Problem 1 – Q - Learning

[50 points]
An enterprising student builds a q-learning agent to buy and sell real estate. The agent begins owning Nothing. When the agent owns Nothing, it can try to buy land (BuyL) or buy a mansion (BuyM). If successful, the agent then owns Land or a Mansion. When the agent owns property, he can only try to Sell. Once the property is sold, the scenario ends in the Terminal state. Assume a discount factor of ϒ = 1.

a) At first, the student makes decisions for two episodes and the agent just learns Q(s, a). Fill in the table with Q-value estimates after each observation. Use a learning rate of α = 0.5. All estimates begin at 0. Terminal state T has no actions and a value of 0.

Observation				**New Q Estimates**			
State s	**Action** a	**Successor** s'	**Reward** r	$Q(N, BuyL)$	$Q(N, BuyM)$	$Q(L, S)$	$Q(M, S)$
			Initial values:	0	0	0	0
N	$BuyL$	L	-10				
L	Sell	T	20				
N	$BuyL$	N	-2				
N	$BuyM$	M	-20				
M	Sell	M	-8				
M	Sell	T	60				

b) What is the optimal policy from this q-learning agent, and what is the q-learning agent's estimate of the expected sum of discounted future rewards starting in N under this policy?

c) If we instead use these observations to estimate a transition model, what would it be?

$T(N, BuyL, N) =$		$T(L, Sell, L) =$	
$T(N, BuyL, L) =$		$T(L, Sell, T) =$	
$T(N, BuyM, N) =$		$T(M, Sell, M) =$	
$T(N, BuyM, M) =$		$T(M, Sell, T) =$	

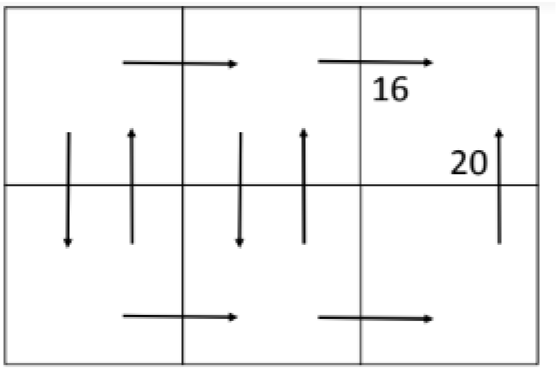
d) If the same sequence of 6 observations in (a) repeated indefinitely and the q-learner very slowly decreased its learning rate, what would Q*(N,BuyM) converge to? Justify your answer.

Problem 2 – Reinforcement Learning

[50 points]

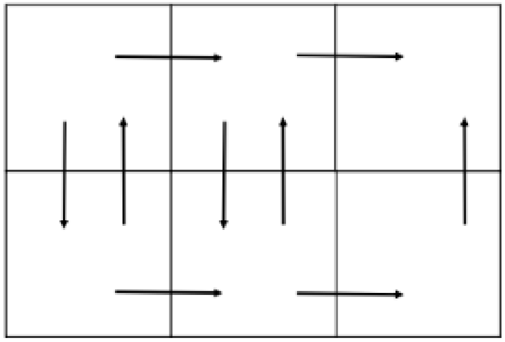
Given the following deterministic world where:

- Possible moves shown by arrows
- Reward is indicated by numbers near the arrows (or zero if there is no number). For example, the reward of moving from the bottom right-state to the top-right state is 20.

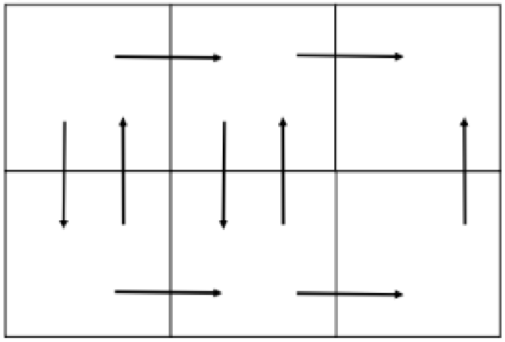


1. Find the optimal $Q^*(s, a)$ values for the discount factors 0.5 and 0.9 i.e., fill the following figures

$\gamma = 0.5$

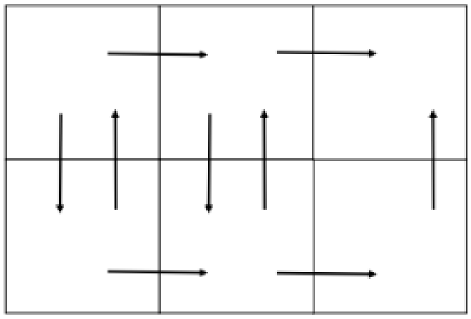


$\gamma = 0.9$

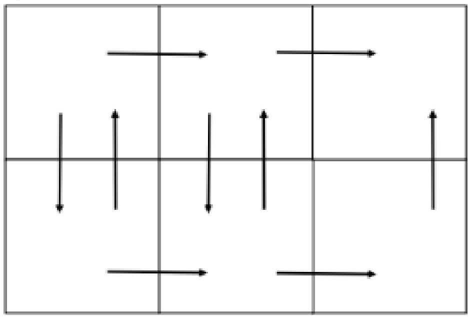


2. Based on the former section, find the optimal $V^*(s)$ values for the discount factors 0.5 and 0.9. What is the optimal policy? I.e., fill and mark the following figures -

$\gamma = 0.5$



$\gamma = 0.9$



3. In general, if we are doing reinforcement learning, why would we prefer an agent to get front-loaded instantaneous rewards which are good indicators of total rewards as opposed to getting all rewards at the end of the game? Give an answer which holds even if the discount factor is 1.