Question 6

6.3

Lets start from C

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

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

Yes. For example, ~~using $\alpha = 1$ would make Monte~~

MC at $\alpha = .01$ performs better than TD at $\alpha = 0.01$

~~For lower~~ For higher values of α, TD seems to perform better than MC. This could be due to the fact that individual rewards don't affect $V(S)$ as much as episode returns.

6.5

This may be happening due to the whole stochasticity of the process, and due to the updates happening before the final return is generated

$$V(s) \leftarrow V(s) + \alpha(R + V(s') - V(s))$$

$V(s)$ is affected more when α is larger and R is fluctuating as well. This may be why we observe the increase in error. With smaller α, the learning takes place more steadily, but slowly