Assignment2: Problem

Consider grid-world example with termination:

XX	1	2	3
4	5	6	7
8	9	10	11
12	13	14	XX

Over the equiprobable policy, following policy was found to be greedy

XX	L	L	L/D
U	L/U	L/D	D
U	L/R	D/R	D
U/R	R	R	XX

Transition dynamic is now probabilistic:

(i) If say the state = 1, action is then A
Pr(0|1,a) = 0.7
Pr(2|1, a) = PR(5|1, a) = Pr(1|1,) = 0.1

(ii) If state = 5, action is then a
Pr (1|5, a) = Pr(4|5,a) = 0.4
Pr(9|5,a) = Pr(6|5,a) = 0.1

(iii) …

Apply Monte-carlo first visit method over 70 independent simulation runs to estimate Vpi(s) S = {1…14}

Randomize the initial state for each trajectory

Reward Structure = -1 for all states

= 0 for State XX

Plot for all States: 14 Coverage Plots (Vpi ^I (s) }

Tabulate Final values:

States	Vpi(s)
1	
…	
14	

Part2: Repeat the exercise for every visit case