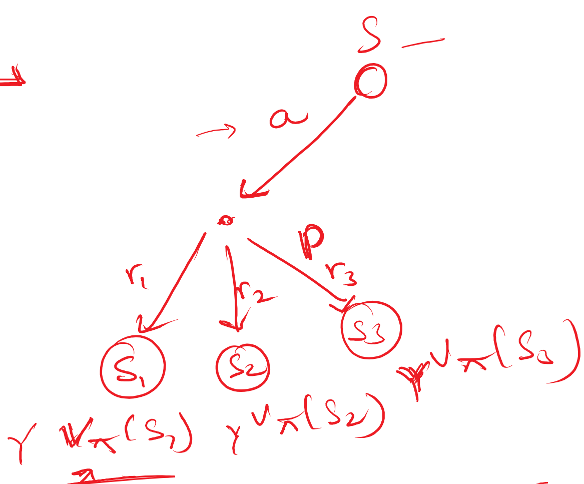
$v_\pi(s)$



$$v_\pi(s) = \pi(a_1|s)\cdot q_\pi(s,a_1) + \pi(a_2|s)\cdot q_\pi(s,a_2) + \pi(a_3|s)\cdot q_\pi(s,a_3).$$

$$v_\pi(s) = \sum_a \pi(a|s)\, q_\pi(s,a).$$

q_π from v_π



$$q_\pi(s,a) = p(s_1,r_1|s,a)\left[r + \gamma\, v_\pi(s_1)\right] + p(s_2,r_2|s,a)\left[r + \gamma\, v_\pi(s_2)\right] + p(s_3,r_3|s,a)\left[r + \gamma\, v_\pi(s_3)\right]$$

$$q_\pi(s,a) = \sum_{s',r} p(s',r|s,a)\left[r + \gamma\, v_\pi(s')\right]$$

Reinforcement Learning

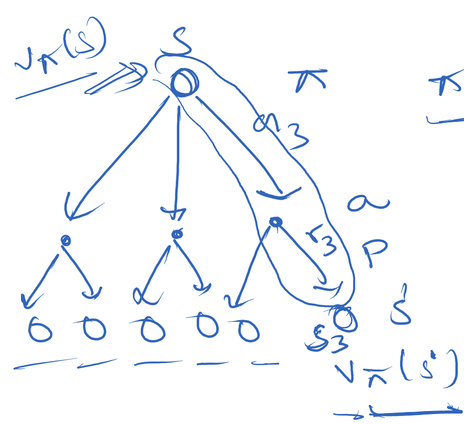
Bellman's Equations

Bellman's Equation for state values

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

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi[G_t | S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} | S_t = s] + \gamma \mathbb{E}_\pi[G_{t+1} | S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} | S_t = s] + \gamma \mathbb{E}_\pi[\mathbb{E}_\pi[G_{t+1} | S_{t+1} = s'] | S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} | S_t = s] + \gamma \mathbb{E}_\pi[v_\pi(S_{t+1} = s') | S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1} = s') | S_t = s] \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s', r | s, a)[r + \gamma v_\pi(s')]
\end{aligned}
$$

$$v_\pi(s') = E_\pi[G_{t+1} | S_{t+1} = s']$$



$$\pi(a_3|s) \cdot p(r_3, s_3 | s, a)$$

$$[r_3 + \gamma v_\pi(s')]$$

$= 0$

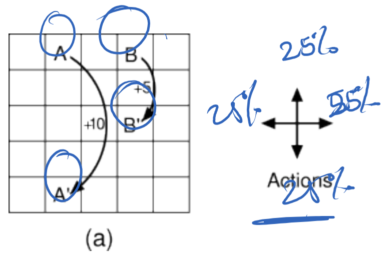
$\pi(a|s) = \frac{1}{\ }$

0

-1

$\pi(\cdot|s) = \frac{1}{4}$

$\gamma = 0.9$



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)

$$\frac{1}{4} \times 1\,[0 + 0.9\,(2.3)] + \frac{1}{4} \times 1\,[0 + 0.9\,(0.7)] + \cancel{\frac{1}{4}\,[0 + 0.9 \times (0.4)]} + \cancel{\frac{1}{4}\,[0 + 0.9 \times (0.4)]}$$

$$= 0.67 \sim 0.7$$

Bellman's Equation for action values

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

$$= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a]$$

$$= \mathbb{E}_\pi[R_{t+1} | S_t = s, A_t = a] + \gamma \mathbb{E}_\pi[G_{t+1} | S_t = s, A_t = a]$$

$$= \mathbb{E}_\pi[R_{t+1} | S_t = s, A_t = a] + \gamma \mathbb{E}_\pi[\mathbb{E}_\pi[G_{t+1} | S_{t+1} = s'] | S_t = s, A_t = a]$$

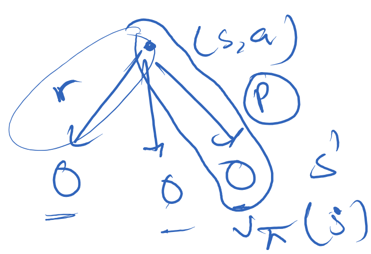
$$= \mathbb{E}_\pi[R_{t+1} | S_t = s, A_t = a] + \gamma \mathbb{E}_\pi[v_\pi(S_{t+1} = s') | S_t = s, A_t = a]$$

$$= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1} = s') | S_t = s, A_t = a]$$

$$= \sum_{s',r} p(s', r | s, a)[r + \gamma v_\pi(s')]$$

$$= \sum_{s',r} p(s', r | s, a)[r + \gamma \textstyle\sum_{a'} \pi(a'|s') q_\pi(s', a')]$$

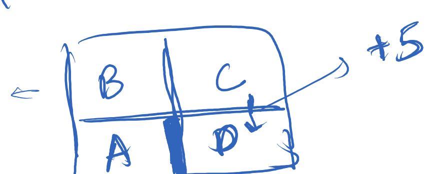
$$v_\pi(s) = \sum_a \pi(a|s)\, q_\pi(s,a)$$



$$p(s', r | s, a)\,[r + \gamma\, v_\pi(s')]$$

$\pi(\cdot|s) = 0.25$

$\gamma = 0.7$



$$V(A) = \frac{1}{4} \times 1\,[0 + 0.7 \cdot V(B)] + \frac{3}{4} \times 1\,[0 + 0.7\, V(A)]$$

$$V(B) = \frac{1}{4} \times 1\,[0 + 0.7 \cdot V(A)] + \frac{1}{4} \times 1\,[0 + 0.7\, V(C)] + 2 \times 1\,[0 + 0.7\, V(B)]$$

$$V(B) = \frac{1}{4}\times 1\left[0 + 0.7\cdot V(A)\right] + \frac{1}{4} \ldots$$

$$+ \frac{2}{4}\times 1\left[0 + 0.7\, V(B)\right]$$

$$V(C) = \frac{1}{4}\times 1\left[0 + 0.7\times V(B)\right] + \frac{1}{4}\times 1\left[5 + 0.7\, V(D)\right]$$

$$+ \frac{2}{4}\times 1\left[0 + 0.7\times V(C)\right]$$

$$V(D) = 0$$

Optimality

Sunday, February 16, 2020 11:41 PM

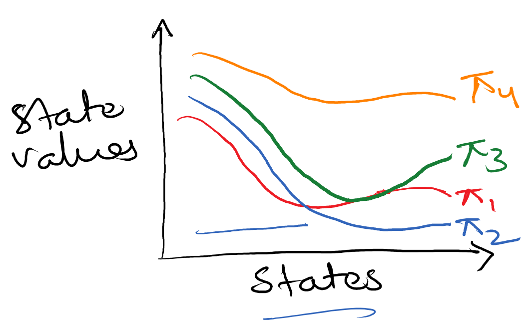
Reinforcement Learning

Optimality

Optimal Policy

$\pi_1 \quad \pi_2$

$\pi_1 \geq \pi_2$ iff $v_{\pi_1}(s) \geq v_{\pi_2}(s) \ \forall s \in S$



$\pi_2 \ngeq \pi_1$

$\pi_1 \ngeq \pi_2$

$\pi_3 \geq \pi_1$

$\pi_3 \geq \pi_2$

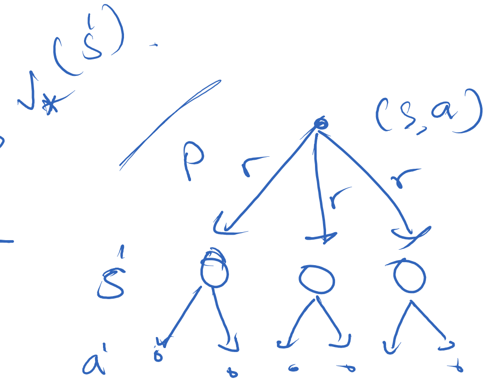
$\pi_4 \geq \pi_1, \pi_2, \pi_3$

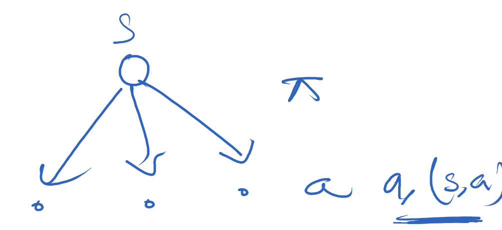
$v_* = \max_\pi v_\pi(s) \ \forall s \qquad \pi_*$

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

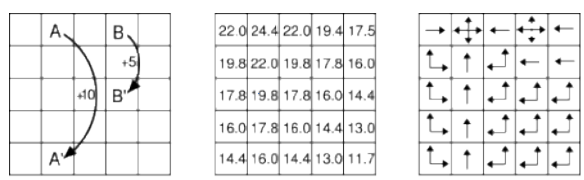
Bellman optimality equations

$$v_*(s) = \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma v_*(s')]$$
$$q_*(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma \max_{a'} q_*(s',a')]$$





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



a) gridworld b) V^* c) π^*