Exercise 1 - Mars-Rover prediction

In this example we empirically compare the prediction abilities of TD(0) and constant-MC when applied to the following Markov reward process (MRP):

s_1	s_2	s_3	s_4	s_5	s_6	s_7
1						10

We will often use MRPs when focusing on the prediction problem, in which there is no need to distinguish the dynamics due to the environment from those due to the agent. In this MRP, all episodes start in the centre state, s_4, then proceed either left or right by one state on each step, with equal probability. Episodes terminate either on the extreme left or the extreme right (blue). When an episode terminates on the right, a reward of $+10$, when it terminates on the left a reward of 1 occurs; all other rewards are zero.

- Show a graph of the values learned after various numbers of episodes (say 0,1, 10, 100) on a single run of TD(0) with a step size $\alpha = 0.1$.
- Show a graph, where a comparison of the TD(0) and the First Visit MC learning curves (as a function of the number of episodes) for various values of α are drawn. As performance measure take the root mean-squared (RMS) error between the value function learned and a good estimate for the value function (can be found with various methods - name at least two), averaged over the five non-terminal states, then averaged over 100 runs. In all cases the approximate value function should be initialised to the intermediate value $V(s) = 5.5$, for all s. Which method performs better on this task?
- The obtained results of the previous exercise show a dependence on the value of the step-size parameter, α. Do you think the conclusions about which algorithm is better would be affected if by a wide range of α α values? Is there a different, fixed value of at which either algorithm performs significantly well? Why or why not?
- Produce another figure, where V(s) is drawn for each of the five non-terminal states individually. Now assume the parameter α decays from 0.5 over 250 episodes to 0.01. Compare TD(0) and the First Visit MC, what do you observe?