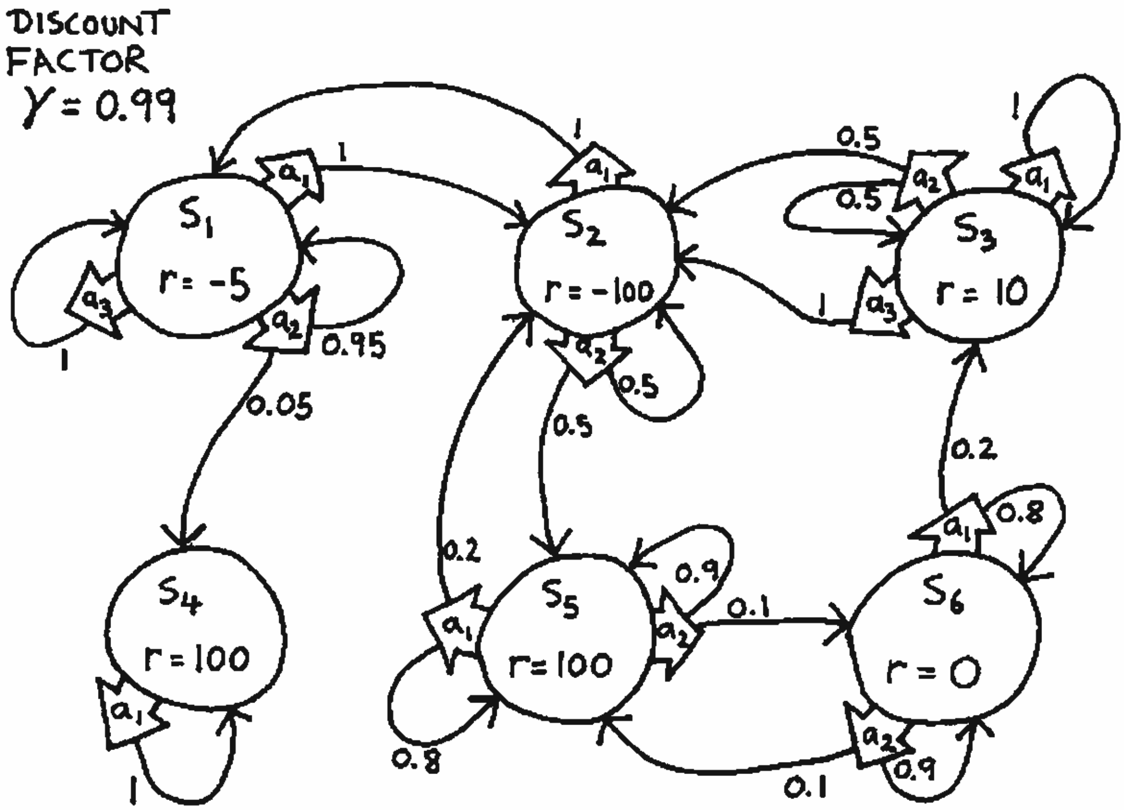
Take home Quiz V

Solution

NAME:
DATE:

NOTE: You don't need to turn this in.

a) By thinking carefully, and perhaps hitting a few keys on your calculator, it should be possible to deduce the optimal policy for the following MDP without needing to run Value Iteration.



You must write down the optimal π policy below:

$\pi(S_1) = a_2$
$\pi(S_2) = a_1$
$\pi(S_3) = a_3$
$\pi(S_4) = a_1$
$\pi(S_5) = a_1$
$\pi(S_6) = a_1$

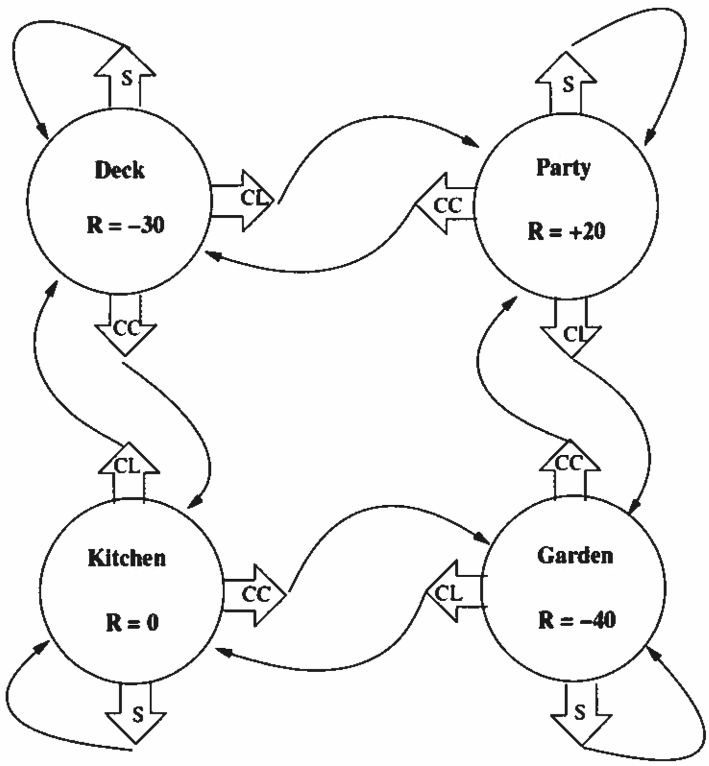
Reasoning

γ is almost 1 (i.e. its 0.99)

∴ long-term reward will play a key-role as γ is not being decayed.

→ Best state to reach is therefore S4 because it has r=100 and when in it we don't need to leave, so no further state to reduce the utility.

2. You are a wildly implausible robot who wanders among the four areas depicted below. You hate rain and get a reward of -30 on any move that starts in the deck and -40 on any move that starts in the Garden. You like parties, and you are indifferent to kitchens



Actions: All states have three actions: Clockwise (**CL**), Counter-Clockwise (**CC**), Stay (**S**). Clockwise and Counter-Clockwise move you through a door into another room, and Stay keeps you in the same location. All transitions have are deterministic (probability 1.0).

a) How many distinct policies are there for this MDP?

$3^4 = 81$ i.e 3 actions 4 states.

b) Let J*(Room)= expected discounted sum of future rewards assuming you start in "Room" and subsequently act optimally. Assuming a discount factor γ=0.5 , give the J* values for each room.

Looking through the problem, I can deduce that optimal policies: $\pi(K) = S$, $\pi(G) = CC$
$\pi(D) = CL$, $\pi(P) = S$

Therefore: $J^*(D) = -30 + \frac{1}{2} J^*(P)$
$J^*(P) = 20 + \frac{1}{2} J^*(P)$
$J^*(K) = 0 + \frac{1}{2} J^*(K)$
$J^*(G) = -40 + \frac{1}{2} J^*(P)$

Solving for J's:

$J^*(Deck) = -10$

$J^*(P) = 40$

$J^*(K) = 0$

$J^*(G) = -20$

Recall
D – Deck
P – Party
K – Kitchen
G – Garden

c) The optimal policy when the discount factor, γ, is small but non-zero (e.g. γ=0.1) different from the optimal policy when is large e.g. γ=0.9). If we began with γ=0.1 , and then gradually increased , what would be the threshold value of above which the optimal policy would change?

As γ increases, π changes from S in kitchen to CL. this will occur when value for taking action S in Kitchen is equal to value of taking action CL

$\therefore \quad J^S(K) = 0 + \gamma J^S(K) = 0$

$J^{CL}(K) = 0 + \gamma J^*(D)$

But optimal policy in Deck is CL even without considering γ.

$J^*(D) = -30 + \gamma J^*(P)$

also: $J^*(P) = 20 + \gamma J^*(P)$

$\therefore \quad J^*(P) = \frac{20}{1-\gamma}$

This then resolves to:

$\gamma\left(-30 + \gamma\left(\frac{20}{1-\gamma}\right)\right) = 0$

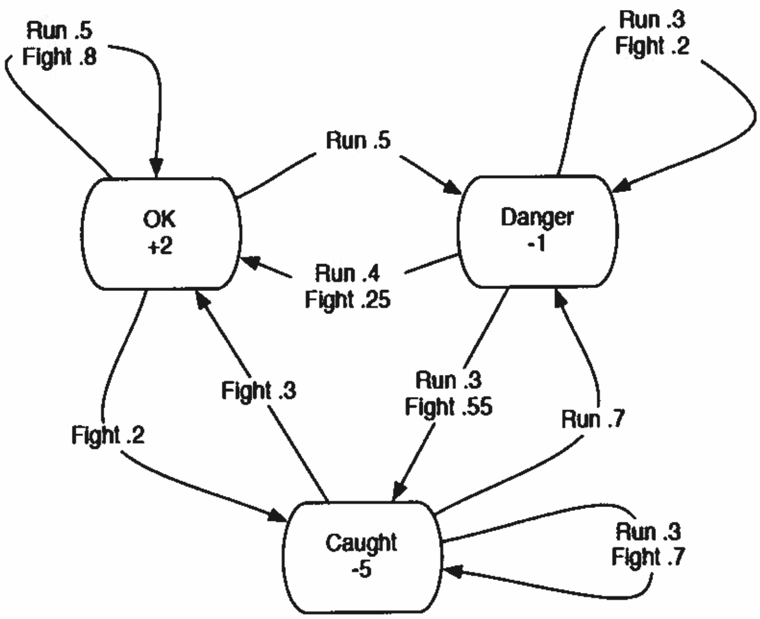
$-30 + \gamma\left(\frac{20}{1-\gamma}\right) = 0$

$20\gamma = 30(1-\gamma)$

$\gamma = 3/5 = 0.6$

3. A boy is being chased around the school yard by bullies and must choose whether to Fight or Run.
    - There are three states:
        - Ok (O), where he is fine for the moment.
        - Danger (D), where the bullies are right on his heels.
        - Caught (C), where the bullies catch up with him and administer noogies.
    - He begins in state O 75% of the time.
    - He begins in state D 25% of the time.

The graph of the MDP is given here:



a) Fill out the table with the results of value iteration at k =2 with a discount factor $\gamma =$ 0. 9. Note to calculate values at k=1, I used initial values as 0.

k	$J^k(O)$	$J^k(D)$	$J^k(C)$
1	2	-1	-5
2	2.54	-1.9	-6.98

$$J^2(O) = 2 + 0.9 \max_a \begin{Bmatrix} (0.5*2)+(0.5*-1), \\ (0.8*2)+(0.2*-5) \end{Bmatrix}$$

$$= 2 + 0.9\,(0.6) = 2.54$$

$$J^2(D) = -1 + 0.9 \max \begin{Bmatrix} (0.4 * 2) + (0.3 * -1) + (0.3 * -5), \\ (.25 * 2) + (0.2 * -1) + (0.55 * -1) \end{Bmatrix}$$

$$= -1 + 0.9\,(-1)$$

$$= -1 - 0.9 = \underline{\underline{-1.9}}$$

$$J^2(C) = -5 + 0.9 \max \begin{Bmatrix} (0.7 \times -1) + (0.3 * -5) \\ (0.3 * 2) + (0.7 * -5) \end{Bmatrix}$$

$$-5 + 0.9\,(-2.2)$$

$$= \underline{\underline{-6.98}}$$

__Note:__ please check whether these calculations are correct; I did them off my head.

b) At k = 2 with γ =0. 9 what policy would you select? Is it necessarily true that this is the optimal policy? At k = 3 what policy would you select? Is it necessarily true that this is the optimal policy?

O $\xrightarrow{Do}$ Fight
D $\xrightarrow{Do}$ Run
C $\xrightarrow{Do}$ Run

value iteration that has not converged is not guarantee to find an optimal policy; so this policy is not optimal.

c)

I. Suppose you have a robot trying to reach a goal and avoid cliffs in a small grid world. It can only move North, South, East, or West, but occasionally fails to move in the intended direction. If you were to model this using an MDP and were trying to solve it optimally, should you use value iteration or policy iteration? Justify your answer in one sentence.

value iteration

Reason: we have many states & few actions. we know value iteration is cheaper than policy iteration, so use that.

II. Now suppose that the robot can teleport to any grid cell but the teleportation causes it to land in neighboring grid cells near the target with some probability. Of you were to model this using an MDP and were trying to solve it optimally should you use value iteration or policy iteration? Justify your answer in one sentence.

Policy iteration. It is generally better when there are many actions.

Justification: Instead of 4 actions, we now have an action to teleport to each different square of the grid, which increases the action.