This question has three parts, each of which can be answered concisely, but be prepared to explain and justify your concise answer.

1. Suppose you have a policy π and its action-value function, q_π, then you greedify q_π to produce the deterministic policy π':

$$\pi'(s) = \arg\max_a q_\pi(s, a) \qquad \forall s \in \mathcal{S}.$$

(a) What do you know about the relationship between π and π'?

(b) Now suppose you notice that π' is the same as π. What then do you know about the two policies?

(c) Now suppose you notice that π' is different from π. Do you know anything more about the two policies other than what you reported in part (a)?

2. The goal of reinforcement learning can be seen as producing a ______________, which maps from _____________ to _______________.

3. From state x, taking action 1 always produces a reward of 2 and sends you to a state y from which a return of 10 is always received. The discount parameter gamma is 0.9. What is v_π(y)? What is q_*(x,1)?

4. Suppose the discount rate γ is 0.5 and the following sequence of rewards is observed: R_1=7, R_2=6, R_3=-4, R_4=4, R_5=8, R_6=2, followed by the terminal state. What are the following returns?

G_6?

G_5?

G_4?

G_3?

G_2?

G_1?

G_0?

5. Given a choice between two actions, we (should) always pick the one with the larger _______.

a) reward
b) return
c) value

6. An episodic task begins and ends.
A _______ task goes on and on.

a) continuous
b) discounted
c) continuing
d) average reward

7. Suppose the discount rate gamma is 0.5 and the following sequence of rewards is observed: R_1=1, R_2=6, R_3=-12, R_4=16, followed by the terminal state. What are the following returns?

G_4?
G_3?
G_2?
G_1?
G_0?

8. Suppose the discount rate gamma is 0.5 and the following sequence of rewards is observed: R_1=1, followed by an infinite sequence of rewards of +13. What are the following returns?

G_2?
G_1?
G_0?

Question 9. Give a definition of v_π in terms of q_π.

Question 10. Give a definition of q_π in terms of v_π.

Question 11. Give a definition of v_* in terms of q_*.

Question 12. Give a definition of q_* in terms of v_*.

Question 13. Give a definition of π_* in terms of q_*.

Question 14. Give a definition of π_* in terms of v_*.

Question 15. Sketch the backup diagrams for the following tabular learning methods:

(a) TD(0)

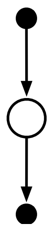
(b) One-step Expected Sarsa

(c) single-step full backup of v_π

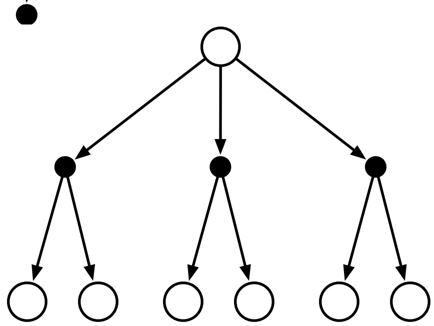
(d) 2-step Tree backup

(e) Monte Carlo backup for v_π

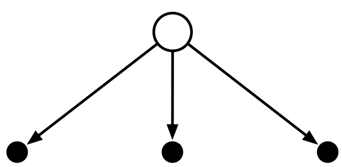
Question 16. Write the update that corresponds to the following backup diagrams:



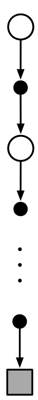
(a)



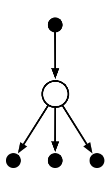
(b)



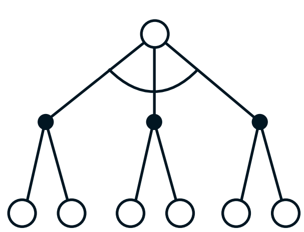
(c)



(d)

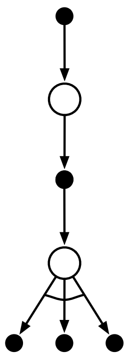


(e)



(f)

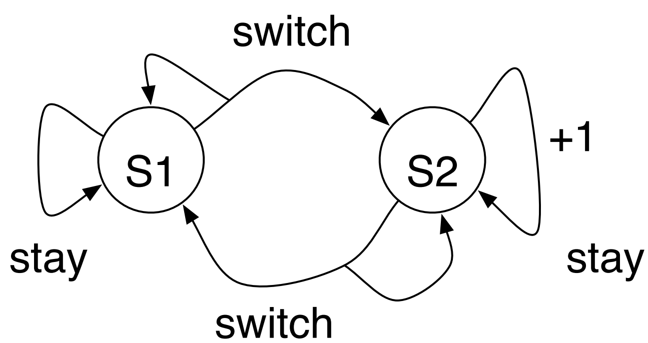
(g) 

(h) 

Question 17. For a finite continuing discounted MDP with discount factor γ, suppose you know two numbers $r_{\min}$ and $r_{\max}$ such that for all $r \in \mathcal{R}$, $r_{\min} \leq r \leq r_{\max}$. Give epresssions for two numbers $v_{\min}$ and $v_{\max}$ such that $v_{\min} \leq v_\pi(s) \leq v_{\max}$ for all states $s \in S$ and all policies π.

Question 18. Markov Decision Processes

Consider the MDP in the figure below. There are two states, *S1* and *S2*, and two actions, *switch* and *stay*. The *switch* action takes the agent to the other state with probability 0.8 and stays in the same state with probability 0.2. The *stay* action keeps the agent in the same state with probability 1. The reward for action *stay* in state *S2* is 1. All other rewards are 0. The discount factor is $\gamma = \frac{1}{2}$.



(a) What is the optimal policy?

(b) Compute the optimal value function by solving the linear system of equations corresponding to the optimal policy.

(c) Suppose that you are doing synchronous value iteration to compute the optimal state-value function. You start with all value estimates equal to 0. Show the value estimates after 1 and 2 iterations respectively.

(d) Suppose you are doing TD-learning. You start with all value estimates equal to 0, and you observe the following trajectory (sequence of states, actions and rewards):

$$S1, switch, 0, S2, stay, +1, S2$$

Assuming the learning rate $\alpha = 0.1$, show the TD-updates that are performed.

Question 19. From state A, the first action leads deterministically to rewards of 2, 4, and 9 followed by a return to A, whereas the second action leads deterministically to a reward of 3 followed by an immediate return to state A. For what values of γ is the first action the better action? To solve this you may have to use the formula for solving quadratic equations.

Question 20. What is generalized policy iteration? Refer to all three words of the phrase in your explanation.