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:

1. 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

1. 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

1. …

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

Results:

Tabulated for simulation run of 70 Episodes

Multi-visit

| 0.0 | -0.48 | -1.78 | -3.19

| -0.62 | -4.54 | -4.2 | -3.76

| -1.9 | -2.61 | -2.6 | -1.76

| -6.02 | -3.35 | -1.31 | 0.0

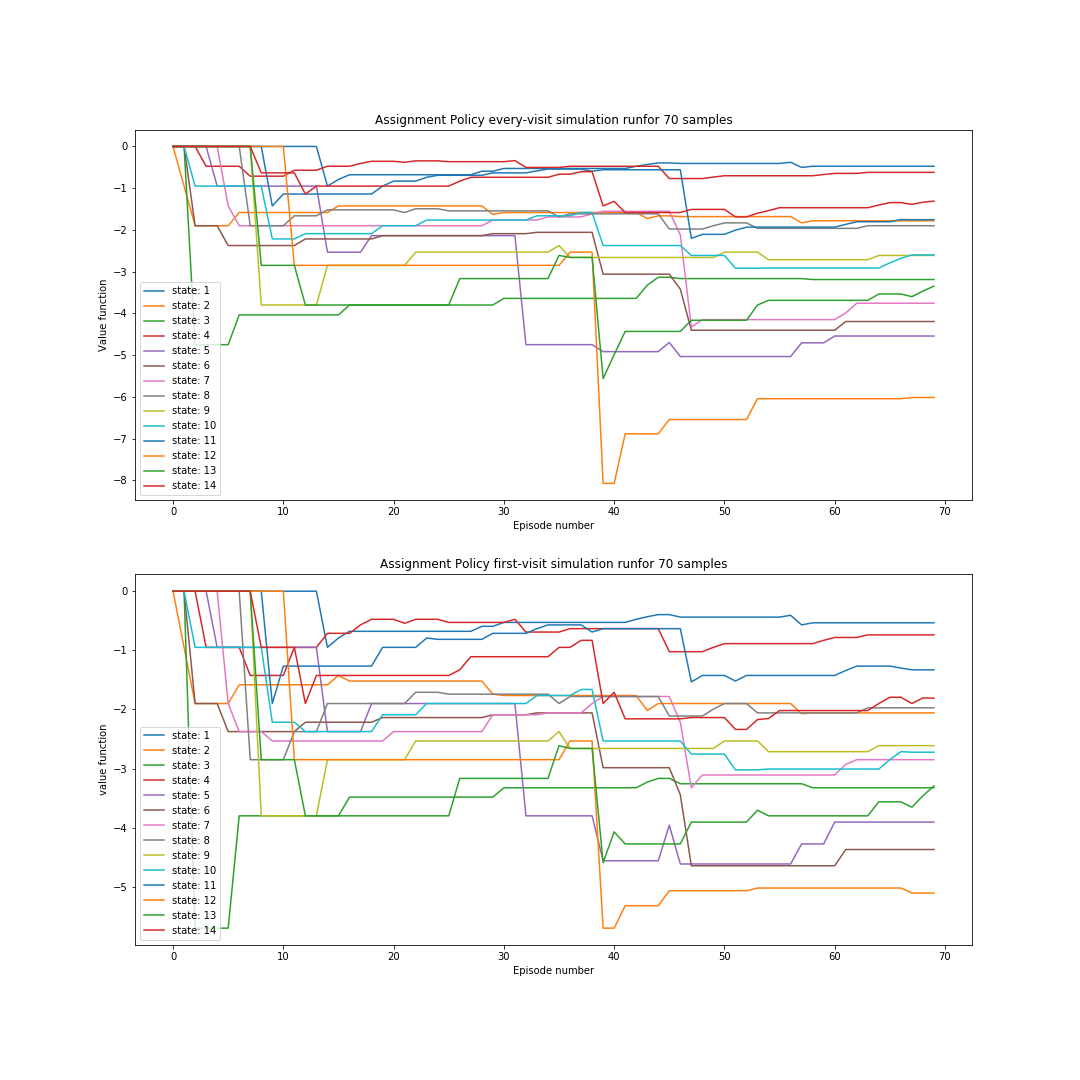
First visit

| 0.0 | -0.53 | -2.06 | -3.32

| -0.74 | -3.91 | -4.37 | -2.85

| -1.97 | -2.61 | -2.72 | -1.33

| -5.11 | -3.29 | -1.81 | 0.0



**Values for: Assignment Policy – 2800 Episodes**

Multi-visit

| 0.0 | -0.78 | -2.41 | -4.78

| -0.9 | -2.78 | -3.49 | -2.75

| -2.54 | -3.46 | -2.41 | -0.89

| -3.49 | -2.51 | -0.86 | 0.0

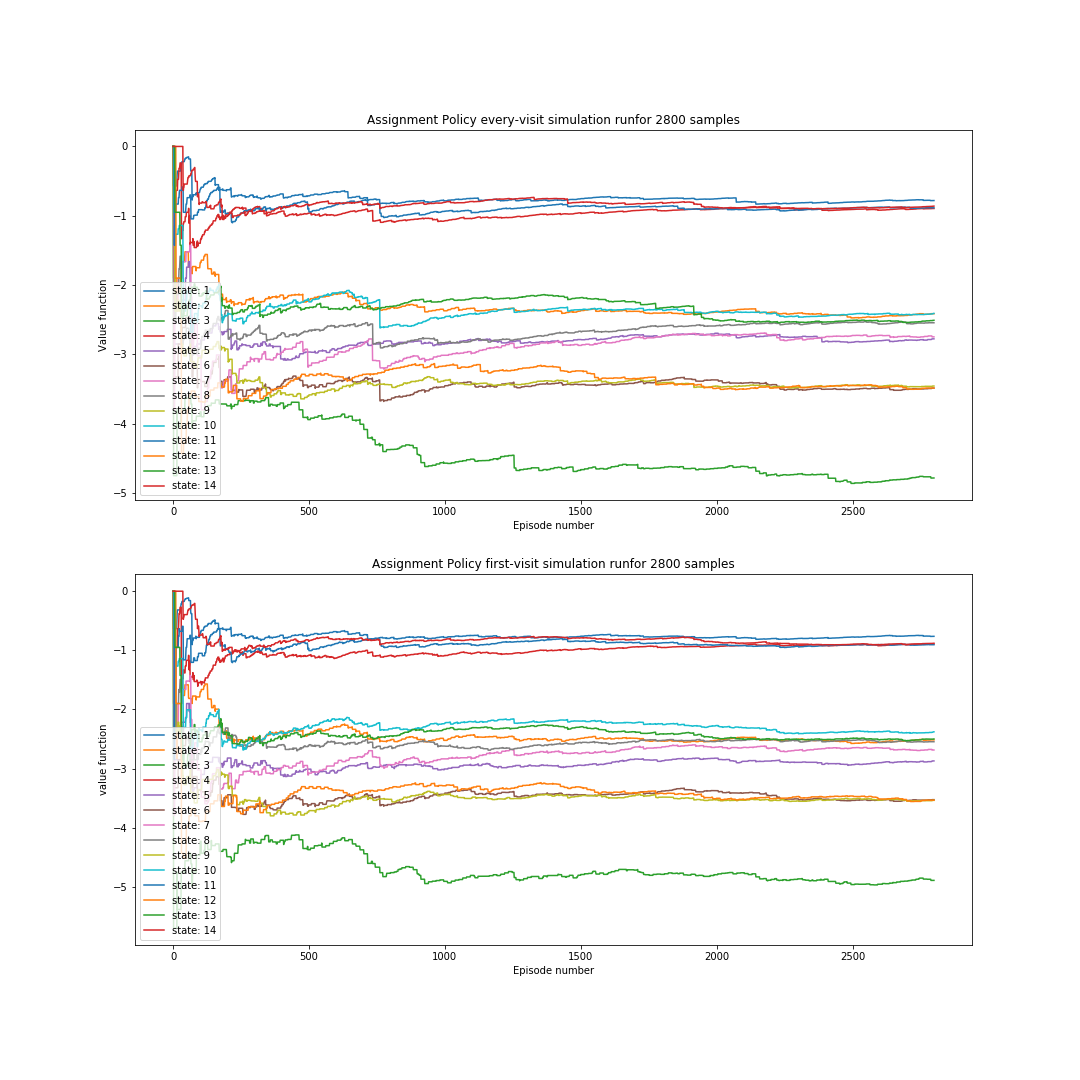
First visit

| 0.0 | -0.76 | -2.52 | -4.89

| -0.89 | -2.87 | -3.53 | -2.69

| -2.55 | -3.54 | -2.38 | -0.91

| -3.54 | -2.5 | -0.88 | 0.0



It can be clearly seen that with large number of 2800 episodes convergences of both the every-visit and first-visit is good

First visit seem to perform better.