Name: ID:

1. (10 points) Consider the grid in Figure 1. The arrows indicate all possible actions where labels correspond to the reward of the action (unlabeled actions have 0 reward). Apply the Q learning algorithm to this grid, assuming the table of $\widehat{Q}$ is initialized to zero. In the table below are the Q values you will update. Note that the updates start in cell B1, go clockwise until reaching B2, then repeats starting in B1. For this problem only fill out the updates in the table - you do not need to present a table of the final action rewards. For this problem let γ = .9.

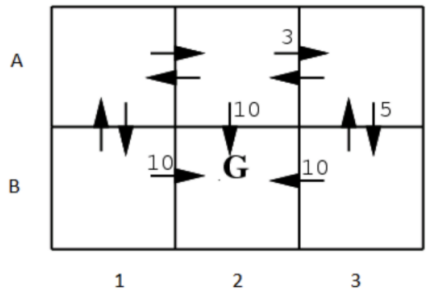

Figure 1: Grid for problems 7-9

a. (10 points) Fill the blanks in the table with numeric formulas. (example: 3 +.9*5)

Q(B1, Up) = r(B1, Up) + γ*max(Q(A1,right), Q(A1,down)) = 0 + 0.9 * max(0, 0) = 0

Q(A1, Right) = r(A1, Right)+γ*max(Q(A2,left), Q(A2,right), Q(A2,down)) = 0 + 0.9 * max(0, 0, 0) = 0

Q(A2, Right) = r(A2, Right)+γ*max(Q(A3,left), Q(A3,down)) = 3 + 0.9 * max(0, 0) = 3

Q(A3, Down) = r(A3, Down)+γ*max(Q(B3,up), Q(B3,left)) = 5 + 0.9 * max(0, 0) = 5

Q(B3, Left) = r(B3, Left)+0 = 10 = 10

Q(B1, Up) = r(B1, Up) + γ*max(Q(A1,right), Q(A1,down)) = 0 + 0.9 * max(0, 0) = 0

Q(A1, Right) = r(A1, Right)+γ*max(Q(A2,left), Q(A2,right), Q(A2,down)) = 0 + 0.9 * max(0, 3, 0) = 0.9*3

Q(A2, Right) = r(A2, Right)+γ*max(Q(A3,left), Q(A3,down)) = 3 + 0.9 * max(0, 5) = 3 + 0.9*5

Q(A3, Down) = r(A3, Down)+γ*max(Q(B3,up), Q(B3,left)) = 5 + 0.9 * max(0, 10) = 5 + 0.9*10

Q(B3, Left) = r(B3, Left)+0 = 10 = 10

Step	Start State	Subsequent State	Update to Q(s,a)
1	B1	A1	Q(B1,Up)=0
2	A1	A2	Q(A1,Right)=0
3	A2	A3	Q(A2,Right)=3
4	A3	B3	Q(A3,Down)=5
5	B3	B2	Q(B3,Left)=10
6	B1	A1	Q(B1,Up)=0
7	A1	A2	Q(A1,Right)=.9*3=2.7
8	A2	A3	Q(A2,Right)=3+.9*5=7.5
9	A3	B3	Q(A3,Down)=5+.9*10=14
10	B3	B2	Q(B3,Left)=10

b. (4 points) Suppose we run Q-learning twice on the grid in Figure 1, once with γ = .99 and once with γ = .01. How will the policies generated by the rewards differ with respect to how the agent reaches B2? Assume the agent starts at B1. (Note that we're not asking you to manually run Q-learning, we just want to know how the two policies will qualitatively differ.)

γ = .99: The agent tends to circle around B2 before going to it.
γ = .01: The agent tends to move immediately move to B2 from B1
Q(s,a) = r + γ * max(Q(s',a'))
γ is the evaluates the future rewards.
If it is equal to 1, the agent values future reward as much as current reward. If an agent does something good, this is as valuable as doing this action directly. For example, in this question, the total reward for clockwise path is 3+5+10=18
If it is equal to 0, it will cause the agent to only value immediate rewards.
argmax(Q(B1,right), Q(B1, up)) = argmax(10,0)

c. (1 points) Suppose now that the values in Figure 1 are not the state-action rewards, but rather the values of a learnt Q-table (where missing values indicate a reward of 0). Construct a policy using these values, assuming the agent starts in cell A2.

A1: right (or down)
A2: down (full points)
A3: down
B1: right
B3: left