MAS 2022-23 HW 5 MDP Solutions v2

Multi-agent systems (Vrije Universiteit Amsterdam)

HW5

HW5

1

Bellman equations for deterministic policy ~~[illegible]~~

General

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

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

① Deterministic policy

Under the policy π, each state is mapped to unique action a_s

$$\pi : s \longmapsto a_s$$

Hence, summation over action collapses in singleton.

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

$q_\pi(s,a)$ = value when taking action a in state s (action a is arbitrary, not necessarily dictated by policy!) and THEN following policy π!

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

Notice: $v_\pi(s) = \sum_a \pi(a|s)\, q(s,a) = q_\pi(s,a_s)$

② Deterministic policy and Transition.

We now have the following deterministic mappings:

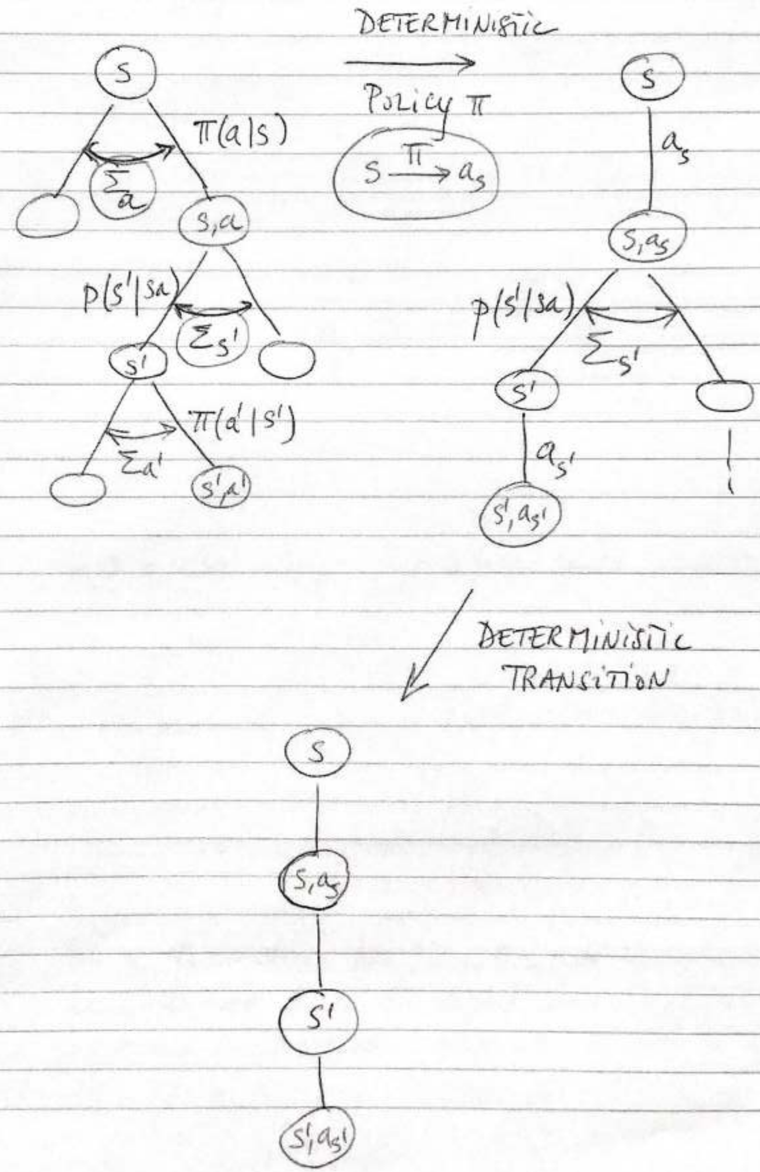
$$s \xmapsto{\pi} a_s \quad \text{unique}$$

$$s \xmapsto{a_s} s_{a_s} \quad \text{unique.}$$

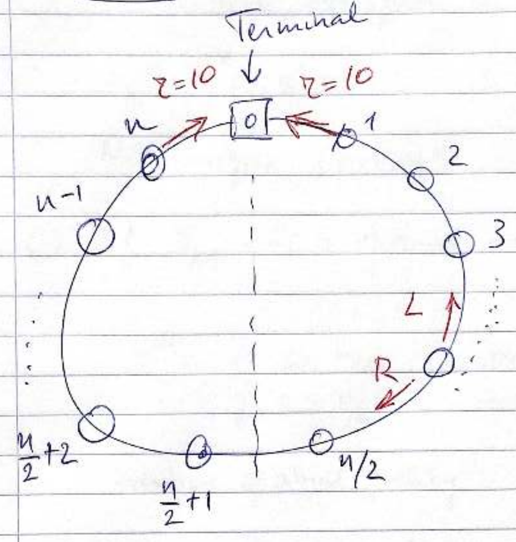
$$v_\pi(s) = r(s, a_s, s_{a_s}) + \gamma v(s_{a_s})$$

$$q_\pi(s, a) = r(s, a, s_a) + \gamma q(s_a, a_{s_a})$$

Backup diagrams



MDP 1 Circular state space



Non-terminal transitions yield reward $r_{NT} = 0$.

n = even
$\Downarrow$
$\frac{n}{2}$ integer.

Node 0 = absorbing : transition yields $r = 10$.

$\gamma = 1$ (no discounting)

π = equiprobable policy : each action has prob ½
two actions in each node
move clockwise (R)
move counterclockwise (L)

① Since transitions btw non-terminal states incur no cost and the agent will eventually end up in absorbing state 0 we conclude:

$$v_\pi(s) = 10 \quad \forall s.$$

$$q_\pi(s,a) = 10 \quad \forall s, a.$$

② Optimal policy: any policy that ensures eventual absorption in state 0.

$q^*(s,a) = 10 \qquad v^*(s) = 10.$

<u>Not</u> unique. (policy).

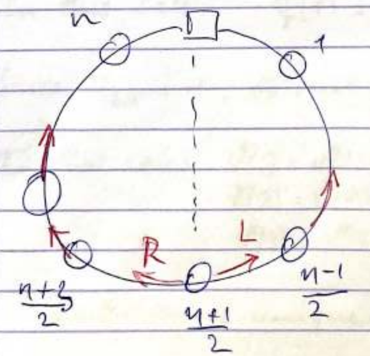
③ If $r_{NT} = -1$: optimal → go to terminal state <u>asap</u>.

π^*: if $s \leq n/2$: action: go L with prob = 1
if $s \geq n/2 + 1$ — go R —

Unique optimal policy.

④ If $\gamma < 1$: go to terminal asap to be optimal.
(similar to 3).

⑤ If n = odd ($r_{NT} = -1$, $\gamma = 1$)



Optimal policy is NO LONGER unique!

if : $s \leq \frac{n-1}{2}$ → go L.

if $s > \frac{n+1}{2}$ → go R

if $s = \frac{n+1}{2}$ one can choose btw L and R.

HW5

6

Markov decision process (MDP) MDP2

1. Equi-prob. policy π

$$v_\pi(2) = v_\pi(5) = 10 \quad \text{because of symmetry.}$$

Equal prob to end up in A (reward 20) and B (reward 0).

The other values can be computed using the Bellman eq:

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

$$v(1) = \frac{1}{2}(20 + 10) = 15 \quad ; \quad v(3) = \frac{1}{2}(10 + 0) = 5$$

$$v(6) = v(1) \quad , \quad v(5) = v(2) \quad , \quad v(4) = v(3)$$

2. An optimal policy is <u>any</u> policy that avoids absorption by B.
So not unique.
In this case: $v_\pi^*(1) = \dots = v^*(6) = 20$.

3. Since $r_{NT} = -1$, optimal policy = "go to A as fast as you can!".

In that case:
$$v^*(1) = v^*(6) = 20$$
$$v^*(2) = v^*(5) = 19$$
$$v^*(3) = v^*(4) = 18$$

This policy is unique.

6.3 Continued

7 ~~8~~

4 $r_{NT} = -10.$

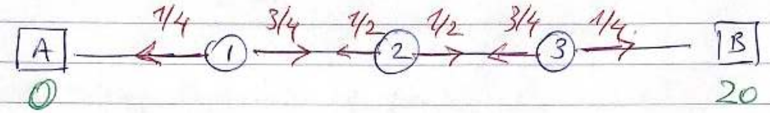
Optimal values
$$v^*(1) = v^*(6) = 20$$
$$v^*(2) = v^*(5) = 10$$
$$v^*(3) = v^*(4) = 0$$

Policy is NOT unique since in 3 and 4 it does not matter which direction you choose.

8

HW5, question 4 : MDP3

① State value f$^{\text{ion}}$ $v_\pi(s)$ under given policy π



Since the transitions are deterministic we can simplify the Bellman eq:

$$s \xrightarrow{a} s_a$$

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

$$= \sum_a \pi(a|s)\left[r(s,a,s_a) + \gamma v(s_a)\right]$$

Denote, $v_\pi(1) = v_1$, $v_\pi(2) = v_2$, $v_\pi(3) = v_3$

Notice: $v_\pi(A) = 0 = v_\pi(B)$,

Then:

$$v_1 = \frac{1}{4}(0+0) + \frac{3}{4}(-2+v_2) = -\frac{3}{2} + \frac{3}{4}v_2$$

$$v_2 = \frac{1}{2}(-2+v_1) + \frac{1}{2}(-2+v_3) = \frac{v_1+v_3}{2} - 2$$

$$v_3 = \frac{3}{4}(-2+v_2) + \frac{1}{4}(20+0) = \frac{3}{4}v_2 + 5 - \frac{3}{2}$$

$$= \frac{3}{4}v_2 + \frac{7}{2}$$

Summing v_1 and v_3:

$$v_1 + v_3 = \left(-\frac{3}{2} + \frac{3}{4} v_2\right) + \left(\frac{3}{4} v_2 + \frac{7}{2}\right)$$
$$= \frac{3}{2} v_2 + 2$$

Substituting this into eq. for v_2:

$$v_2 = \frac{1}{2}(v_1 + v_3) - 2 = \frac{1}{2}\left(\frac{3}{2} v_2 + 2\right) - 2$$
$$= \frac{3}{4} v_2 - 1$$

$$\Rightarrow \boxed{v_2 = -4}$$

$$\Rightarrow \boxed{v_1 = -\frac{3}{2} + \frac{3}{4} v_2 = -\frac{3}{2} + \frac{3}{4}(-4) = -\frac{9}{2}}$$

$$\Rightarrow \boxed{v_3 = \frac{3}{4} v_2 + \frac{7}{2} = \frac{3}{4}(-4) + \frac{7}{2} = \frac{1}{2}}$$

(2) Compute state-action value $q_\pi(2,R)$ and $q_\pi(3,L)$

$$q_\pi(2,R) = -2 + v_\pi(3) = -2 + \frac{1}{2} = -3/2$$

$$q_\pi(3,L) = -2 + v_\pi(2) = -2 + \left(-\frac{9}{2}\right) = -\frac{13}{2}$$

(3) Optimal policy: go R in each state.
Unique!