Question 6

6.3

Lets start from C

$V(B) = 0.5 \qquad V(D) = 0.5$

Reward is 0 for either side.

So $V(C) = 0.5 + 0.1(0 + 0.5 - 0.5)$

$= 0.5$ estimate remains same

Suppose we went to D

$V(D) = 0.5 + 0.1(0 + 0.5 - 0.5)$

$= 0.5$, whether we went to C or E

estimate remains same

Suppose we went to E

reward is 1 for going to the terminal state and 0 otherwise.

If we went to the terminal state,

$V(E) = 0.5 + 0.1(1 + \overset{0}{\cancel{0.5}} - 0.5)$

$V[E] = 0.55$

But according to the figure, $V[E] = 0.5$

which means that this episode terminated at the bottom.

$V[A] = 0.5 + 0.1(0 + 0 - 0.5)$

$= 0.5 - 0.05$

$= 0.45$

Change is -0.05

6.4

It would be better to take a wider range of α values to determine the better learning method.

$$Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma Q(S',A$$

$$V(S) \leftarrow V(S) + \alpha[R + \gamma V(S') - V(S)]$$

α determines how much effect the reward would have on the update of $V(S)$.
With a smaller α, there would be less oscillations in the mse, but it might take infinite time for convergence. Both algorithms cannot be compared with each other against any fixed alpha

6.5

V(s) is affected by how large α is. Large α may cause V(s) to move towards its optimal value rapidly, but it may also overshoot the optimal value due to a large step size. Since this is essentially a stepwise movement towards an optima, it does not depend on initial values.