Reinforcement Learning

Mid Semester Exam

08/10/2024

Sanjit K. Kaul

Instructions: You have about 120 minutes to work on the questions. Answers with no supporting steps will receive no credit. No resources, other than a pen/pencil, are allowed. In case you believe that required information is unavailable, make a suitable assumption.

Question 1. 15 **marks** You tweak the policy iteration algorithm. Post policy evaluation you carry out policy improvement in a modified way. Specifically, you carry out the greedy policy improvement for exactly one state from the set of states, instead of doing it for all the states. Your choice of state for improvement is uniform and random (without any bias). Is the resulting policy guaranteed to be an improvement? Prove your claim from first principles.

Question 2. 20 **marks** An environment has two non-terminal states s_1 and s_2 and zero or more terminal states. We have two policies π_1 and π_2. It is known that the value function $v_{\pi_1}(\ )$ of policy π_1 satisfies $v_{\pi_1}(s_1) = v_*(s_1)$ and $v_{\pi_1}(s_2) < v_*(s_2)$. Further, the value function $v_{\pi_2}(\ )$ of policy π_2 satisfies $v_{\pi_2}(s_1) < v_*(s_1)$ and $v_{\pi_2}(s_2) = v_*(s_2)$. Assume the MDP is given by the functions $p(s', r|s, a)$, where s is the current state, a is action chosen in the state, s' is the next state and r is the reward obtained by the agent. Answer the following questions.

(a) Derive the optimal policy using the value functions of π_1 and π_2 and the MDP.

(b) Consider a policy π such that $\pi(s_1) = \pi_1(s_1)$ and $\pi(s_2) = \pi_2(s_2)$. Come up with conditions (make suitable reasonable assumptions about the MDP) under which π is an optimal policy. You must show that π is in fact optimal given your assumptions.

Question 3. 25 **marks** An environment has three non-terminal states $0, 1,$ and 2 and terminal states -1 and 3. In each of the non-terminal states, the agent may choose to either go forward or down. Choosing down has the environment transition to terminal state -1 with a reward -1. Choosing forward in 0 transitions the environment to 1 with the agent getting a reward of 0. Choosing forward in 1 transitions the environment to 2 with the agent getting a reward of 0. Choosing forward in 2 transitions the environment to terminal state 3 with the agent getting a reward of 1. Draw the MDP.

Consider two policies $\mu_1(a|s)$ and $\mu_2(a|s)$. Policy μ_1 chooses forward and down with equal probability. Policy μ_2 chooses forward with probability $2/3$ and down with probability $1/3$. Calculate the expected value and variance of the return, starting in state 0, when using policy μ_2, given that the returns available to you were obtained via episodes generated using policy μ_1.

Question 4. 20 **marks** A MDP has three states 0, 1, and 2. In each state, an agent may choose the action left or the action right. We are given the following results of a policy evaluation. We have $Q(0, \text{left}) = -1$,

$Q(0, \text{right}) = 2$, $Q(1, \text{left}) = 0$, $Q(1, \text{right}) = 1$, $Q(2, \text{left}) = -1$, $Q(2, \text{right}) = 4$. We are also giving the following two episodes of data. Episode 1 is $(0, \text{right}, 1)$, $(1, \text{right}, 4)$, $(2, \text{right}, -1)$, $(2, \text{right}, -2)$, $(2, \text{left}, 1)$. Episode 2 is $(1, \text{right}, 4)$, $(2, \text{right}, -1)$, $(2, \text{right}, -2)$, $(2, \text{left}, 1)$, $(1, \text{left}, 3)$, $(0, \text{left}, 1)$. As always, each tuple is (S_t, A_t, R_{t+1}).

Use the two episodes to calculate an improved policy that must be ϵ-soft.

Question 5. 20 **marks** An agent interacts with its environment using its camera. At every time t the agent captures an image using its camera. The agent can choose to process the image locally or send it to the cloud for processing. If it chooses to process a captured image locally, at $t+1$ it receives a reward of 1 and it is again able to make a choice for an image captured at $t+1$. If the agent chooses to send a captured image to the cloud, it must wait for the cloud to return the processed image and can't process any images while it waits for the cloud. An image sent to the cloud at t is processed and returned by the cloud at a certain time $t+k$, $k \geq 1$. Specifically, for an image sent at time t, at each time $t+k$ the processed image is returned to the agent independently with probability $p < 1$. For every time step spent by the cloud processing the image, the agent pays a cost of 1 (reward of -1). On receiving the processed image, the agent receives a reward of $10/k$ in addition to the cost it must pay for the time step.

Propose a suitable set of states for a MDP that can model the agent's interaction with the environment. Define the MDP (as a table or graphically). Note that there is only one state in which the agent chooses an action. Calculate the value of the state under the assumption that the agent always chooses to process the image locally. Further, calculate the value of the state under the assumption that the agent always chooses to send to the cloud. Assume a discount factor $\gamma = 0.8$.

Question 1. 15 **marks** You tweak the policy iteration algorithm. Post policy evaluation you carry out policy improvement in a modified way. Specifically, you carry out the greedy policy improvement for exactly one state from the set of states, instead of doing it for all the states. Your choice of state for improvement is uniform and random (without any bias). Is the resulting policy guaranteed to be an improvement? Prove your claim from first principles.

Suppose we have the policy π with value function $v_\pi(s)$, $s \in S$, where S is the set of all states.

Suppose $s_0 \in S$ is our choice of state for improvement. Let the greedy improvement for s_0 result in policy π'.

We have

$$\pi'(s) = \pi(s) \quad \text{for states } s \neq s_0,$$

$$\pi'(s_0) = \underset{a \in A(s_0)}{\operatorname{argmax}}\; E\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s_0, A_t = a\right]$$

Stating the policy clearly. Up to

5/15

Note that

$$\max_{a \in A(s_0)} E\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s_0, A_t = a\right]$$
$$\geq E_\pi\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s_0\right]$$

That is

$$E_{\pi'}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s_0\right]$$
$$\geq E_\pi\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s_0\right]$$

Also,

$$E_{\pi'}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s\right]$$
$$= E_\pi\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s\right],$$

for all $s \neq s_0$.

In summary,

$$T_{\pi'} v_\pi(s) \geq T_\pi v_\pi(s) \quad \text{for all states } s.$$

Showing that $T_{\pi'} v_\pi(s) \geq T_\pi v_\pi(s)$ Up to

5/15

Further, for every state s,

$$T_{\pi'}\left(T_{\pi'} v_\pi(s)\right) \geq T_\pi v_\pi(s),$$

$$T_{\pi'}\left(T_{\pi'}\left(T_{\pi'} v_\pi(s)\right)\right) \geq T_\pi v_\pi(s),$$

and so on, we get

$$v_{\pi'}(s) \geq T_\pi v_\pi(s) = v_\pi(s), \quad s \in S.$$

$\therefore$ The resulting policy is an improvement.

Using the property of repeat applications of $T_{\pi'}$ resulting in $v_{\pi'}(s)$ to show improvement. Up to

5/15

Question 2. 20 **marks** An environment has two non-terminal states s_1 and s_2 and zero or more terminal states. We have two policies π_1 and π_2. It is known that the value function $v_{\pi_1}(\,)$ of policy π_1 satisfies $v_{\pi_1}(s_1) = v_*(s_1)$ and $v_{\pi_1}(s_2) < v_*(s_2)$. Further, the value function $v_{\pi_2}(\,)$ of policy π_2 satisfies $v_{\pi_2}(s_1) < v_*(s_1)$ and $v_{\pi_2}(s_2) = v_*(s_2)$. Assume the MDP is given by the functions $p(s', r|s, a)$, where s is the current state, a is action chosen in the state, s' is the next state and r is the reward obtained by the agent. Answer the following questions.

(a) Derive the optimal policy using the value functions of π_1 and π_2 and the MDP.

(b) Consider a policy π such that $\pi(s_1) = \pi_1(s_1)$ and $\pi(s_2) = \pi_2(s_2)$. Come up with conditions (make suitable reasonable assumptions about the MDP) under which π is an optimal policy. You must show that π is in fact optimal given your assumptions.

(a) The optimal policy π_* satisfies

$$\pi_*(s) = \underset{a\in\mathcal{A}(s)}{\operatorname{argmax}} \sum_{s',r} (r + V_*(s'))\, p(s',r|s,a)$$

Specifically,

$$\pi_*(s_1) = \underset{a\in\mathcal{A}(s_1)}{\operatorname{argmax}} \left[\sum_r (r+V_*(s_1))\,p(s_1,r|s_1,a) + \sum_r (r+V_*(s_2))\,p(s_2,r|s_1,a)\right]$$

$$= \underset{a\in\mathcal{A}(s_1)}{\operatorname{argmax}} \left[\sum_r (r+V_{\pi_1}(s_1))\,p(s_1,r|s_1,a) + \sum_r (r+V_{\pi_2}(s_2))\,p(s_2,r|s_1,a)\right] \quad \text{— ①}$$

Rewriting in terms of V_{π_1} & V_{π_2}.

$$\pi_*(s_2) = \underset{a\in\mathcal{A}(s_2)}{\operatorname{argmax}} \left[\sum_r (r+V_*(s_1))\,p(s_1,r|s_2,a) + \sum_r (r+V_*(s_2))\,p(s_2,r|s_2,a)\right]$$

$$= \underset{a\in\mathcal{A}(s_2)}{\operatorname{argmax}} \left[\sum_r (r+V_{\pi_1}(s_1))\,p(s_1,r|s_2,a) + \sum_r (r+V_{\pi_2}(s_2))\,p(s_2,r|s_2,a)\right] \quad \text{— ②}$$

① & ② define the optimal policy.

Writing the Bellman optimality equations! Up to 5/20

Making appropriate substitutions Up to 5/20

If just the final equations ① & ② are provided & are correct. 10/20

(b) We want conditions that ensure $\pi_*(s_1) = \pi_1(s_1)$ & $\pi_*(s_2) = \pi_2(s_2)$.

Consider ① & ② above.

Setting $p(s_2|s_1,a) = 0$ & $p(s_1|s_2,a) = 0$, we can rewrite ① & ② as

$$\pi_*(s_1) = \underset{a\in\mathcal{A}(s_1)}{\operatorname{argmax}} \left[\sum_r (r+\gamma V_*(s_1))\,p(s_1,r|s_1,a)\right]$$

and

$$\pi_*(s_2) = \underset{a\in\mathcal{A}(s_2)}{\operatorname{argmax}} \left[\sum_r (r+\gamma V_*(s_2))\,p(s_2,r|s_2,a)\right]$$

The sufficient conditions 5/20

Also note that, for our setting of MDP values,

$$V_{\pi_1}(s_1) = \sum_r (r+\gamma V_{\pi_1}(s_1))\,p(s_1,r|s_1,\pi_1(s_1))$$

$$V_{\pi_2}(s_2) = \sum_r (r+\gamma V_{\pi_2}(s_2))\,p(s_2,r|s_2,\pi_2(s_2))$$

Since $V_*(s_1) = V_{\pi_1}(s_1)$ and $V_*(s_2) = V_{\pi_2}(s_2)$,

$$V_{\pi_1}(s_1) = \sum_r (r+\gamma V_{\pi_1}(s_1))\,p(s_1,r|s_1,\pi_1(s_1))$$

$$= \underset{a\in\mathcal{A}(s_1)}{\operatorname{argmax}} \sum_r (r+\gamma V_{\pi_1}(s_1))\,p(s_1,r|s_1,a)$$

$\therefore$ $\pi(s_1)$ gives the optimal action for state s_1.

Similarly,

$$V_{\pi_2}(s_2) = \sum_r (r+\gamma V_{\pi_2}(s_2))\,p(s_2,r|s_2,\pi_2(s_2))$$

$$= \underset{a\in\mathcal{A}(s_2)}{\operatorname{argmax}} \sum_r (r+\gamma V_{\pi_2}(s_2))\,p(s_2,r|s_2,a).$$

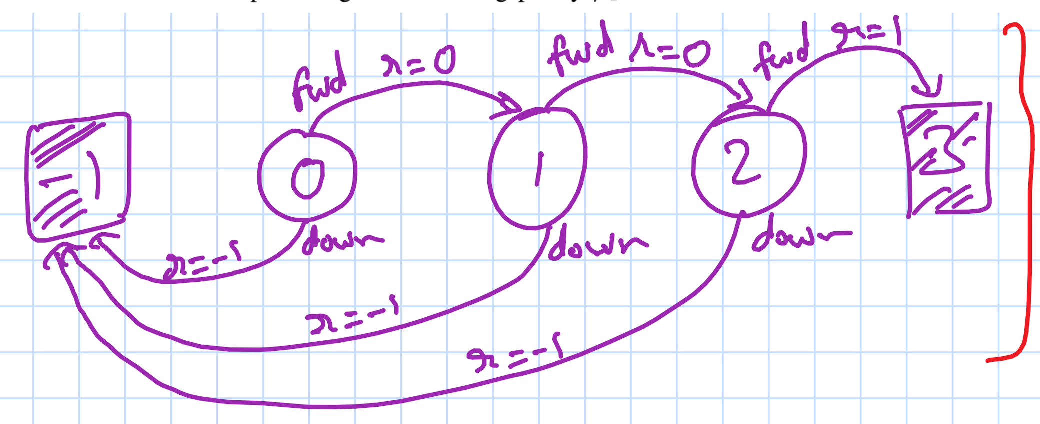
$\pi_2(s_2)$ gives the optimal action for s_2.

Showing that the suggested sufficient conditions result in $\pi_*(s_1) = \pi_1(s_1)$ & $\pi_*(s_2) = \pi_2(s_2)$. 5/20

$$\boxed{\text{Setting } p(s_2|s_1,a)=0 \text{ \& } p(s_1|s_2,a)=0 \text{ are sufficient to ensure a policy } \pi \text{ that satisfies } \pi(s_1)=\pi_1(s_1) \text{ \& } \pi(s_2)=\pi_2(s_2) \text{ is optimal.}}$$

Question 3. 25 **marks** An environment has three non-terminal states $0, 1,$ and 2 and terminal states -1 and 3. In each of the non-terminal states, the agent may choose to either go forward or down. Choosing down has the environment transition to terminal state -1 with a reward -1. Choosing forward in 0 transitions the environment to 1 with the agent getting a reward of 0. Choosing forward in 1 transitions the environment to 2 with the agent getting a reward of 0. Choosing forward in 2 transitions the environment to terminal state 3 with the agent getting a reward of 1. Draw the MDP.

Consider two policies $\mu_1(a|s)$ and $\mu_2(a|s)$. Policy μ_1 chooses forward and down with equal probability. Policy μ_2 chooses forward with probability $2/3$ and down with probability $1/3$. Calculate the expected value and variance of the return, starting in state 0, when using policy μ_2, given that the returns available to you were obtained via episodes generated using policy μ_1.



Up to 5/25 for the MDP.

Consider the return $G_t | S_t = 0$.

G_t takes the value -1 or 1.

$G_t = 1$ if the agent chooses fwd in 0 followed by fwd in 1 & fwd in 2.

$G_t = -1$ if the agent chooses down in 0
or if the agent chooses fwd followed by down
or if the agent chooses fwd, fwd, down.

Our behavior policy is μ_1. To estimate the expected return using μ_2, we must scale the return obtained using μ_1 by the IS factor.

Scale $G_t = 1$ by $\dfrac{\mu_2(\text{fwd}|0)\,\mu_2(\text{fwd}|1)\,\mu_2(\text{fwd}|2)}{\mu_1(\text{fwd}|0)\,\mu_1(\text{fwd}|1)\,\mu_1(\text{fwd}|2)}$

Scale $G_t = -1$ by:

$\dfrac{\mu_2(\text{down}|0)}{\mu_1(\text{down}|0)}$ if the seq of actions is down.

$\dfrac{\mu_2(\text{fwd}|0)\,\mu_2(\text{down}|0)}{\mu_1(\text{fwd}|0)\,\mu_1(\text{down}|0)}$ if the seq of actions is fwd, down.

$\dfrac{\mu_2(\text{fwd}|0)\,\mu_2(\text{fwd}|1)\,\mu_2(\text{down}|0)}{\mu_1(\text{fwd}|0)\,\mu_1(\text{fwd}|1)\,\mu_1(\text{down}|0)}$ if the seq of actions is fwd, fwd, down.

$$E_{\mu_2}[G_t | S_t = 0] = (1)\frac{\mu_2(\text{fwd}|0)\,\mu_2(\text{fwd}|1)\,\mu_2(\text{fwd}|2)}{\mu_1(\text{fwd}|0)\,\mu_1(\text{fwd}|1)\,\mu_1(\text{fwd}|2)}\,\mu_1(\text{fwd}|0)\,\mu_1(\text{fwd}|1)\,\mu_1(\text{fwd}|2)$$

Up to 2.5/10

$$+(-1)\Big[\frac{\mu_2(\text{down}|0)}{\mu_1(\text{down}|0)}\,\mu_1(\text{down}|0)$$

2.5/10

$$+\frac{\mu_2(\text{fwd}|0)\,\mu_2(\text{down}|0)}{\mu_1(\text{fwd}|0)\,\mu_1(\text{down}|0)}\,\mu_1(\text{fwd}|0)\,\mu_1(\text{down}|0)$$

2.5/10

$$+\frac{\mu_2(\text{fwd}|0)\,\mu_2(\text{fwd}|1)\,\mu_2(\text{down}|0)}{\mu_1(\text{fwd}|0)\,\mu_1(\text{fwd}|1)\,\mu_1(\text{down}|0)}\,\mu_1(\text{fwd}|0)\,\mu_1(\text{fwd}|1)\,\mu_1(\text{down}|0)\Big]$$

2.5/10

Up to 10/25 for the four terms

$$= (1)(2/3)(2/3)(2/3) + (-1)(1/3) + (-1)(2/3)(1/3) + (-1)(2/3)(2/3)(1/3)$$

$$= (2/3)(2/3)(1/3) - \frac{2}{3}\frac{1}{3} - \frac{1}{3}$$

$$= \frac{2}{9}\frac{1}{3}\left(-\frac{1}{3}\right) - \frac{1}{3} = -\frac{1}{3}\left(\frac{2}{9} + 1\right)$$

$$= \frac{-11}{27} \,/\!/$$

$$\mathrm{Var}_{\mu_2}[G_t | S_t = 0] = E_{\mu_2}[G_t^2 | S_t = 0] - \left(E_{\mu_2}[G_t | S_t = 0]\right)^2$$

$$E_{\mu_2}[G_t^2 | S_t = 0] = (1)\frac{\mu_2(\text{fwd}|0)\,\mu_2(\text{fwd}|1)\,\mu_2(\text{fwd}|2)}{\mu_1(\text{fwd}|0)\,\mu_1(\text{fwd}|1)\,\mu_1(\text{fwd}|2)}$$

2.5/10

$$+(1)\frac{\mu_2(\text{down}|0)}{\mu_1(\text{down}|0)}$$

2.5/10

$$+(1)\frac{\mu_2(\text{fwd}|0)\,\mu_2(\text{down}|0)}{\mu_1(\text{fwd}|0)\,\mu_1(\text{down}|0)}$$

2.0/10

$$+(1)\frac{\mu_2(\text{fwd}|0)\,\mu_2(\text{fwd}|1)\,\mu_2(\text{down}|0)}{\mu_1(\text{fwd}|0)\,\mu_1(\text{fwd}|1)\,\mu_1(\text{down}|0)}$$

2.5/10

Up to 10/25.

$$= (4/3)^3 + 2/3 + \left(\frac{4}{3}\right)(2/3) + \left(\frac{4}{3}\right)^2\left(\frac{2}{3}\right)$$

$$\text{Variance} = (4/3)^3 + 2/3 + \left(\frac{4}{3}\right)(2/3) + \left(\frac{4}{3}\right)^2\left(\frac{2}{3}\right) - \left(\frac{11}{27}\right)^2 /\!/$$

Question 4. 20 **marks** A MDP has three states 0, 1, and 2. In each state, an agent may choose the action left or the action right. We are given the following results of a policy evaluation. We have $Q(0, \text{left}) = -1$, $Q(0, \text{right}) = 2$, $Q(1, \text{left}) = 0$, $Q(1, \text{right}) = 1$, $Q(2, \text{left}) = -1$, $Q(2, \text{right}) = 4$. We are also giving the following two episodes of data. Episode 1 is $(0, \text{right}, 1)$, $(1, \text{right}, 4)$, $(2, \text{right}, -1)$, $(2, \text{right}, -2)$, $(2, \text{left}, 1)$. Episode 2 is $(1, \text{right}, 4)$, $(2, \text{right}, -1)$, $(2, \text{right}, -2)$, $(2, \text{left}, 1)$, $(1, \text{left}, 3)$, $(0, \text{left}, 1)$. As always, each tuple is (S_t, A_t, R_{t+1}).

Use the two episodes to calculate an improved policy that must be ϵ-soft.

Using the two episodes, we can update the Q values in the following manner. Here we will assume first-visit MC and $\alpha = 0.2$. This is essentially policy eval, given the episodes & current Q-values.

$$Q(0, \text{left}) = -1 + (0.2)(1 - (-1)) = -0.6.$$

$$Q(0, \text{right}) = 2 + (0.2)(3 - 2) = 2 + 0.2 = 2.2.$$

$$Q(1, \text{left}) = 0 + (0.2)(4 - 0) = 0.8.$$

$$Q(1, \text{right}) = 1 + (0.2)\left(\frac{2+6}{2} - 1\right)$$

$$= 1 + (0.2)(3) = 1.6 \,//$$

$$Q(2, \text{left}) = -1 + (0.2)\left(\frac{1+5}{2} - (-1)\right)$$

$$= -1 + (0.2)(4) = -0.2 \,//$$

$$Q(2, \text{right}) = 4 + (0.2) = 4.2$$

Any legitimate way of updating Q-values starting with the given and using the 2 episodes is fine. I used the sample mean.

The Q-values above must be used to come up with an improved policy.

$$\pi(a|0) = \begin{cases} \epsilon/2 & \text{left,} \\ 1 - \epsilon/2 & \text{right.} \end{cases}$$

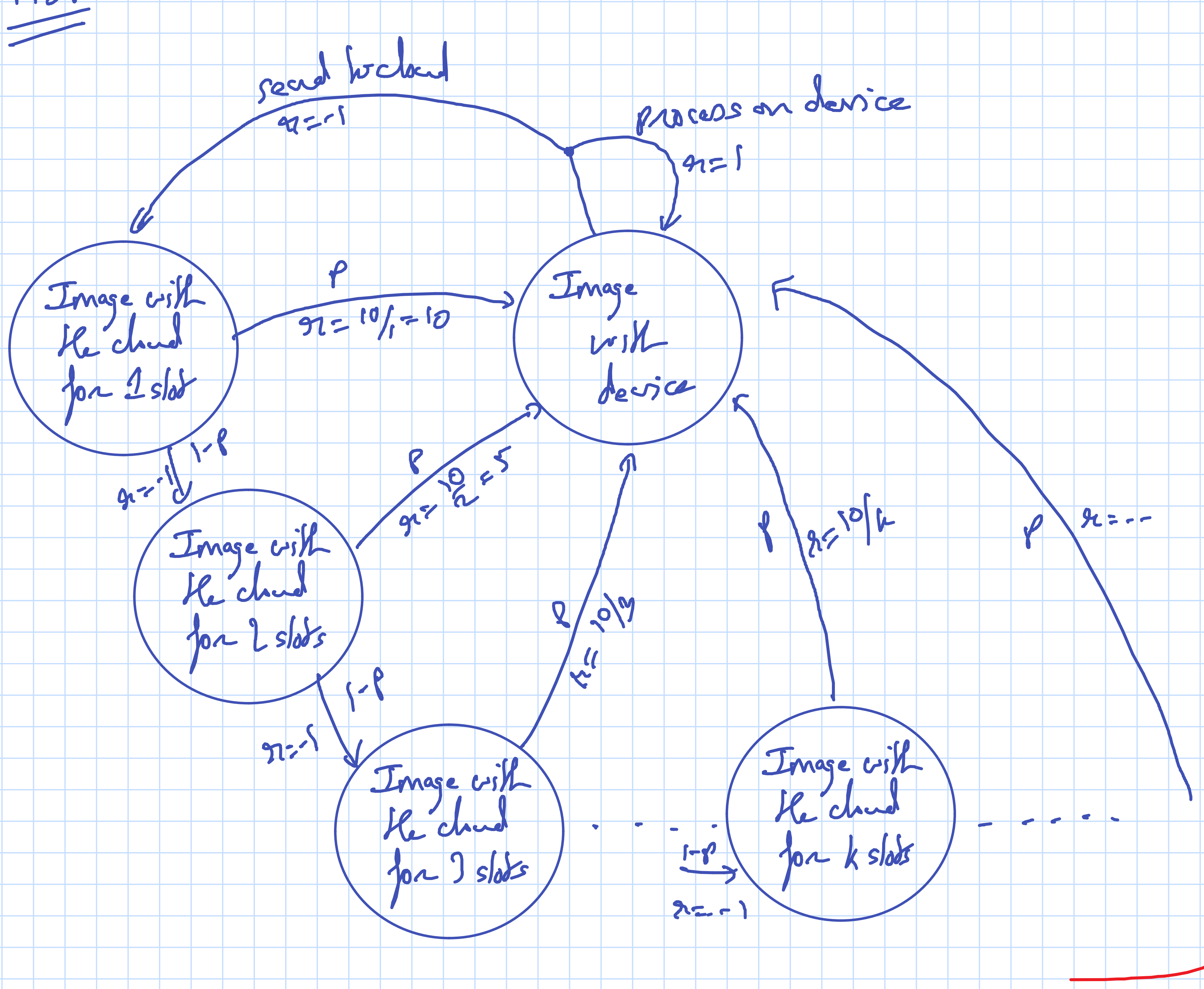
$$\pi(a|1) = \begin{cases} \epsilon/2 & \text{left,} \\ 1 - \epsilon/2 & \text{right.} \end{cases}$$

$$\pi(a|2) = \begin{cases} \epsilon/2 & \text{left,} \\ 1 - \epsilon/2 & \text{right.} \end{cases}$$

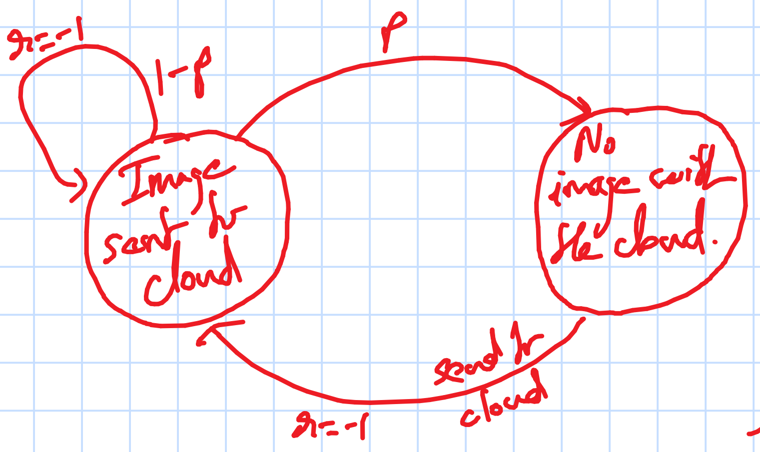
Question 5. 20 **marks** An agent interacts with its environment using its camera. At every time t the agent captures an image using its camera. The agent can choose to process the image locally or send it to the cloud for processing. If it chooses to process a captured image locally, at $t+1$ it receives a reward of 1 and it is again able to make a choice for an image captured at $t+1$. If the agent chooses to send a captured image to the cloud, it must wait for the cloud to return the processed image and can't process any images while it waits for the cloud. An image sent to the cloud at t is processed and returned by the cloud at a certain time $t+k$, $k \geq 1$. Specifically, for an image sent at time t, at each time $t+k$ the processed image is returned to the agent independently with probability $p < 1$. For every time step spent by the cloud processing the image, the agent pays a cost of 1 (reward of -1). On receiving the processed image, the agent receives a reward of $10/k$ in addition to the cost it must pay for the time step.

Propose a suitable set of states for a MDP that can model the agent's interaction with the environment. Define the MDP (as a table or graphically). Note that there is only one state in which the agent chooses an action. Calculate the value of the state under the assumption that the agent always chooses to process the image locally. Further, calculate the value of the state under the assumption that the agent always chooses to send to the cloud. Assume a discount factor $\gamma = 0.8$.

MDP



Up to 10/20



2/20 for attempt of such kind.

Value of the state "image with device", in case image is always processed at the device is

$$1 + (0.8)1 + (0.8)^2 1 + \dots$$

$$= 1 \cdot \frac{1}{1-0.8} = 5$$

2/20

Assuming we always send to the cloud:

Let's call the above policy π.

$$V_\pi(s) = E\left[R_{t+1} + \gamma V_\pi(s) \mid S_t = s\right]$$

$$= E\left[R_{t+1} \mid S_t = s\right] + (0.8) V_\pi(s)$$

2/20

Here t are the decision instants.

$$\therefore R_{t+1} = \begin{cases} -1 + 10 & \text{w.p } p \\ -2 + 10/2 & \text{w.p } (1-p)p \\ -3 + 10/3 & \text{w.p } (1-p)^2 p \\ \vdots & \vdots \\ -k + 10/k & \text{w.p } (1-p)^k \frac{10}{k} \\ \vdots & \vdots \end{cases}$$

$$E\left[R_{t+1}\right] = -\frac{1}{p} + 10\boxed{\left(p + \frac{(1-p)p}{2} + \frac{(1-p)^2 p}{3} + \dots\right)}$$

$< \frac{1}{p}$, for $p \neq 0$.

$$= -\frac{1}{p} + 10p\left(1 + \frac{(1-p)}{2} + \frac{(1-p)^2}{3} + \dots\right)$$

$= c$, where c is some constant, for $p \neq 0$. $c < 10 - 1/p$.

$$V_\pi(s) = c + (0.8) V_\pi(s)$$

$$V_\pi(s)(0.2) = c.$$

$$\Rightarrow V_\pi(s) = \frac{c}{0.2} = 5c < 50 - 5/p$$

Great approach for calculating the expected reward.

6/20