第四次理论作业

1.给定如下图（a）所示的2×3网格。状态1为初始状态，状态6为目标状态。当到达状态6时，智能体获得+10的奖励值，并结束这一局的交互。到达状态1-5均会导致-1的奖励值。每个状态可以执行up、down、left、right四个动作。假设当前的所有状态的Q值表如下图（b）所示，并且智能体采取贪心策略(Greedy)。试回答以下问题。

4	5	Finish 6
Start 1	2	3

(a)

Q(1,up)=4	N/A	N/A	Q(1,right)=3
Q(2,up)=6	N/A	Q(2,left)=3	Q(2,right)=8
Q(3,up)=9	N/A	Q(3,left)=7	N/A
N/A	Q(4,down)=2	N/A	Q(4,right)=5
N/A	Q(5,down)=6	Q(5,left)=5	Q(5,right)=8

(b)

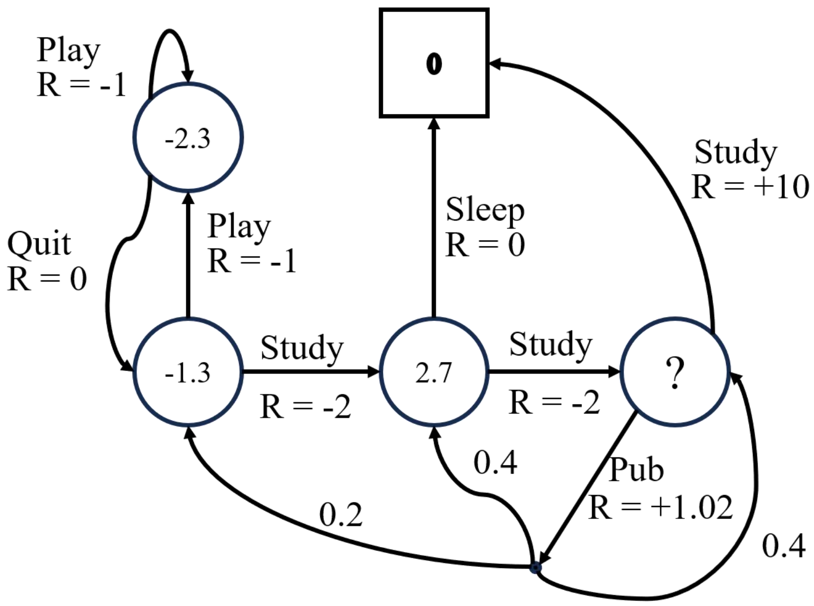
(a) 假设环境模型已知，状态转移概率为1，折扣因子$\gamma=0.9$。请根据值函数定义，计算执行一次更新以后的Q(3,left)值　是多少？

(b) 假设学习率$\alpha=0.2$，折扣因子$\gamma=0.8$，给定轨迹：状态1->right->状态2->up->状态5->right->6，请使用时序差分（Temporal Difference）公式更新这条轨迹上的Q值函数。

2.给定如下图所示的MDP，其中图中的节点（圆圈或正方形）均表示状态，且节点里面的数值表示此状态迭代若干次后的值函数V。除了问号节点外，节点与节点之间的边表示可行的动作，边上的R表示执行此动作带来的奖励值。对于每个状态而言，不同动作被选择的概率均相同。问号节点可以执行Study和Pub两个动作，执行Study动作会带来奖励值+10，并转移到正方形节点，执行Pub动作会带来奖励值+1，且分别以0.2、0.4、0.4的概率转移到下一个状态。

（a）假设$\gamma=1$，计算图中问号处节点的值函数V。

（b）根据V值和Q值之间的关系，计算图中所有状态动作对的Q值。



3.考虑如下的三个 MDP：

（1）MDP 1:

1）转移函数：$P(s_1|s_1,a_1)=1, P(s_2|s_1,a_2)=1, P(s_2|s_2,a_1)=1, P(s_2|s_2,a_2)=1$

2）奖励函数：$R(s_2|s_1,a_2)=1$，否则为 0

（2）MDP 2:

1）转移函数：$P(s_1|s_1,a_1)=1, P(s_2|s_1,a_2)=1, P(s_2|s_2,a_1)=1, P(s_2|s_2,a_2)=1$

2）奖励函数：$R(s_2|s_2,a_1)=1, R(s_2|s_2,a_2)=1$，否则为 0

（3）MDP 3:

1）转移函数：$P(s_1|s_1,a_1)=1, P(s_2|s_1,a_2)=1, P(s_1|s_2,a_1)=1, P(s_2|s_2,a_2)=1$

2）奖励函数：$R(s_1|s_1,a_1)=1, R(s_2|s_2,a_2)=1$，否则为 0

（a）假设$\gamma=1$，对于状态s_1来说，是否存在一个策略使得状态s_1的值函数无限大？如果不存在这样一个策略，请给出使得状态s_1的值函数最大的策略。

（b）当$\gamma=1-\epsilon$，其中ϵ是一个很小的正数，给出上述三个 MDP 的最优策略。若有存在多个最优策略，只需给出其中的一个。