1. Copying from Wikipedia.

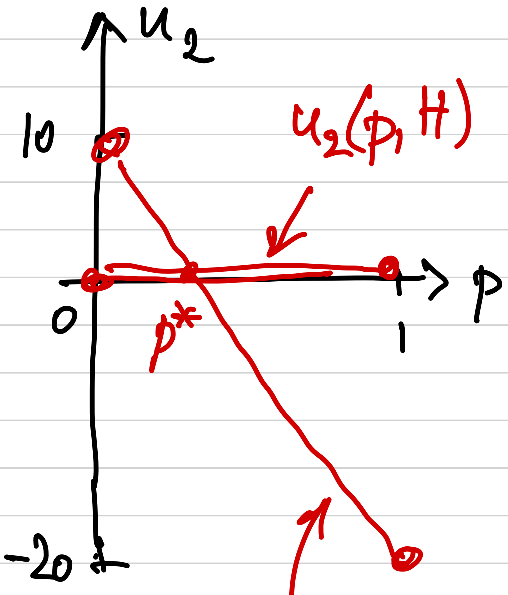
		Student H q	W $(1-q)$
TA	C p	$5, \underline{0}$	$\underline{7}, -20$
	N $(1-p)$	$\underline{10}, 0$	$2, \underline{10}$

C = check
N = no check
H = honest
W = Wikipedia

No pure NE!

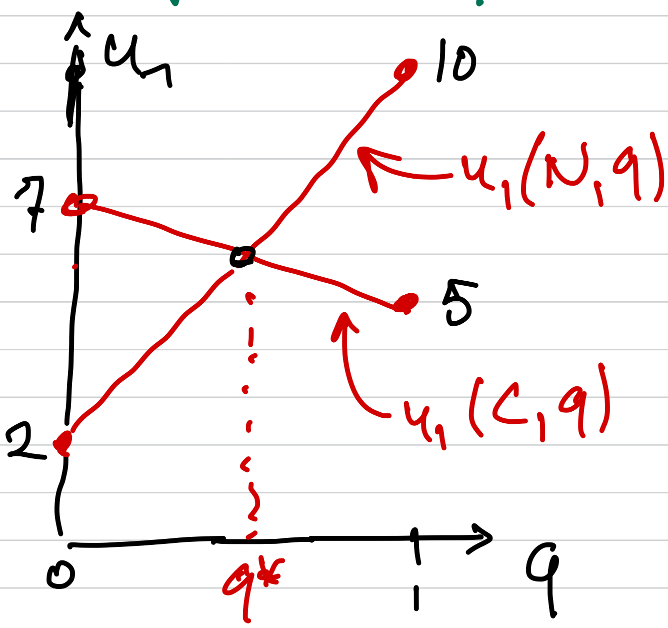
Mixed NE

parameter p



$p^* = \frac{1}{3}$

parameter q



$$\left.\begin{aligned} u_1(N, q) &= 2(1-q) + 10q \\ &= 8q - 2 \\ u_1(C, q) &= 7(1-q) + 5q \\ &= -2q + 7 \end{aligned}\right\} q^* = \tfrac{1}{2}$$

② utilities for NE

$$u_1(p^*, q^*) = \frac{1}{6}\cdot 5 + \frac{2}{6}\cdot 10 + \frac{1}{6}\cdot 7 + \frac{2}{6}\cdot 2$$

$$u_2(p^*, q^*) = \frac{1}{6}\cdot 0 + \frac{2}{6}\cdot 0 + \frac{1}{6}(-20) + \frac{2}{6}\cdot 10$$

5, 0	7, −20
10, 0	2, 10

p^* \	1/2	1/2
1/3	1/6	1/6
2/3	2/6	2/6

$p^* = \frac{1}{3}$, $q^* = \frac{1}{2}$

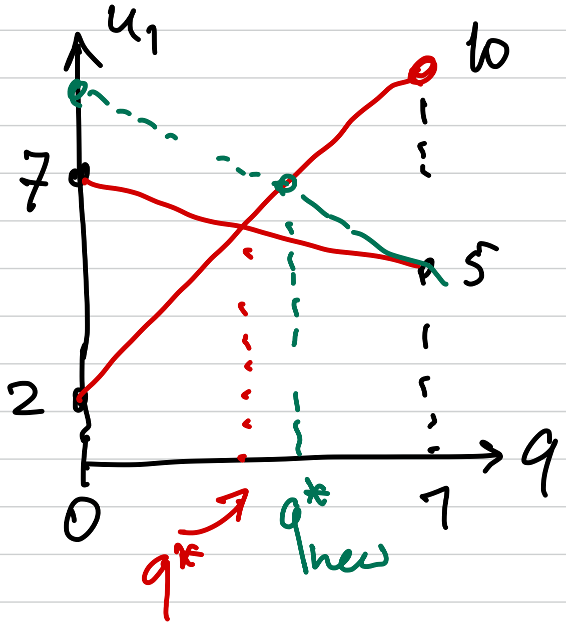
$$u_1(p^*, q^*) = \frac{1}{6}\cdot(5+20+7+4) = \frac{36}{6} = 6.$$

$$u_2(p^*, q^*) = \frac{1}{6}(0+0-20+20) = 0$$

(Compare to fig on previous page)

3/ What can be done to increase prob that student will be honest.

prob = q
(see mixing parameters).



Currently TA gets reward 7 if he finds a cheating student.
If we increase that reward : $q^* \uparrow$.

$$q^*_{new} > q^*$$

2. MDP 1

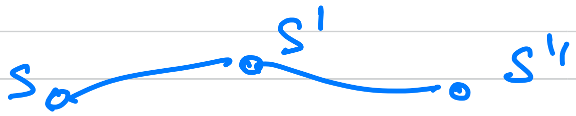
① $P_\pi(s, s') = \sum_a \pi(a|s)\, p(s'|sa)$

prob that agent will transition $s \to s'$ in one step (under policy π)

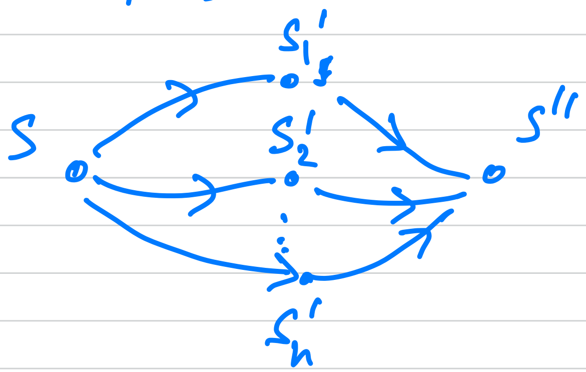
$r_\pi(s)$ = expected immediate reward in state s under policy π.

② (i) $P(s, s')\, P(s', s'')$

prob that you will transition from $s \to s'$ and then $s' \to s''$

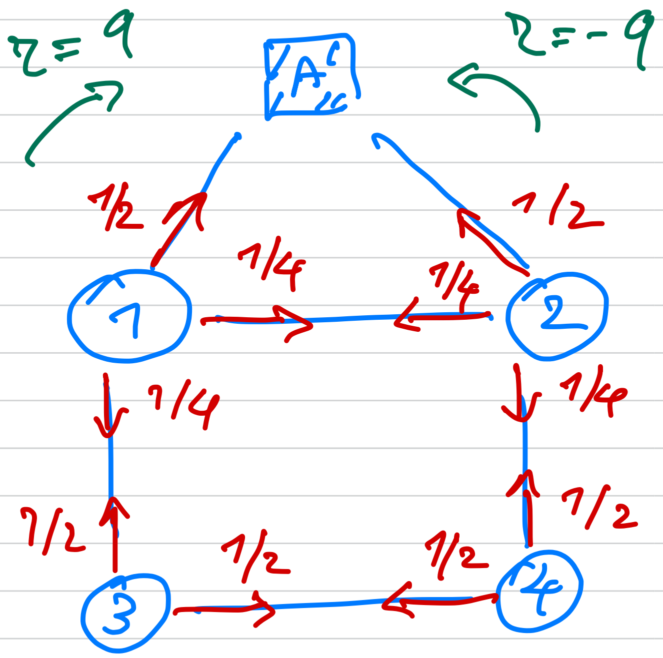


(ii) $P^2(s, s'')$



Prob of transitioning from $s \to s''$ in 2 steps along any possible path.

③



to s'

P

from $s=$	A	1	2	3	4
A	1	0	0	0	0
1	1/2	0	1/4	1/4	0
2	1/2	1/4	0	0	1/4
3	0	1/2	0	0	1/2
4	0	0	1/2	1/2	0

$r(A) = 0$ absorbing state.

$r(1) = \frac{1}{2} \cdot 9 + \frac{1}{4}(-1) + \frac{1}{4}(-1) = \frac{1}{2} 9 + \frac{1}{2}(-1) = 4.$

$r(2) = \frac{1}{2} \cdot (-9) + \frac{1}{2}(-1) = -5.$

$r(3) = -1$, $r(4) = -1.$

④ If γ small $\rightarrow \gamma^2 \simeq 0$.

Bellman eq. $\quad v = \gamma P v + r$

$$v = \gamma P(\gamma P v + r) + r$$
$$= \underbrace{\gamma^2}_{\simeq 0} P^2 v + \gamma P r + r.$$

$$= \gamma P r + r$$

(P, r are known, see previous page).

<u>Alternatively</u>:

$$v = \gamma P v + r \Rightarrow (I - \gamma P) v = r$$
$$\Rightarrow v = (I - \gamma P)^{-1} r$$
$$= (I + \gamma P + \underbrace{\gamma^2 P^2 + \dots}_{\simeq 0}) r$$
$$= r + \gamma P r.$$

3. Optimal v^* assuming $\gamma = 1/3$.

Solution: move to A via 1 as fast as possible.

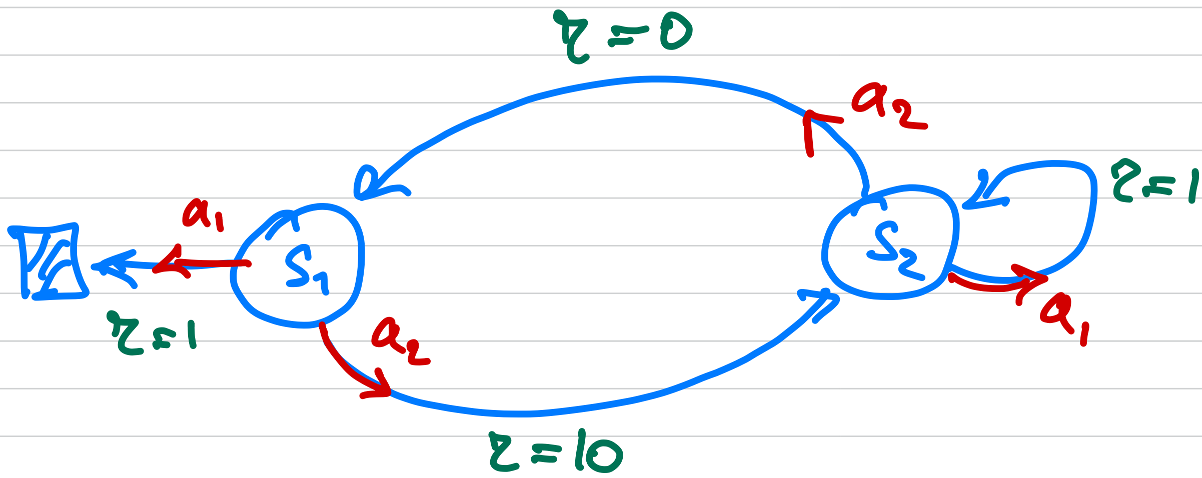
$v^*(1) = 9$

$v^*(2) = r + \gamma v^*(1) = -1 + 1/3 \cdot 9 = 2$

$v^*(3) = -1 + \gamma v^*(1) = 2$

$v^*(4) = -1 + \gamma v^*(3) = -1 + \frac{1}{3} \cdot 2 = -\frac{1}{3}$

Question 3 : MDP 2



① By picking action a_2 in both states the agent would loop forever.
→ infinite horizon

② $\gamma = 0.9$ → optimal policy π^* ?

There are 4 possible (deterministic) policies.

$$\begin{cases} \pi_1 : s_1 \to a_1 , \; s_2 \to a_1 \\ \pi_2 : s_1 \to a_1 , \; s_2 \to a_2 \\ \pi_3 : s_1 \to a_2 , \; s_2 \to a_1 \\ \pi_4 : s_1 \to a_2 , \; s_2 \to a_2 . \end{cases}$$

Next we compute the value f^{π} for each policy.

notation
$v_{\pi_1} = v_1$

$\rightarrow \pi_1 : s_1 \rightarrow a_1 , \; s_2 \rightarrow a_1$

$v_1(s_1) = 1$ (go to absorbing node)

$v_1(s_2) = 1 + \gamma + \gamma^2 + \dots = \frac{1}{1-\gamma}$ (infinite loop)

$\rightarrow \pi_2 : s_1 \rightarrow a_1 , \; s_2 \rightarrow a_2$

$v_2(s_1) = 1$

$v_2(s_2) = 0 + \gamma \cdot 1 = \gamma$

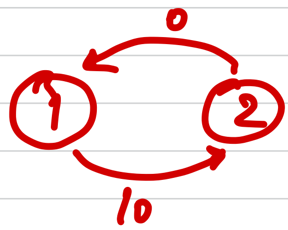
$\rightarrow \pi_3 : s_1 \rightarrow a_2 , \; s_2 \rightarrow a_1$

$v_3(s_1) = 10 + \gamma + \gamma^2 + \dots$

$= 10 + \gamma(1 + \gamma + \gamma^2 + \dots) = 10 + \frac{\gamma}{1-\gamma}$

$v_3(s_2) = 1 + \gamma + \gamma^2 + \dots = \frac{1}{1-\gamma}$

$\rightarrow \pi_4 : s_1 \rightarrow a_2 , \; s_2 \rightarrow a_2$



$v_4(s_1) = 10 + 0.\gamma + 10\gamma^2 + 0.\gamma^3 + \dots$

$= 10 + 10\gamma^2(1 + \gamma^2 + \dots)$

$= 10 + 10\gamma^2 \frac{1}{1-\gamma^2} = 10\left(1 + \frac{\gamma^2}{1-\gamma^2}\right)$

$= 10\left(\frac{1}{1-\gamma^2}\right) = \frac{10}{1-\gamma^2}$

$v_4(s_2) = 0 + 10.\gamma + 0.\gamma^2 + 10\gamma^3 + \dots$

$= 10\gamma(1 + \gamma^2 + \gamma^4 + \dots) = \frac{10\gamma}{1-\gamma^2}$

Overview

	π_1	π_2	π_3	π_4
$v(s_1)$	1	1	$10 + \frac{\gamma}{1-\gamma}$	$\frac{10}{1-\gamma^2}$
$v(s_2)$	$\frac{1}{1-\gamma}$	γ	$\frac{1}{1-\gamma}$	$\frac{10\gamma}{1-\gamma^2}$

$\gamma = 0.9$ $\rightarrow 1-\gamma = 0.1 \rightarrow \frac{1}{1-\gamma} = 10, \ \frac{1}{1-\gamma^2} = \frac{1}{0.19}$

$$\rightarrow \frac{1}{1-\gamma^2} = \frac{1}{0.19} = 5.26.$$

$$\rightarrow \frac{\gamma}{1-\gamma^2} = 4.74$$

	π_1	π_2	π_3	π_4
$v(s_1)$	1	1	19	52.6
$v(s_2)$	10	0.9	10	47.4

optimal policy!

3. Changes in $\gamma \longrightarrow$ change in π^* ??

Consider: $\gamma = 0$

	π_1	π_2	π_3	π_4
$v(s_1)$	1	1	10	10
$v(s_2)$	1	0	1	0

$\rightarrow$ optimal

Conclusion:

$\gamma = 0$ (small) $\longrightarrow$ π_3 optimal

$\gamma \simeq 1$ $\longrightarrow$ π_4 optimal.

4. Vickrey auction

$\longrightarrow$ See Theory

5. Q-learning & SARSA

① $v_\pi(2) = \sum_a \pi(a|s)\, q_\pi(s,a)$ $\quad a = L, R$

$$= \pi(L|2)\, q_\pi(2,L) + \pi(R|2)\, q_\pi(2,R)$$

$$= \frac{3}{4}\cdot 4 + \frac{1}{4}\cdot 5 = \frac{17}{4}$$

$$v_\pi(3) = \frac{2}{3}\cdot 6 + \frac{1}{3}\cdot 8 = \frac{12+8}{3} = \frac{20}{3}$$

② Q-learning update. $2 \xrightarrow{R} 3$

$$\underbrace{q_\pi(2,R)}_{\text{new}} \leftarrow \underbrace{q_\pi(2,R)}_{\text{old}} + \underline{\underline{\alpha}}\left[r + \gamma \max_a q_\pi(3,a) - \underbrace{q_\pi(2,R)}_{\text{old}}\right]$$

$$= 5 + 0.9\left[-1 + \frac{2}{3}\underbrace{\max(6,8)}_{8} - 5\right]$$

$$= 5 + 0.9\left[-1 + \frac{16}{3} - 5\right]$$

$$= 5 + 0.9\left(-\frac{2}{3}\right) = 5 - 0.6 = 4.4.$$

③ Expected Sarsa:

The only thing that changes is that we take a weighted mean rather than max.

$$\frac{2}{3}\cdot 6 + \frac{1}{3}\cdot 8 = \frac{20}{3}$$