4.4.

(1)(a) $Q(B, a_0) = 1 \cdot 1 = 1$

C will choose the greedy action with the probability:- $p = 1 - \varepsilon + \frac{\varepsilon}{N}$

where $\varepsilon = 0.9$ and $N = 2$

$\Rightarrow p = 0.95$; for the state C the greedy action will be a_0 since $n \in (-\infty, \infty)$

$\Rightarrow Q(C, a_0) = 0.95 \times 2 = 1.90$

$Q(C, a_1) = 0.05 \times n$

For A, the greedy action a_1 will be chosen with the probability $p = 0.95$

$\Rightarrow Q(A, a_0) = 0.05 \times \sum_a Q(B, a)$

$= 0.05 \times 1 = 0.05$ — (1)

$Q(A, a_1) = 0.95 \times \sum_a Q(C, a)$

$= 0.95 \times [1.9 + 0.05n]$

$= 1.805 + 0.0475n$ — (2)

(b) Using (1) & (2) from 4.4.1.(a) we can say that the final policy will choose action a_1 more than a_0 in the state A when :-

$Q(A, a_1) \geq Q(A, a_0)$

$\Rightarrow 1.805 + 0.0475n > 0.05$

$\Rightarrow n > -36.94$

for $n < -36.94$, it will choose action a_0.

(2) (a) $Q(B, a_0) = 1$

- $Q(C, a_0) = \begin{cases} 2 & n < 2 \\ 0 & \text{otherwise} \end{cases}$

$Q(C, a_1) = \begin{cases} 0 & \text{if } n < 2 \\ \text{otherwise} \end{cases}$

- ~~$Q(A, a_0) =$~~

Let's assume that we are not using the greedy policy for now. In that case

$Q(A, a_0) = 2$

$Q(A, a_1) = \begin{cases} 4 & n < 2 \\ n+2 & \text{otherwise} \end{cases}$

We can clearly see that $Q(A, a_1)$ is always greater than $Q(A, a_0)$. So in the presence of the greedy policy like in the case of Q-learning:

$Q(A, a_1) = \begin{cases} 4 & n < 2 \\ n+2 & \text{otherwise} \end{cases}$

$Q(A, a_0) = 0$.

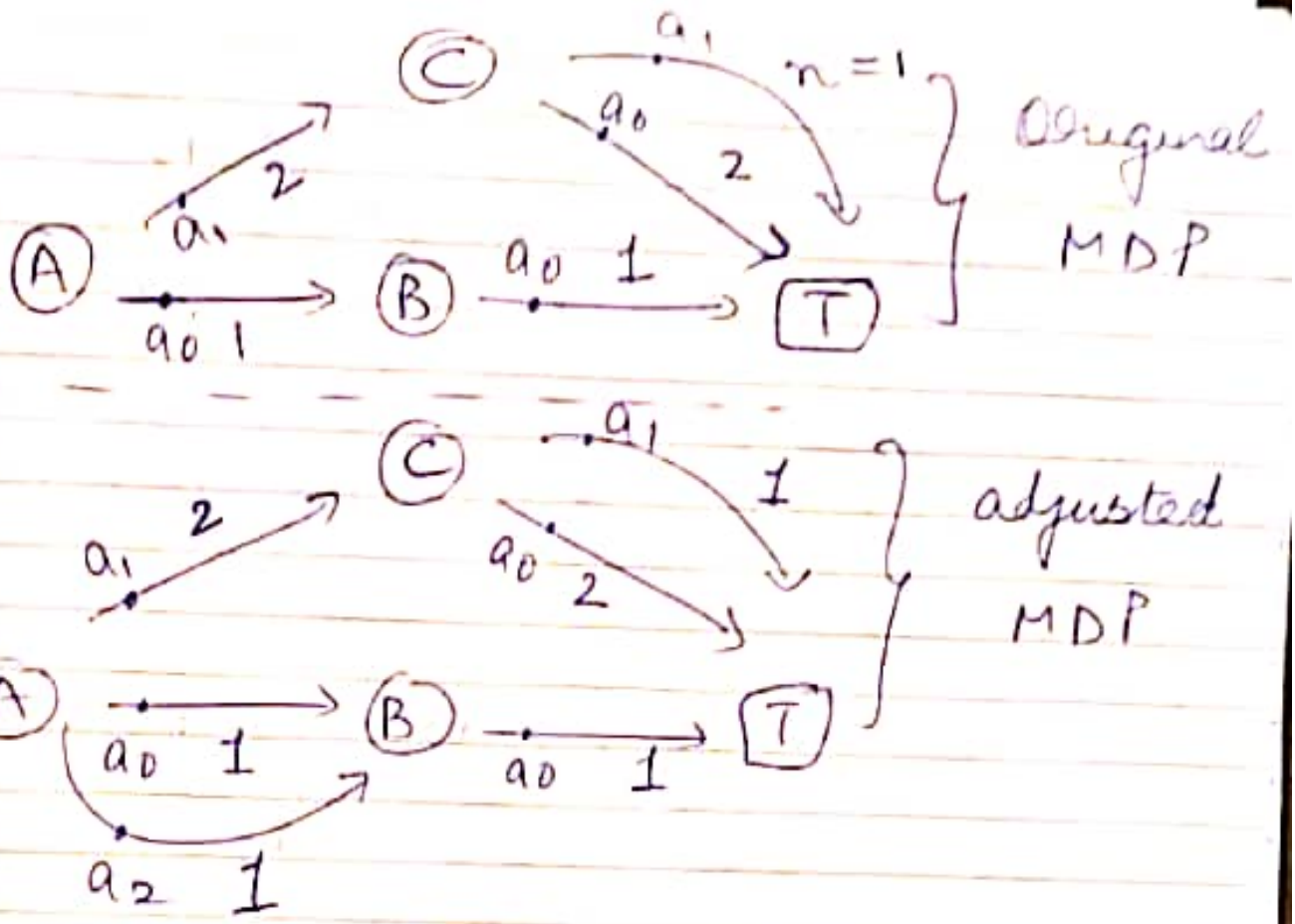
It wasn't clear from the question whether we should assume absolute greedy policy for Q learning or ε-greedy

Above we have showed the final Q-value if the ~~target~~ behaviour policy was absolute greedy.

If ε-greedy policy was to be used as the behaviour policy, then the final Q-values would be the same as in the case of SARSA.

(b) Since $Q(A, a_1)$ is always greater than $Q(A, a_0)$, the final policy will always choose a_1 when in the state A.

(3)



The only difference between the two MDPs is an additional action available for the state A in the adjusted MDP.

This will change the probability of choosing the greedy action when in state A for the adjusted MDP. It will be:

$$p = 1 - \varepsilon + \frac{\varepsilon}{N} \quad \text{where } \varepsilon = 0.1 \text{ & } N = 3$$

$$= 0.933$$

This value is less than the probability of choosing the greedy action when in State A for the original MDP as shown in 4.4.1.

Q-values for all the other states will remain the same. So, the average return after convergence will be less in the case of adjusted MDP.

In the case of Q-learning, the average return after convergence will be the same in both the cases since the greedy action will remain the same at every state.

TO conclude:

	MDP	adjusted MDP	
SARSA	v_1	v_2	$v_1 > v_2$
Q-learning	v_3	v_4	$v_3 = v_4$

where v_1, v_2, v_3 & v_4 are placeholders for average return after convergence.