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'?

$\pi' \geq \pi \quad \equiv V_{\pi'}(s) \geq V_\pi(s) \quad \forall s$

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

both are optimal

(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)?

No
In particular, π may be optimal

2. The goal of reinforcement learning can be seen as producing a policy, which maps from states to actions.

Q,R
act value fn
states & acts
Rewards

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 $v_\pi(y)$? What is $q_*(x,1)$?

$10 = 10$ $\quad 2 + \gamma V_*(y) = 2 + .9 \cdot 10 = 11$

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? 0

G_5? 2

G_4? 9

G_3? 8.5

G_2? $-4 + 4.25 = .25$

G_1? $6 + .125$

G_0? $7 + \frac{6.125}{2}$

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? 0
G_3? ~~-12+8=-4~~ 16
G_2? -12+8=-4
G_1? 6-2=4
G_0? 1+2=3

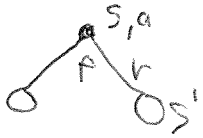
8. Suppose the discount rate gamma is 0.5 and the following sequence of rewards is observed: R_1=1, followed by an infinite sequence of rewards of +13. What are the following returns?

G_2? 26
G_1? 26
G_0? 1+13=14

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

$$v_\pi(s) = \sum_a \pi(a|s)\, q_\pi(s,a)$$

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



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

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_*.

$$\pi_*(s) = \arg\max_a q_*(s,a)$$

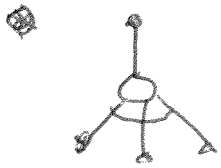
Question 14. Give a definition of π_* in terms of v_*.

$$\pi_*(s) = \arg\max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_*(s')\right]$$

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

(a) TD(0)

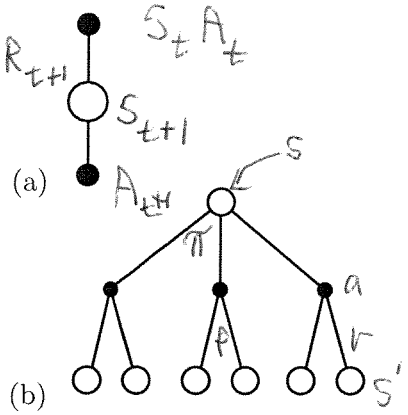
(b) One-step Q-learning



(c) single-step full backup of v_π

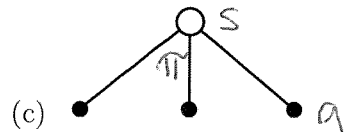
(d) Monte Carlo backup for q_π

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

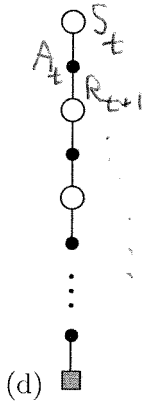


(a)

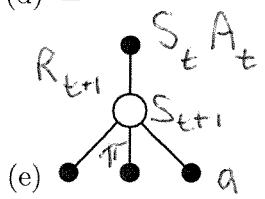
(b)



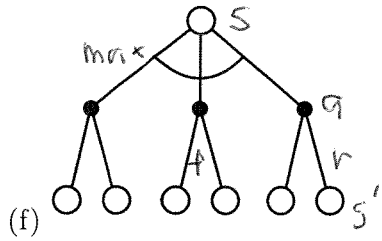
(c)



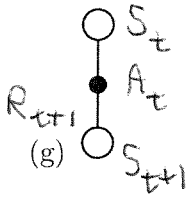
(d)



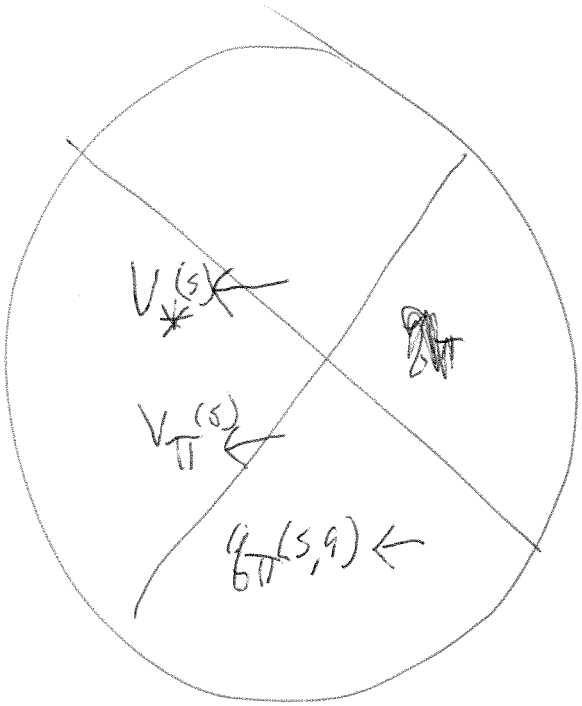
(e)



(f)



(g)



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.

$$V_{max} = \frac{1}{1-\gamma} \cdot r_{max}$$

$$V_{min} = \frac{1}{1-\gamma} r_{min}$$

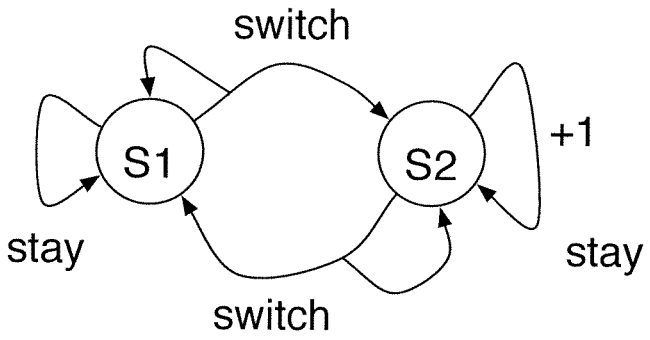
Question 18. What is generalized policy iteration? Refer to all three words of the phrase in your explanation.

~~Iteratively~~

- Continually changing a value fn towards the value fn for a policy,
- While changing the policy toward the greedy policy for the value fn.
- Generalized means the two steps could be done completely and alternately or intermixed more finely (incompletely) randomly, or even from sample experience

Question 19. 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?

s2 → stay

s1 → switch

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

$V_*(s1) = 0.2\left(0 + \gamma V_*(s1)\right) + 0.8\left(0 + \gamma V_*(s2)\right)$

$V_*(s2) = 1 + \gamma V_*(s2)$

Question 20. 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.