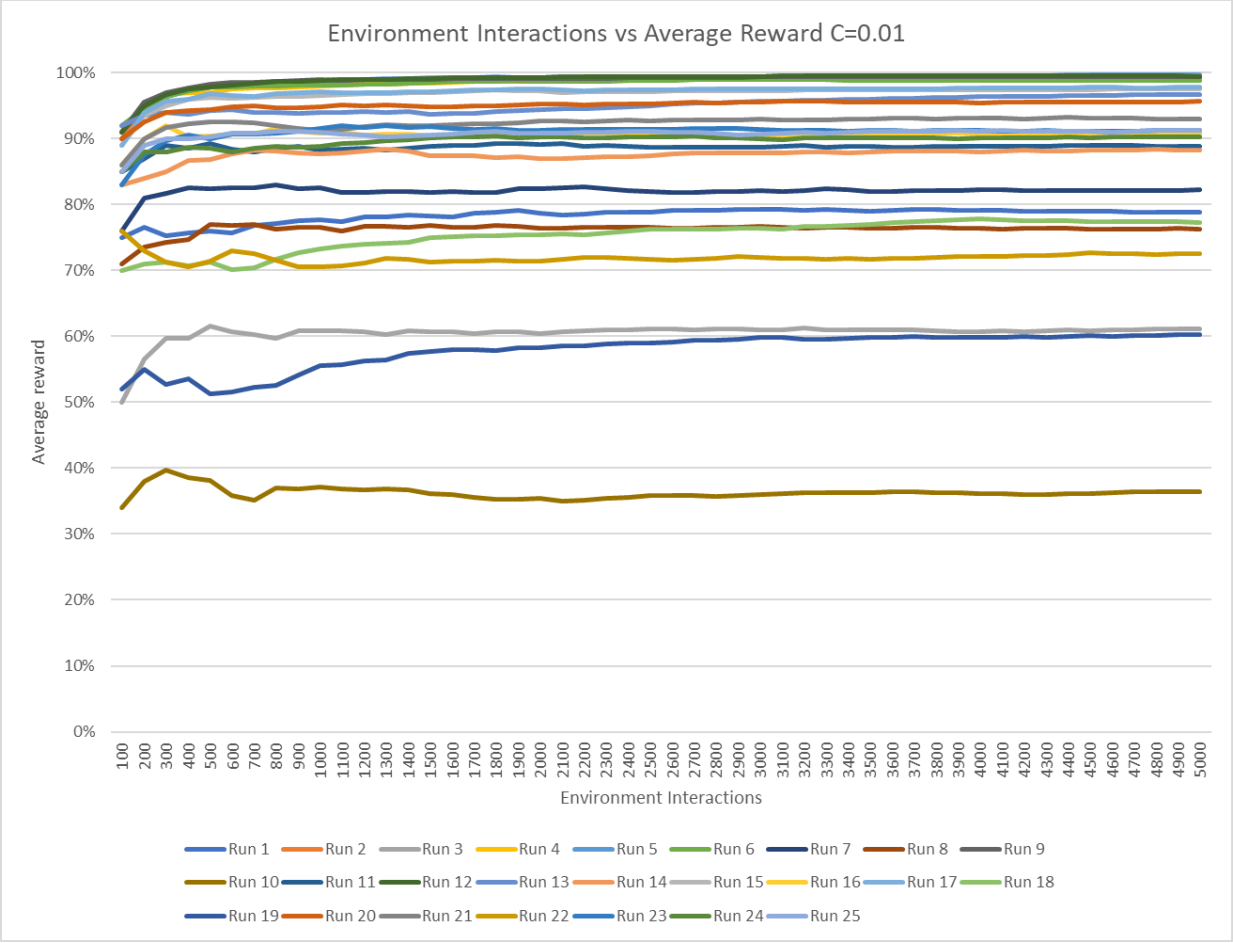
Part 1: UCB algorithm

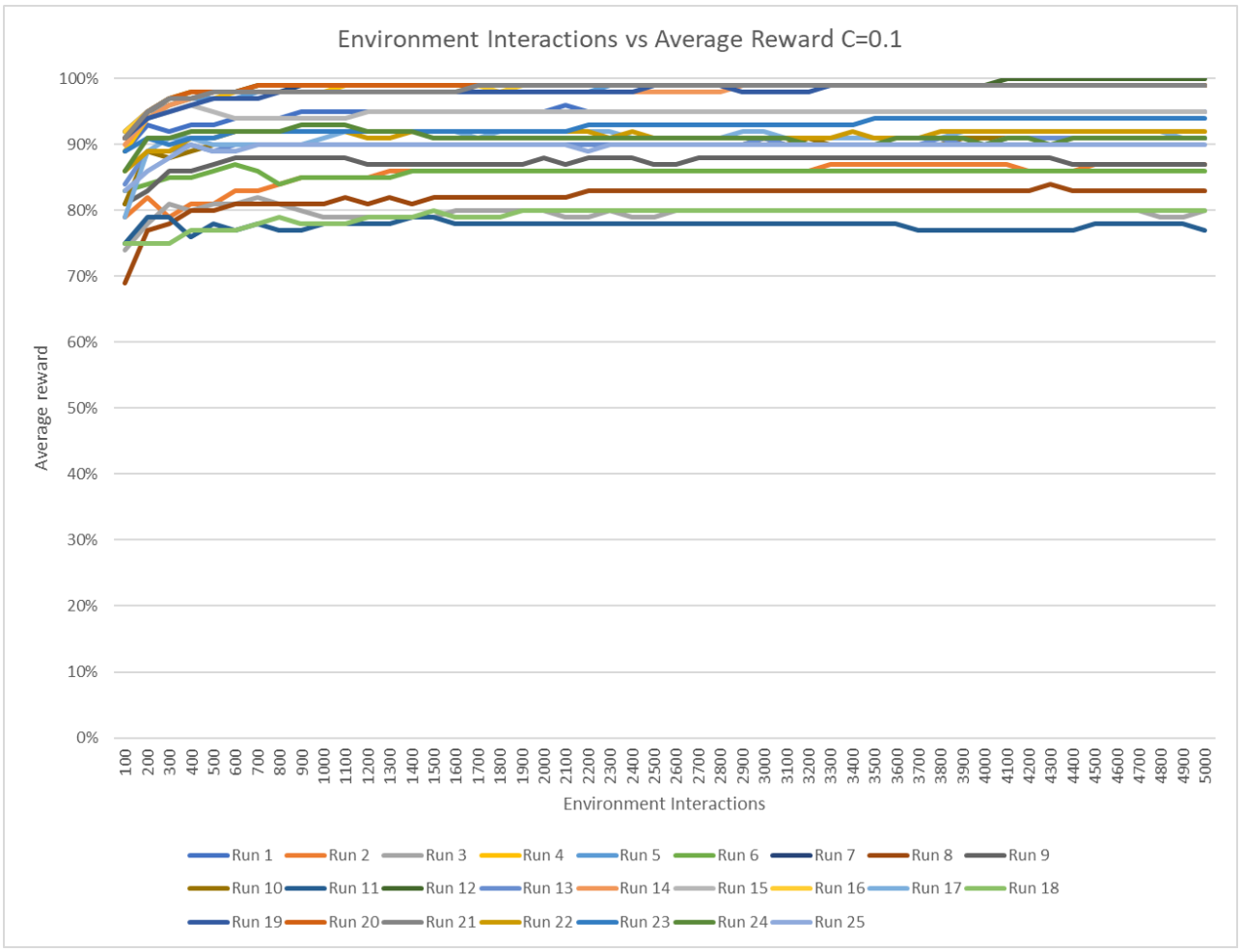
Our results provided in the graphs below show that the exploration constant C impacts average rewards over lifetime by changing how much the agent explores vs exploits.

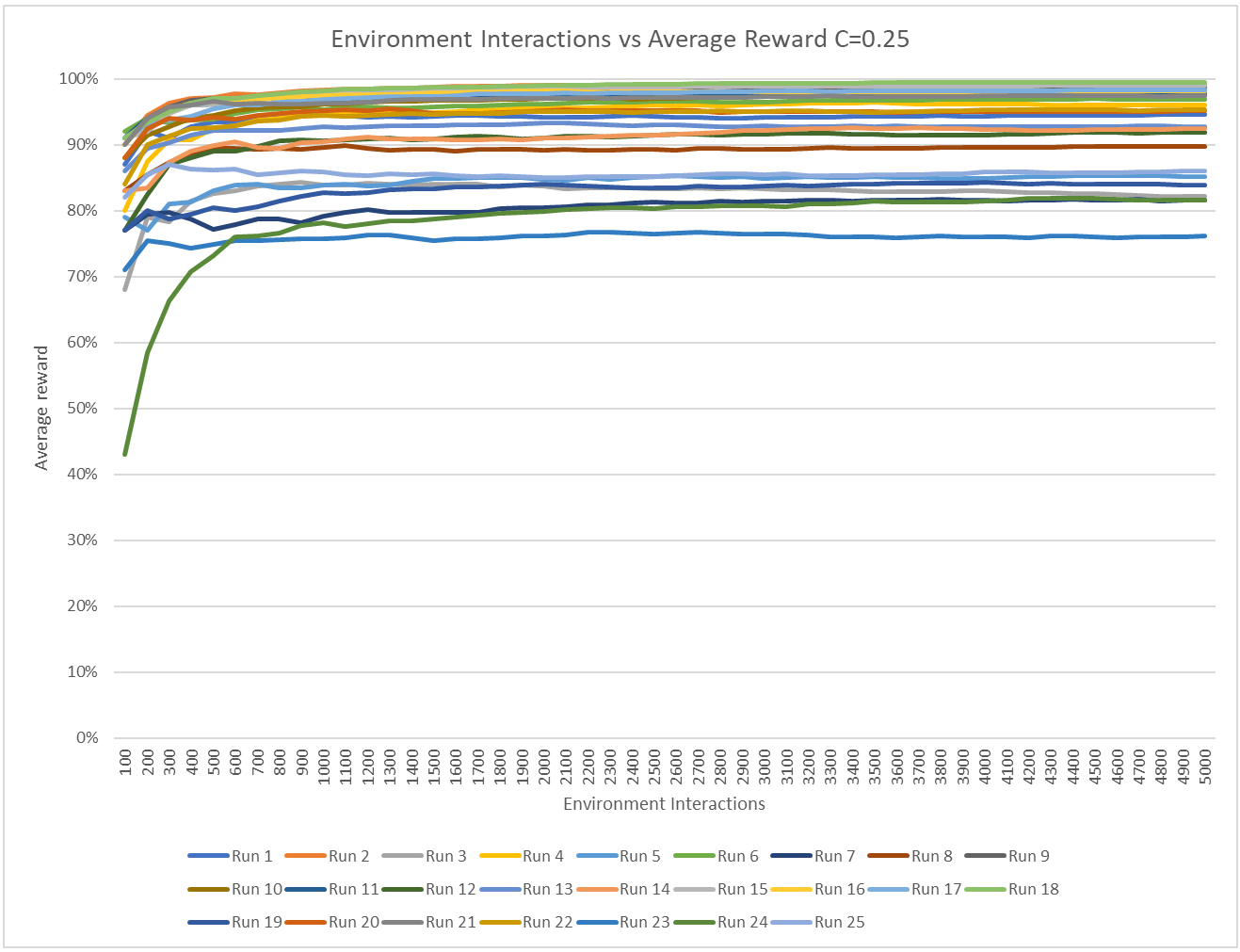
Comparing C=0.01 and C=0.1 we see that C=0.01 has the most runs with the highest average rewards but also has many outliers compared to C=0.1. Looking at the graph for C=0.01 we can see the 3 lowest runs grey, blue and brown where the average reward is around or below 60%. Since the exploration constant C is low, the agent chooses the action that maximizes Q without considering if Q is an accurate estimate of the true values q. For these runs we can say the agent is exploiting too much and not exploring enough.

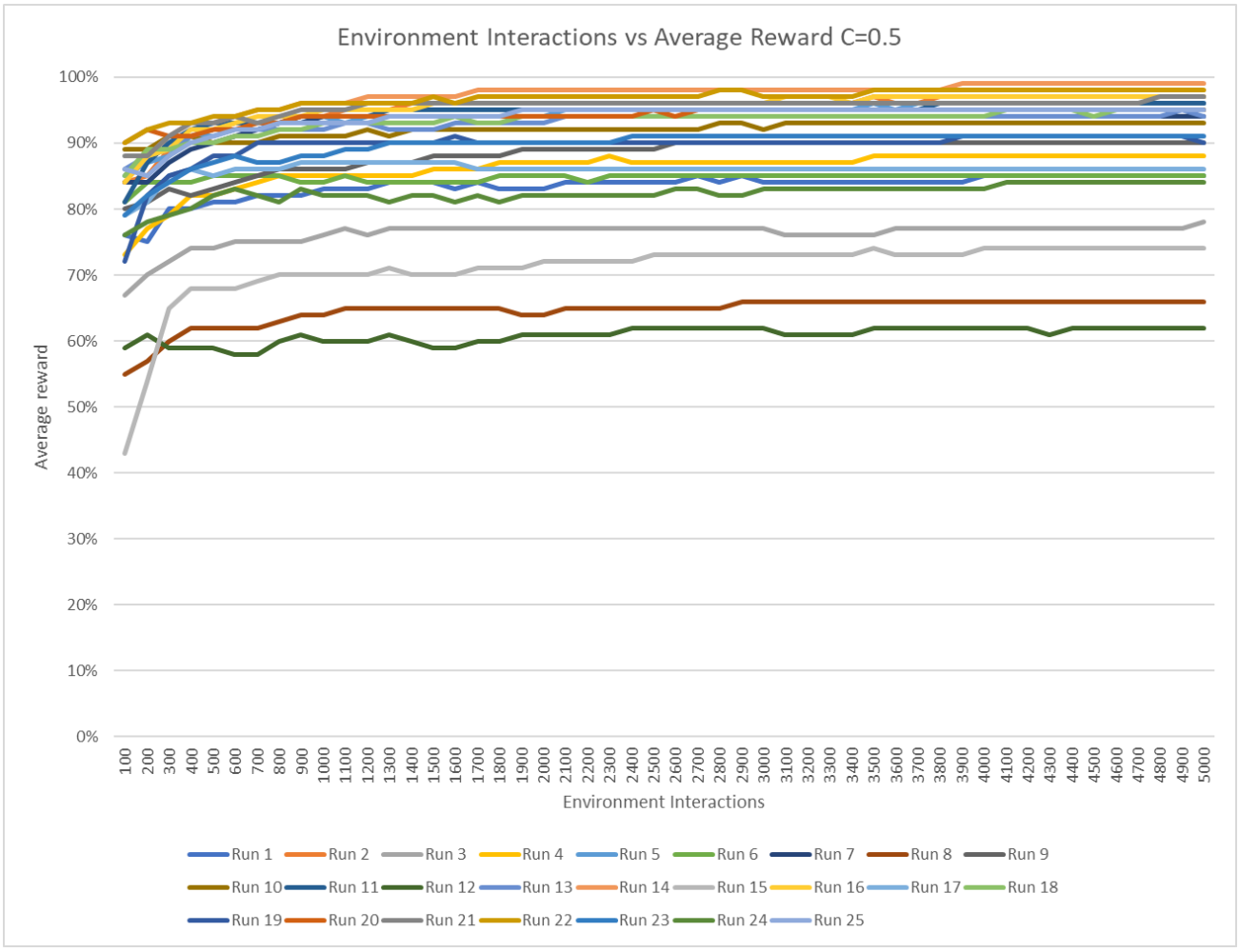
As C increases, we see runs with a wide range of average rewards over lifetime.

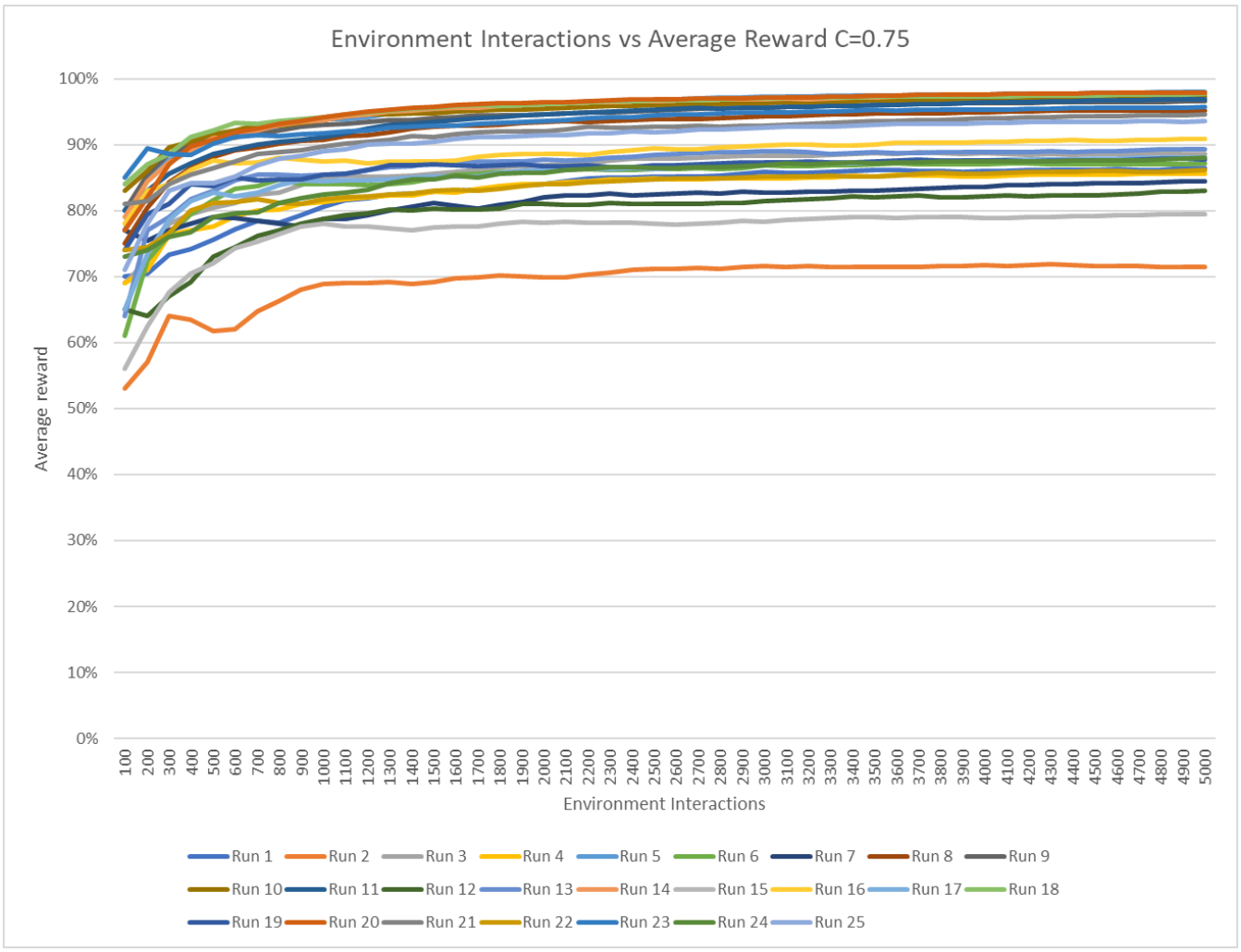
For C=1 we see it has the lowest number runs with high average rewards over lifetime because it explores too much and does not exploit what it has found to be the optimal action. Since C is high, even after finding the true optimal action the agent will try new actions that haven't been tried as much.

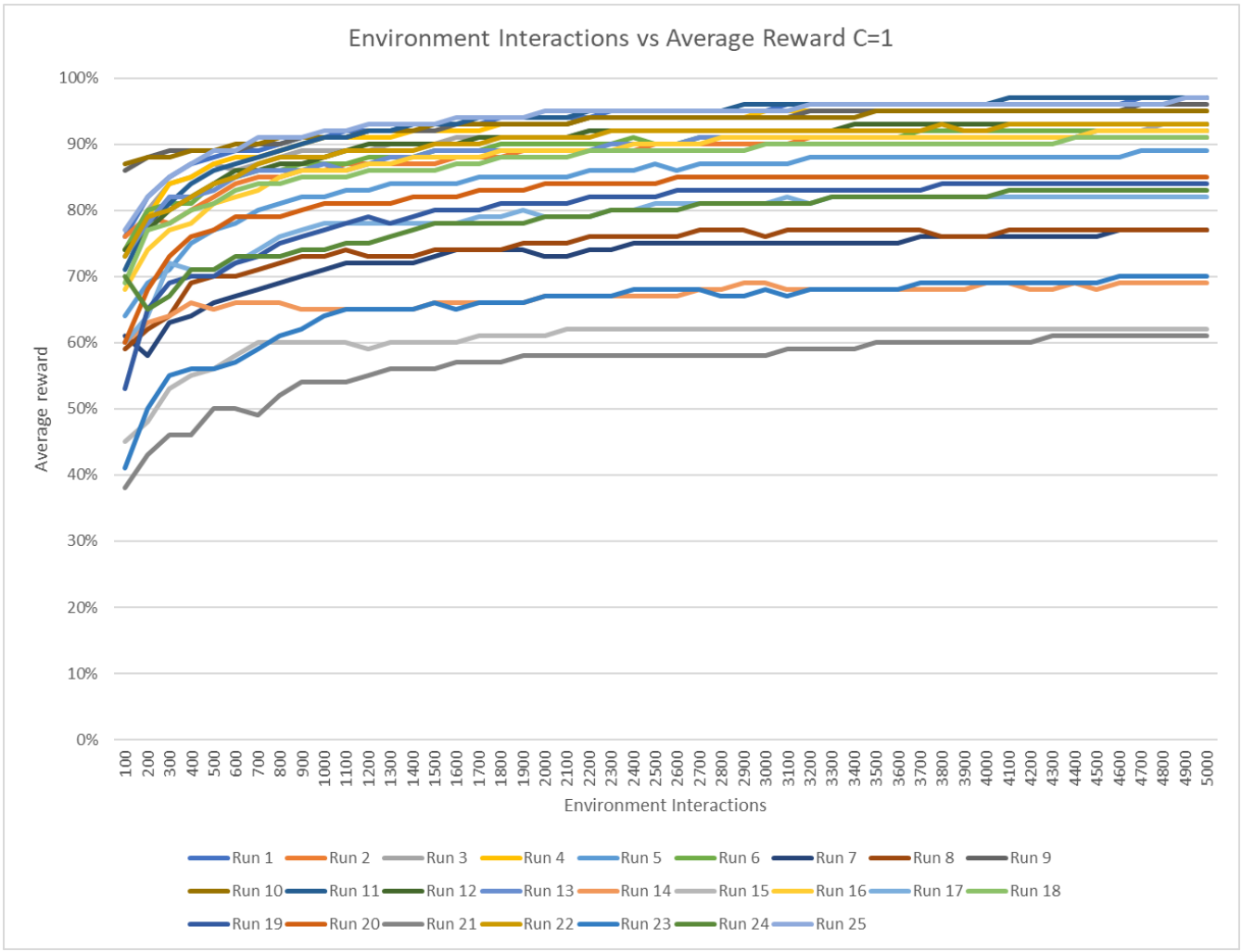



Environment Interactions vs Average Reward C=0.1
Average reward
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
2400
2500
2600
2700
2800
2900
3000
3100
3200
3300
3400
3500
3600
3700
3800
3900
4000
4100
4200
4300
4400
4500
4600
4700
4800
4900
5000
Environment Interactions
Run 1
Run 2
Run 3
Run 4
Run 5
Run 6
Run 7
Run 8
Run 9
Run 10
Run 11
Run 12
Run 13
Run 14
Run 15
Run 16
Run 17
Run 18
Run 19
Run 20
Run 21
Run 22
Run 23
Run 24
Run 25


Environment Interactions vs Average Reward C=0.25
Average reward
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000 2100 2200 2300 2400 2500 2600 2700 2800 2900 3000 3100 3200 3300 3400 3500 3600 3700 3800 3900 4000 4100 4200 4300 4400 4500 4600 4700 4800 4900 5000
Environment Interactions
Run 1 Run 2 Run 3 Run 4 Run 5 Run 6 Run 7 Run 8 Run 9
Run 10 Run 11 Run 12 Run 13 Run 14 Run 15 Run 16 Run 17 Run 18
Run 19 Run 20 Run 21 Run 22 Run 23 Run 24 Run 25


Environment Interactions vs Average Reward C=0.5
Average reward
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000 2100 2200 2300 2400 2500 2600 2700 2800 2900 3000 3100 3200 3300 3400 3500 3600 3700 3800 3900 4000 4100 4200 4300 4400 4500 4600 4700 4800 4900 5000
Environment Interactions
Run 1
Run 2
Run 3
Run 4
Run 5
Run 6
Run 7
Run 8
Run 9
Run 10
Run 11
Run 12
Run 13
Run 14
Run 15
Run 16
Run 17
Run 18
Run 19
Run 20
Run 21
Run 22
Run 23
Run 24
Run 25


Environment Interactions vs Average Reward C=0.75
Average reward
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000 2100 2200 2300 2400 2500 2600 2700 2800 2900 3000 3100 3200 3300 3400 3500 3600 3700 3800 3900 4000 4100 4200 4300 4400 4500 4600 4700 4800 4900 5000
Environment Interactions
Run 1 Run 2 Run 3 Run 4 Run 5 Run 6 Run 7 Run 8 Run 9
Run 10 Run 11 Run 12 Run 13 Run 14 Run 15 Run 16 Run 17 Run 18
Run 19 Run 20 Run 21 Run 22 Run 23 Run 24 Run 25


Environment Interactions vs Average Reward C=1
Average reward
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000 2100 2200 2300 2400 2500 2600 2700 2800 2900 3000 3100 3200 3300 3400 3500 3600 3700 3800 3900 4000 4100 4200 4300 4400 4500 4600 4700 4800 4900 5000
Environment Interactions
Run 1 Run 2 Run 3 Run 4 Run 5 Run 6 Run 7 Run 8 Run 9
Run 10 Run 11 Run 12 Run 13 Run 14 Run 15 Run 16 Run 17 Run 18
Run 19 Run 20 Run 21 Run 22 Run 23 Run 24 Run 25