Two-arm bandits

Two actions ; A and B.

Action A: always reward (+6)

Action B:

if B is of type lucky ($L = l$) or unlucky ($L = \neg l$)

B is lucky CPT
$p(r=+10 \mid L=l) = 4/5$
$p(r=0 \mid L=l) = 1/5$

B is unlucky CPT
$p(r=+10 \mid L=\neg l) = 1/5$
$p(r=0 \mid L=\neg l) = 4/5$

Discount factor $\gamma = 1$

A-priori $p(L=l) = p(L=\neg l) = 0.5$

(a) One step time horizon:

MEU for action A: $\boxed{+6}$ [There is no uncertainty in lever A]

MEU for action B : $\underbrace{+10}_{\text{reward}} \times \left(\underbrace{\frac{1}{2}}_{\text{lucky}} \times \underbrace{\frac{4}{5}}_{p(r=10 \mid L=l)} + \underbrace{\frac{1}{2}}_{\text{unlucky}} \times \underbrace{\frac{1}{5}}_{p(r=+10 \mid L=\neg l)} \right)$

$$+ \ 0 \times \left(\frac{1}{2} \times \frac{1}{5} + \frac{1}{2} \times \frac{4}{5} \right)$$

$$= 10 \times \frac{1}{2} = \boxed{+5}$$

$MEU(a) > MEU(b) \Rightarrow$ Action $\underline{\underline{A}}$ should be taken.

) Two-steps in the future

- Note: There is inherent uncertainty in the type of lever B. As we plan ahead, the resulting rewards can change the ~~likeli~~ belief over the type of lever B. The reward can be treated as an observation that updates the likelihood.

MDP formulation suggested:

State (m, n) where m denotes the # of times lever B was pulled & reward of (+10) was received

n denotes the # of times lever B was pulled & reward of (0) was received.

Example, $(0,0) \xrightarrow[r=+10]{B \text{ is pulled}} (1,0)$

$(0,0) \xrightarrow[r \text{ is } 0]{B \text{ is pulled}} (0,1)$

(ii) In state $(0,1)$, take action B, can land up in

- State $(1,1)$ if the reward is +10 with probability ~~$p(L=l)\,p(r=+10 \mid L=l) + p(L=\neg l)\,p(r=+10 \mid L=\neg l)$~~

$$= p(L=l \mid r=0)\, p(r=+10 \mid L=l) + p(L=\neg l \mid r=0) \cdot p(r=+10 \mid L=\neg l)$$

Note: previously reward was 0 in the first pull, this obs^n will update the bel over the type of lever B is.

$$p(L=l \mid r=0) = \frac{p(r=0 \mid L=l)\, p(L=l)}{p(r=0 \mid L=l)\, p(L=l) + p(r=0 \mid L=\neg l)\, p(L=\neg l)}$$

$$= \left[\frac{1/5 \times \frac{1}{2}}{(1/5 \times 1/2) + (4/5 \times \frac{1}{2})}\right] = \left[\frac{1}{5}\right]$$

$$p(L=\neg l \mid r=0) = 1 - 1/5 = 4/5$$

w, from state $(0,1)$ taking action B can be in

- State $(1,1)$ if reward is $+10$ with probability

$$p(L=l \mid r=0)\, p(r=+10 \mid L=l) + p(L=\neg l \mid r=0)\, p(r=+10 \mid L=\neg l)$$

$$= \frac{1}{5} \times \frac{4}{5} + \frac{4}{5} \times \frac{1}{5} = \frac{8}{25}$$

- State $(0,2)$ if reward is $(+0)$ with probability

$$p(L=l \mid r=0)\, p(r=0 \mid L=l) + p(L=\neg l \mid r=0)\, p(r=0 \mid L=\neg l)$$

$$= \frac{1}{5} \times \frac{1}{5} + \frac{4}{5} \times \frac{4}{5} = \frac{17}{25}$$

(ii) Now, from state $(1,0)$, taking action B can be in

- State $(2,0)$ if reward is $+10$ with probability

$$p(L=l \mid r=+10)\, p(r=+10 \mid L=l) + p(L=\neg l \mid r=+10)\, p(r=10 \mid L=\neg l)$$

We require $p(L=l \mid r=10) = \dfrac{p(r=10 \mid L=l)\, p(L=l)}{p(r=10 \mid L=l)\, p(L=l) + p(r=10 \mid L=\neg l)\, p(L=\neg l)}$

$$= \frac{\frac{4}{5} \times \frac{1}{2}}{\frac{4}{5} \times \frac{1}{2} + \frac{1}{5} \times \frac{1}{2}} = \frac{4}{5}$$

$$p(L=\neg l \mid r=+10) = \frac{1}{5}$$

New state probability is:

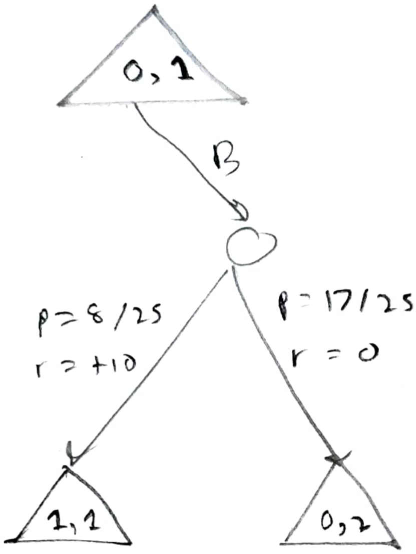
$$\frac{4}{5} \times \frac{4}{5} + \frac{1}{5} \times \frac{1}{5} = \frac{17}{25}$$

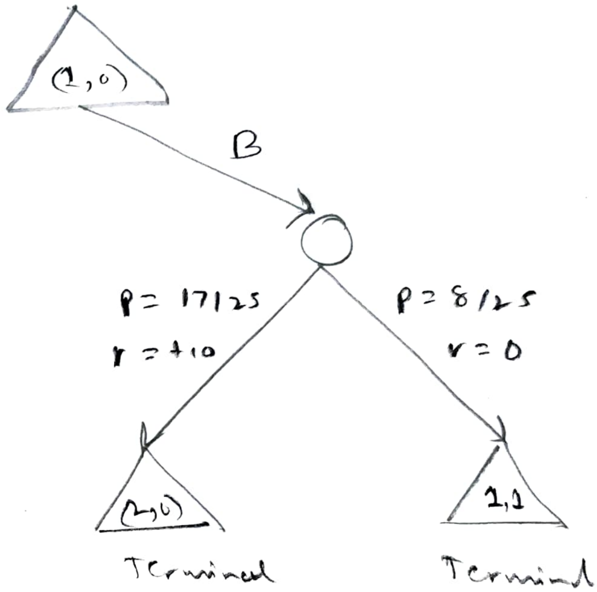
• State (1,1) if reward is +10 with probability

$$p(L=l \mid r=10) \times p(r=0 \mid L=l) + p(L=\neg l \mid r=+10) \times p(r=0 \mid L=\neg l)$$

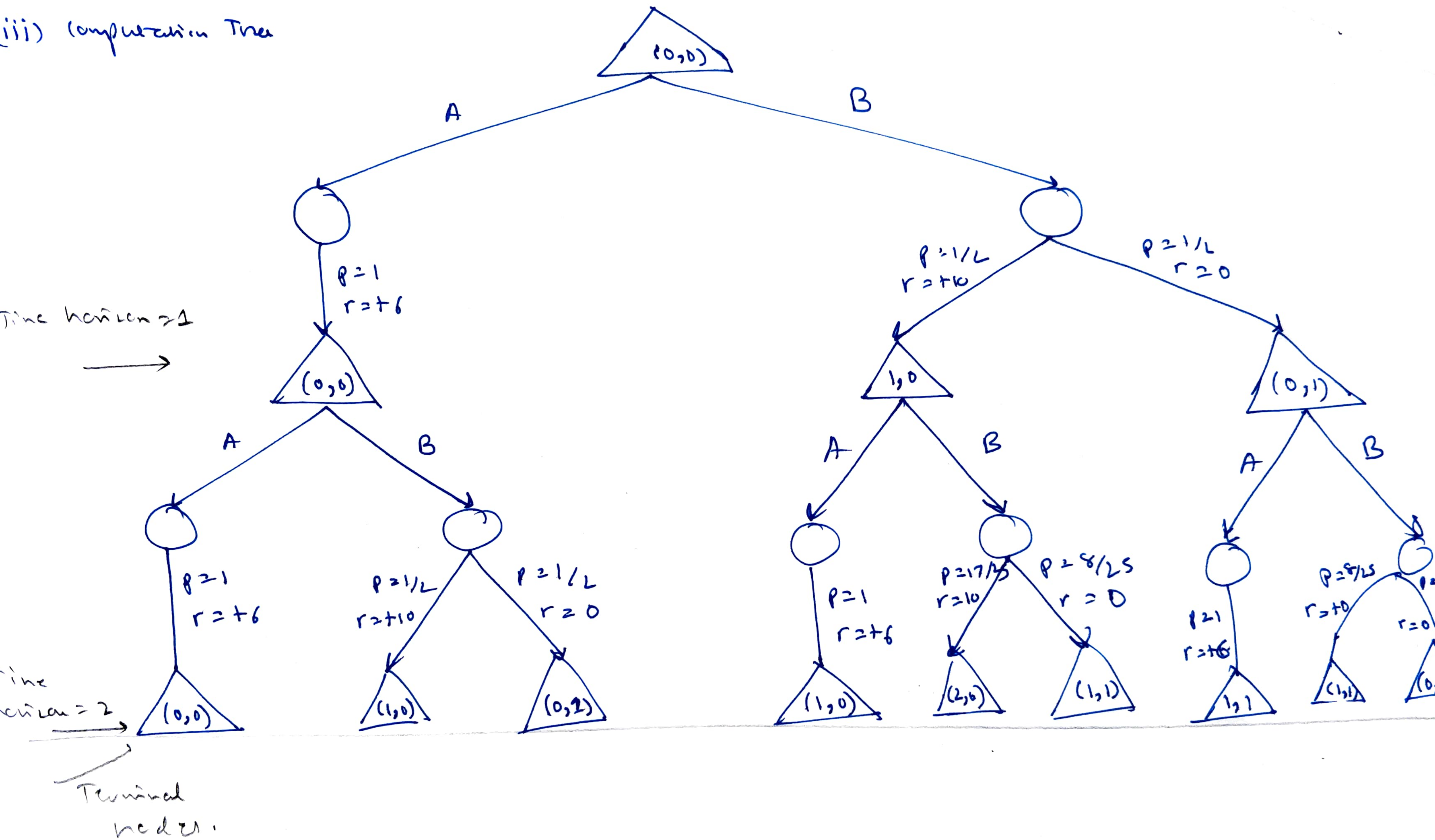
$$= \frac{4}{5} \times \frac{1}{5} + \frac{1}{5} \times \frac{4}{5} = \frac{8}{25}$$

These transitions as trees



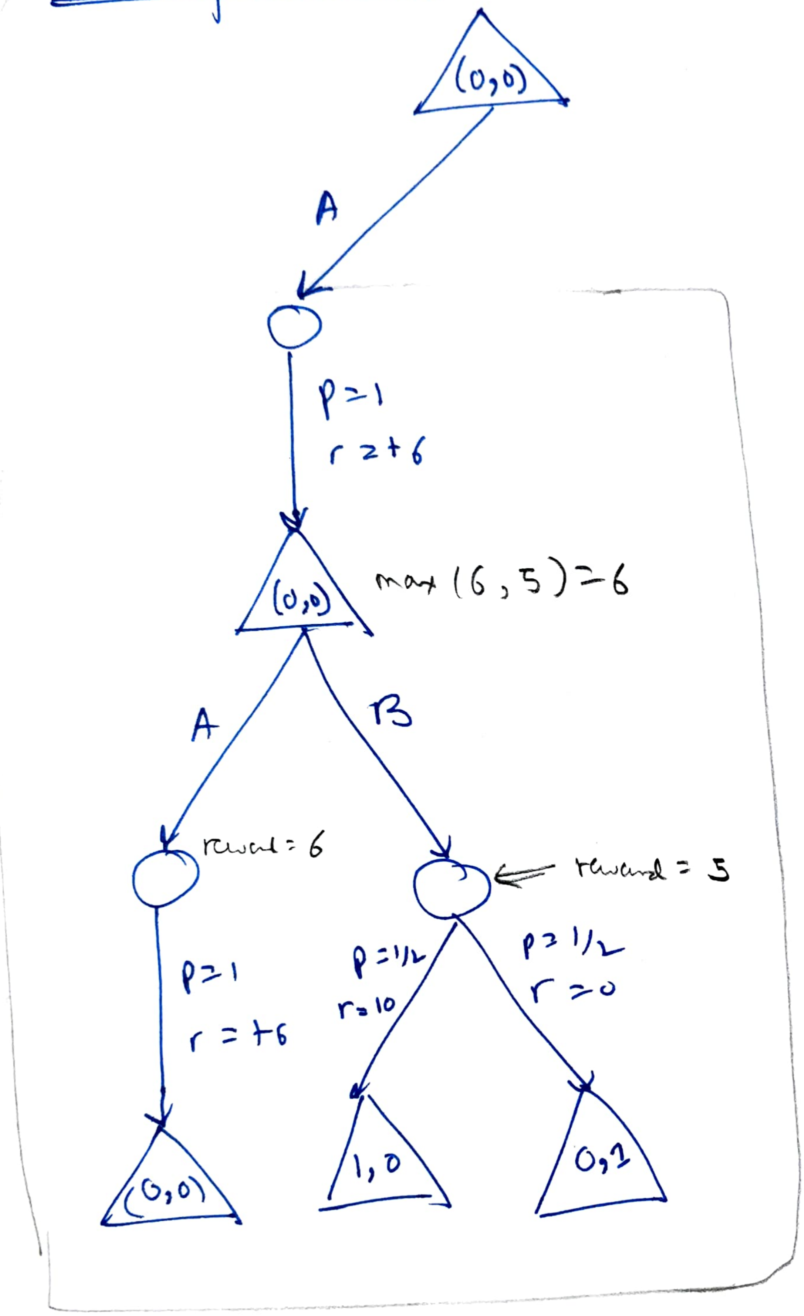


(iii) Computation Tree



) Which action is optimal A or B or both by MEU.

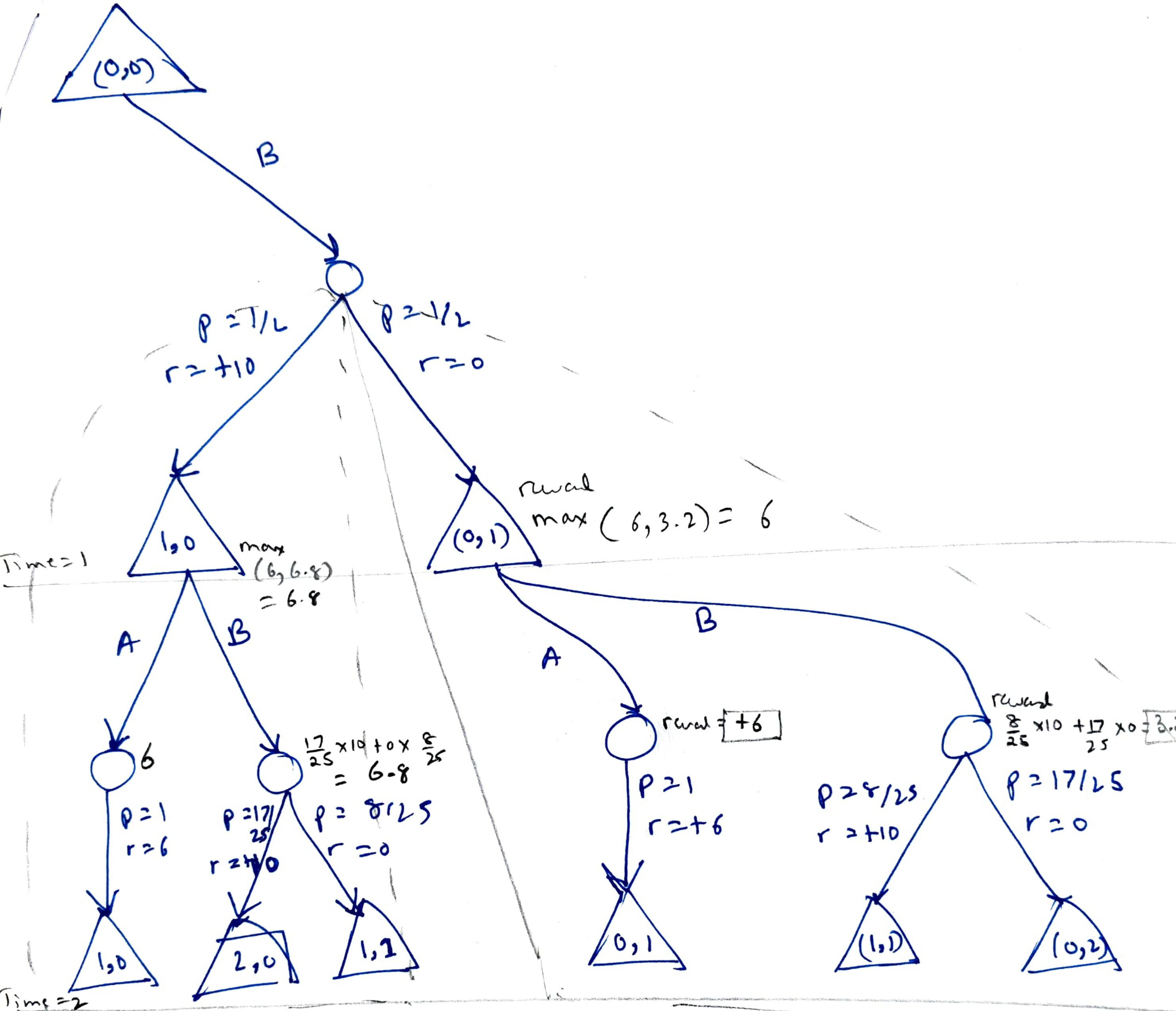
For left sub-tree



The total reward for the sub-tree after taking action A in state (0,0) is $6 + \gamma \cdot 6 = \boxed{12}$

as $\gamma = 1$

For the right sub-tree



Left

Reward = +10 + γ . 6.8

= 10 + 6.8

= $\boxed{16.8}$

Right

Reward = 0 + γ . 6 = $\boxed{6}$

as γ = 1

expected = $\frac{1}{2} \times 16.8 + \frac{1}{2} \times 6$

= 8.4 + 3 = $\boxed{11.4}$

MEU (A) > MEU(B) still action A is optimal.