PROBLEM 2

(5.1) Consider the following MDP. It has states {0,1,2,3,4} with 4 as the starting state. In every state, you can take one of two possible actions: walk (W) or jump (J). The Walk action decreases the state by one. The Jump action has probability 0.5 of decreasing the state by two, and probability 0.5 of leaving the state unchanged. Actions will not decrease the state below zero: you will remain in state 0 no matter which action you will take (i.e., state 0 is a terminal state). Jumping in state 1 leads to state 0 with probability 0.5 and state 1 with probability 0.5. This definition leads to the following transition functions:

- For states $k \geq 1$, $T(k, W, k-1) = 1$
- For states $k \geq 2$, $T(k, J, k-2) = T(k, J, k) = 0.5$
- For state $k = 1$, $T(k, J, k-1) = T(k, J, k) = 0.5$

The reward gained when taking an action is the distance travelled squared: $R(s, a, s') = (s - s')^2$. The discount factor is $\gamma = 0.5$.

(a) **(4 points)** Suppose we initialize $Q_0(s, a) = 0$ for all $s \in \{0, 1, 2, 3, 4\}$ and $a \in \{J, W\}$. Evaluate the Q-values $Q_1(s, a)$ after exactly one Q-value iteration.

	s_0	s_1	s_2	s_3	s_4
J					
W					

(b) **(4 points)** What is the policy that we would derive from $Q_1(s, a)$?

s_1	s_2	s_3	s_4

(c) **(2 points)** What are the values $V_1(s)$ corresponding to $Q_1(s, a)$?

s_0	s_1	s_2	s_3	s_4

(d) **(4 points)** Will the policy change after the second iteration? If your answer is "yes", briefly describe how