For this assignment, I have modularised my code to do the following:

The main functions that can be called are both\_methods, first\_visit\_method and every\_visit\_method.

both\_methods allows us to see the difference in the values obtained for the same trajectories in the different Monte Carlo methods.

Internally, for each episode, these methods accumulate the state trajectories and rewards in different lists. The rewards list isn’t actually used as rewards are uniform in non-terminal states and because gamma is 1.

Inside each episode, there is an action choosing function that utilises the predefined optimal policy and a step taking function which adds the state transition probability asked for by the professor (0.7 chance of going left if action is left, 0.1 chance each of going in the other 3 directions respectively).

Finally, I’ve represented the gridworld in an array format as that seemed the best way in which to conduct the experiment.

Here are the results of 3 different experiments I ran:

First visit method: [0.0, -2.5172413793103448, -4.333333333333333, -5.083333333333333, -1.2903225806451613, -3.0, -3.8, -4.214285714285714, -3.076923076923077, -4.25, -3.1666666666666665, -1.8181818181818181, -3.2222222222222223, -2.6666666666666665, -1.5454545454545454, 0.0]

Every visit method: [0.0, -2.369565217391304, -4.371428571428571, -5.0588235294117645, -1.2903225806451613, -2.9705882352941178, -3.8, -3.9375, -3.2857142857142856, -4.454545454545454, -3.0, -1.7837837837837838, -3.210526315789474, -2.6153846153846154, -1.48, 0.0]

First visit method: [0.0, -1.5, -3.3846153846153846, -4.8125, -2.25, -4.142857142857143, -4.684210526315789, -3.4761904761904763, -3.2222222222222223, -4.2727272727272725, -3.6785714285714284, -1.6, -4.6, -3.142857142857143, -2.0, 0.0]

[0.0, -1.5517241379310345, -3.789473684210526, -4.7368421052631575, -2.206896551724138, -4.2105263157894735, -5.173913043478261, -3.2916666666666665, -3.3, -4.333333333333333, -3.75, -1.588235294117647, -4.6923076923076925, -3.066666666666667, -1.8863636363636365, 0.0]

Every visit method: [0.0, -1.9166666666666667, -3.3529411764705883, -4.333333333333333, -2.7333333333333334, -3.380952380952381, -4.083333333333333, -3.6956521739130435, -4.411764705882353, -4.153846153846154, -3.5555555555555554, -2.4444444444444446, -4.0, -3.6666666666666665, -1.9642857142857142, 0.0]

[0.0, -1.8863636363636365, -3.0, -4.238095238095238, -2.7142857142857144, -4.041666666666667, -4.0, -3.4827586206896552, -4.304347826086956, -4.153846153846154, -3.4, -2.3, -3.7857142857142856, -3.5357142857142856, -2.3947368421052633, 0.0]

What is interesting to see is that the results of both methods are similar in every run of the experiment above, with slight differences in the values.