CS 4804 Homework 8
Solution Sketches

---

1. **(20 points)** The first and third both lead to the same *deterministic* strategy: move to the goal state as quickly as possible. The extra 10 points in the third reward table does not change this strategy because the -1 rewards make it ultimately attractive to move to the goal state quickly regardless of the reward upon getting there. The second formulation, on the other hand, has no penalty for staying in a given state, so the strategy formed will a probabilistic policy that encourages either loitering around or moving to the goal state. Without the -1 reward on every move, there is no incentive to reach the goal quickly.

2. **(30 points)** Assume V(3)=0.

   (a) In state 1, performing action b will result in getting to the goal (and thus stopping the negative rewards) fastest. In state 2, this is still true, but because the negative rewards are doubled from state 1, and it is much 'easier' to go to 1 than 3, the policy will seek to go to 1 with action a.

   (b) Perform policy iteration as follows:
   Initially $\pi_1(1) = b$ and $\pi_1(2) = b$

V_1	a	b
1	$0.8(V(2)-2)+0.2(V(1)-1) = -18$	$0.9(V(1)-1) = -9$
2	$0.8(V(1)-1)+0.2(V(2)-2) = -12$	$0.9(V(2)-2) = -18$

   Now $\pi_2(1) = b$ and $\pi_2(2) = a$

V_2	a	b
1	$0.8(V(2)-2)+0.2(V(1)-1) = -12$	$0.9(V(1)-1) = -9$
2	$0.8(V(1)-1)+0.2(V(2)-2) = -10.5$	$0.9(V(2)-2) = -11.25$

   Since there was no change, we have our policy: $\pi^*(1) = b$ and $\pi^*(2) = a$

   (c) If the initial policy has both states performing action a, then the expected reward will be $-\infty$ for both states and actions, and so policy iteration will not make any progress. Applying a discount factor solves this problem and allows the algorithm to come to the correct solution.

3. **(40 points)** There should be an action for each of the different possible moves that a knight can make (8). The transition table is deterministic since when a knight makes a move, he then always goes there. The entries in the rewards table should all be zero, since there is no positive or negative value to staying on the board for any amount of time. There should also be some positive award for going into the upper right corner, from one of the two states that has an action which goes there. Again, since there is no reason to prefer leaving the board in a small number of moves, there is no discount factor. After running through value or policy iteration, you should discover that all of the values for the states are the same, and in fact any valid move will be chosen with equal probability by the policy. This indicates that any position can eventually reach the goal state.

4. **(10 points)** In order to make it so that the knight reaches the goal as fast as possible, simply make it so that all of the transitions have a negative reward (except for the rewards on the transitions into the goal state). Alternatively, using a discount factor would also produce the same results. Using this formulation, it can be seen that the goal position can be reached in no more than six moves from any other position.