1 Empirical Results

1.1 *Parr's and Russell's* 4 × 3 *Maze*

Number of distinct policies: 256

1.1.1 Algorithm MCESP-SAA

Number of runs=45, Epsilon=0.001, Gamma=1.0, Number of samples=100000, Maximum length of trajectory=400, Maximum number of trajectories=30000

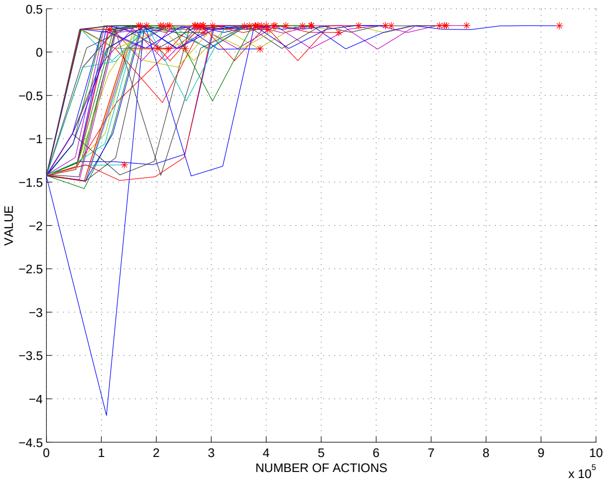

Figure 1: Learning curves for multiple runs (red stars indicate the end of a run)

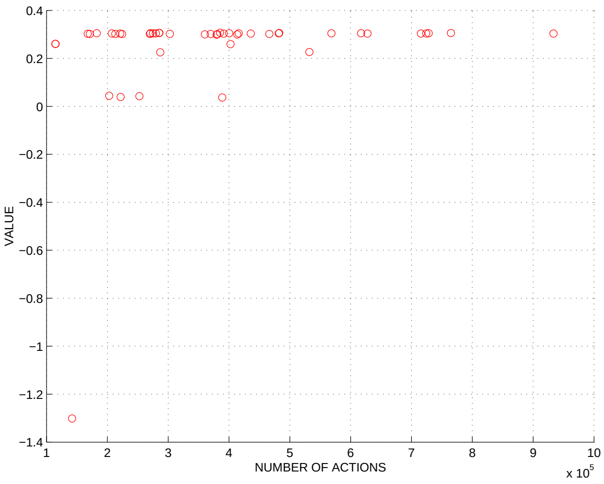

Figure 2: Value of the final policy for multiple runs

The optimal policy has a value of 0.303. Out of the 45 runs, 35 converged to the optimal policy. The mean over 45 runs of the *median* number of steps to goal of the final policy was 12. While the mean over 45 runs of the *average* number of steps to goal of the final policy was 15.

1.1.2 Algorithm MCESP-CE

b=0.1, p=0.6
The other parameters remain unchanged (from the values used for MCESP-SAA)

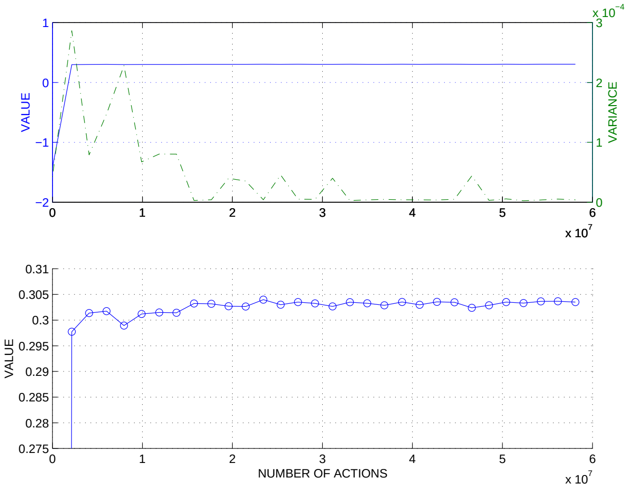

Figure 3: Average learning curve and variance curve

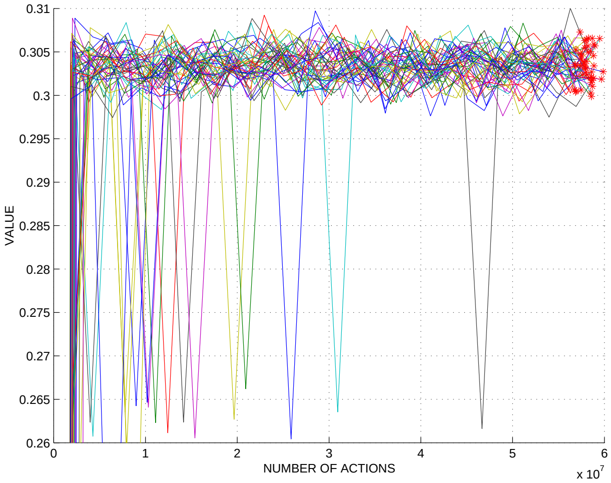

Figure 4: Learning curves for multiple runs

All 45 runs converged to the optimal policy. The mean over 45 runs of the *median/mean* number of steps to goal of the final policy was 12/15

1.1.3 Algorithm Sarsa(0.9)

Number of runs=45, Initial Epsilon=0.2, Final Epsilon=0.0, Rate of epsilon decline=0.000001, Gamma=1.0, Number of samples=100000, Maximum length of trajectory=400, Maximum number of trajectories=30000

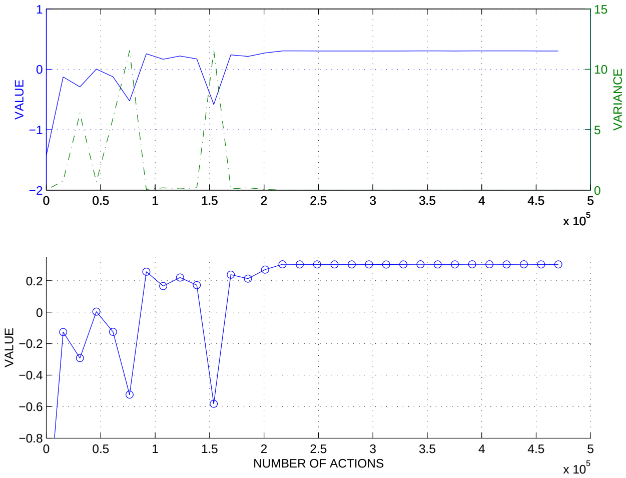

Figure 5: Average learning curve for multiple runs

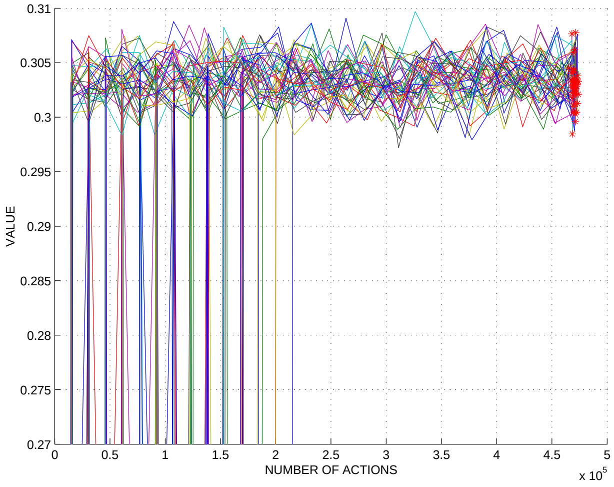

Figure 6: Learning curves for multiple runs

The exploration rate was initialised at 0.2 and reduced by 10^{-5} with the execution of every action. So after 2×10^5 actions were made the agent stopped exploring. All 45 runs converged to the optimal policy. The mean over 45 runs of the *median/mean* number of steps to goal of the final policy was 12/15

The number of trajectories for each run of MCESP and Sarsa were the same (30,000), but MCESP learns slowly initially and this results in a difference in the rates of learning of the two algorithms (MCESP-CE converges after approximately 10^7 actions whereas Sarsa converges after only 10^5 actions).

1.2 *Sutton's Grid World*

Number of distinct policies: 1.15×10^{18}

1.2.1 Algorithm MCESP-SAA

Number of runs=45, Epsilon=0.001, Gamma=0.95, Number of samples=25,000 , Maximum length of trajectory=1000, Maximum number of trajectories=70000

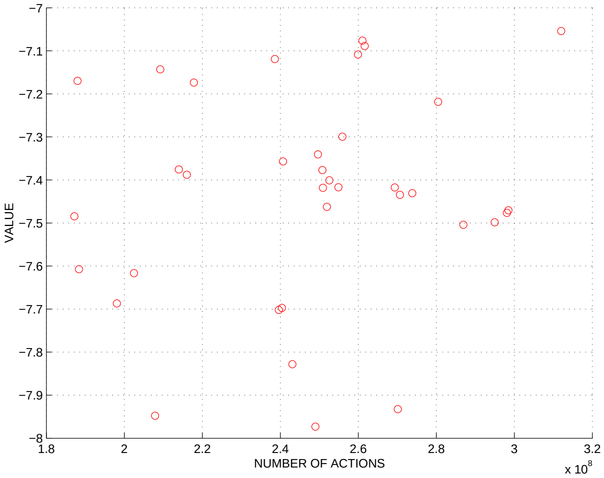

Figure 7: Value of the final policy for 45 runs

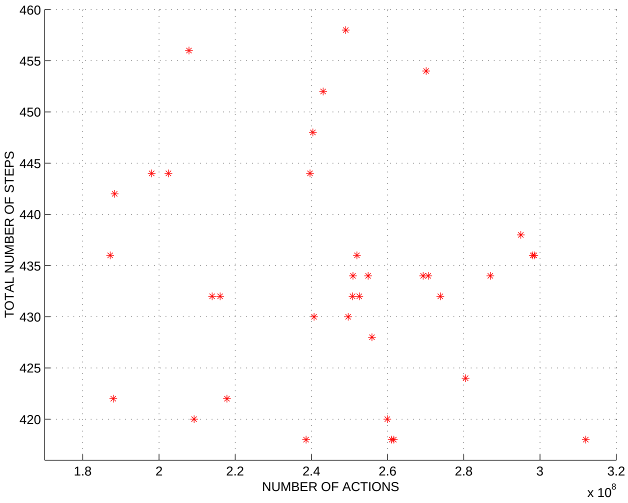

Figure 8: Total number of steps required to reach the goal starting from every possible non-goal state (Optimal number for memoryless policies = 416)

None of the runs converged to the optimal policy. However, 5 runs did come very close, converging to a policy that took 418 steps to reach the goal starting from every possible state. The mean over 45 runs of the *median/mean* number of steps to goal of the final policy was 9/9.1

1.2.2 Gamma=1.0

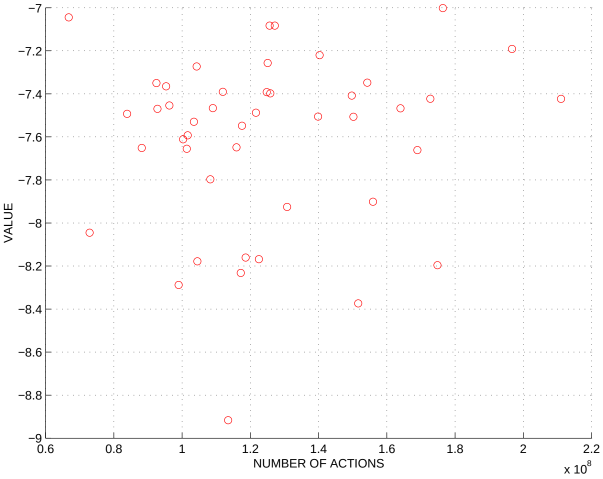

Figure 9: Value of the final policy for 45 runs

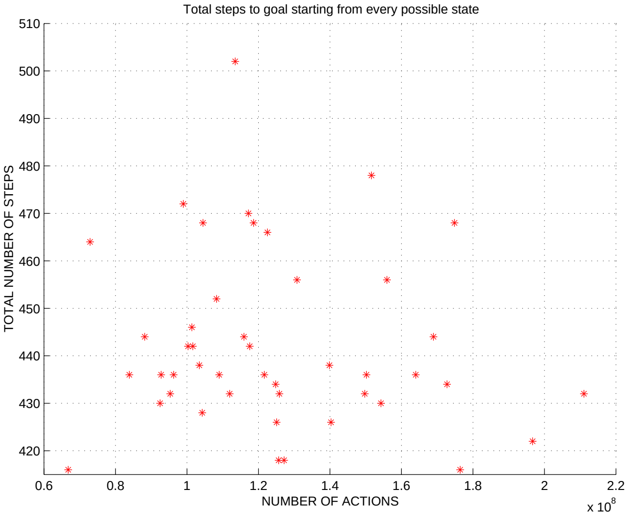

Figure 10: Total number of steps required to reach the goal starting from every possible non-goal state

Two of the runs converged to the optimal policy. The mean over 45 runs of the *median/mean* number of steps to goal of the final policy was 9/9.2

1.2.3 Algorithm MCESP-CE

Gamma=0.95, b=0.1, p=0.6

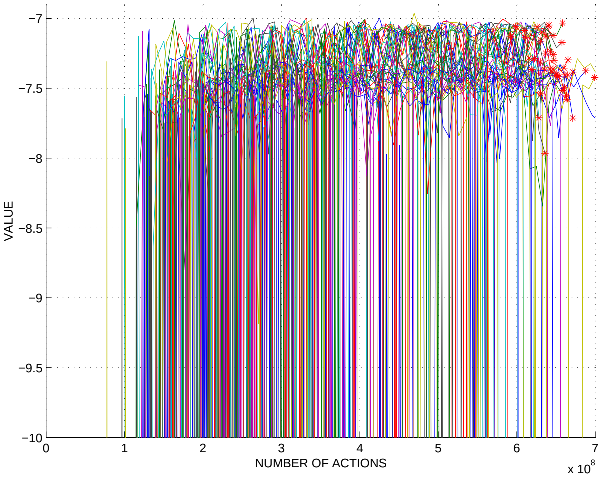

Figure 11: Learning Curve

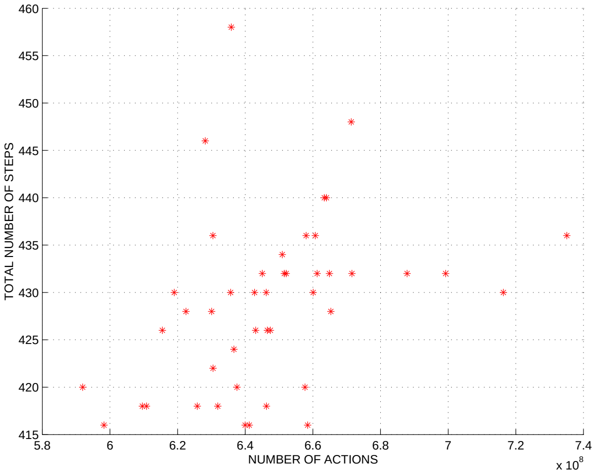

Figure 12: Total number of steps required to reach the goal starting from every possible non-goal state (Optimal number for memoryless policies = 416)

Two of the policies were global optima. Four of the runs converged to the optimal policy. The mean over 45 runs of the *median/mean* number of steps to goal of the final policy was 9.2/9.3

1.2.4 Gamma=1.0

Two were global optima. Nine of the runs converged to the optimal policy. The mean over 45 runs of the *median/mean* number of steps to goal of the final policy was 9.2/9.3

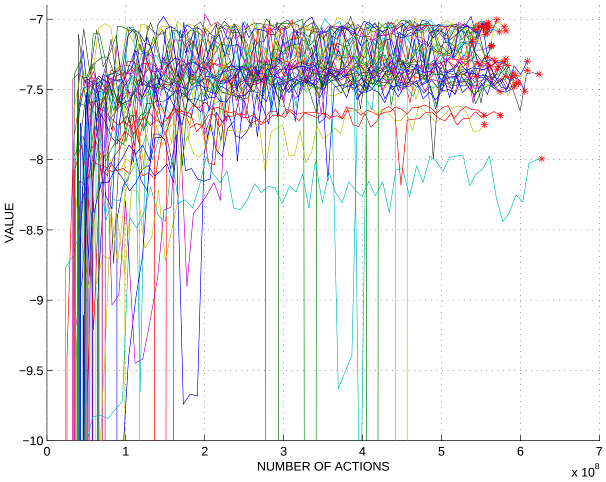

Figure 13: Learning Curves

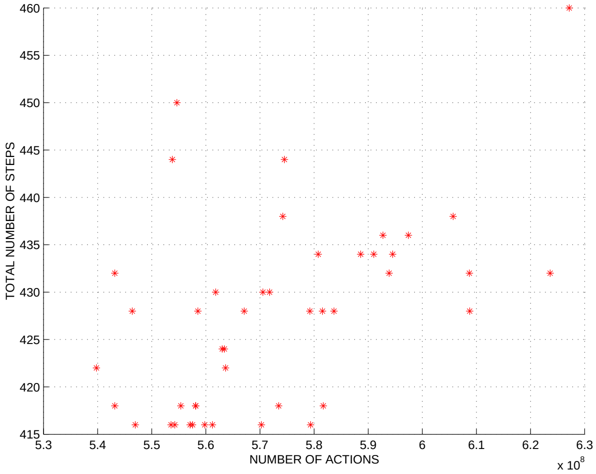

Figure 14: Total number of steps required to reach the goal starting from every possible non-goal state

1.2.5 Gamma=0.95 and No Averaging Zeroes

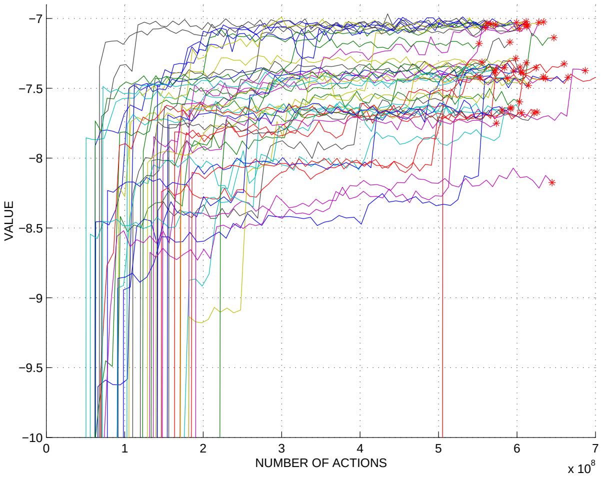

Figure 15: Learning Curves

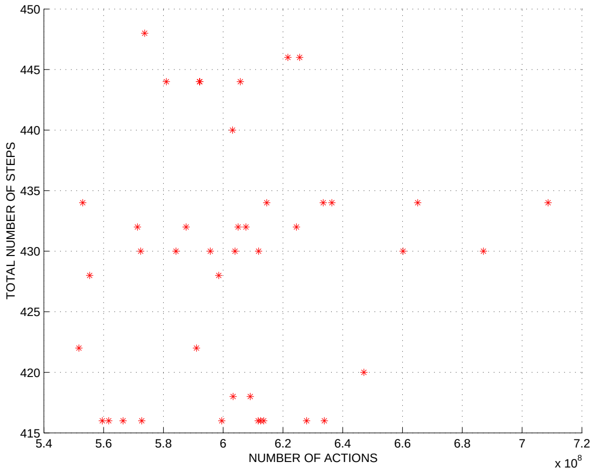

Figure 16: Total number of steps required to reach the goal starting from every possible non-goal state

Ten of the runs converged to the optimal policy. The mean over 45 runs of the *median/mean* number of steps to goal of the final policy was 9.2/9.3

1.2.6 Gamma=1.0 and No Averaging Zeroes

One of the runs converged to the optimal policy. The mean over 45 runs of the *median/mean* number of steps to goal of the final policy was 10/10.1

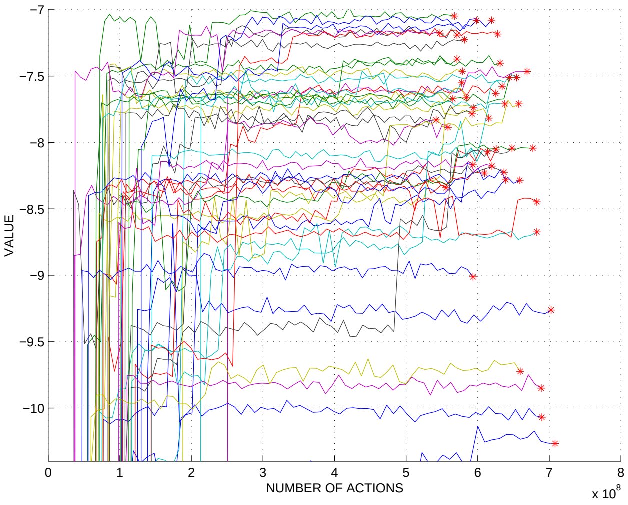

Figure 17: Learning Curves

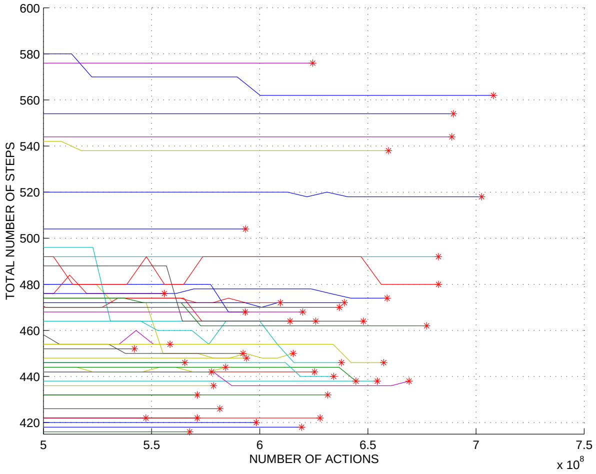

Figure 18: Total number of steps required to reach the goal starting from every possible non-goal state

1.2.7 Algorithm Sarsa(0.9)

```
Number of runs=45, Initial Epsilon=0.2,
Final Epsilon=0.0, Rate of epsilon
decline=0.000001, Gamma=0.95, Number
of samples=25000, Maximum length of
trajectory=1000, Maximum number of
trajectories=70000
```

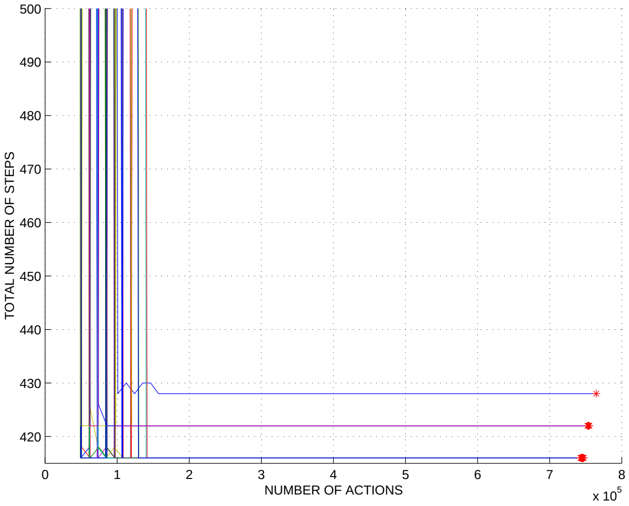

Figure 19: Total number of steps required to reach the goal starting from every possible non-goal state

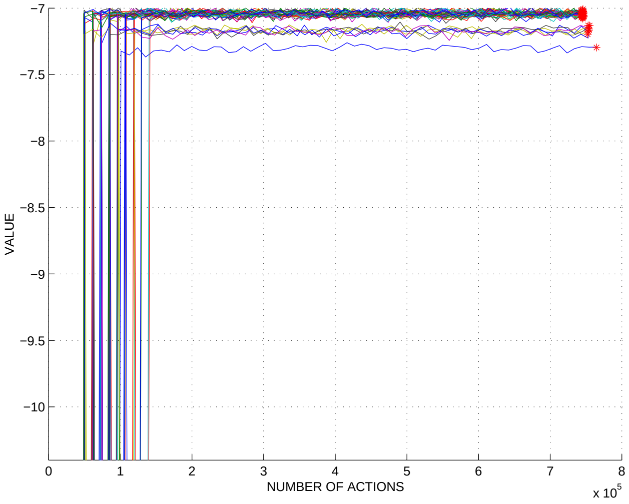

Figure 20: Learning Curves

Thirty five of the runs converged to the optimal policy. The mean over 45 runs of the *median/mean* number of steps to goal of the final policy was 9/9

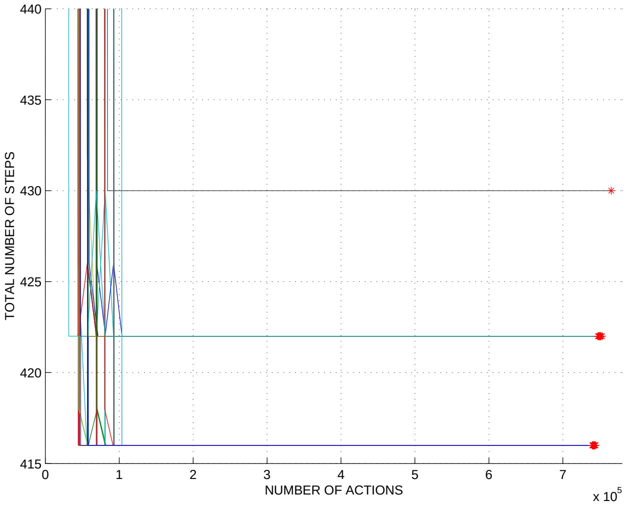

Figure 21: Total number of steps required to reach the goal starting from every possible non-goal state

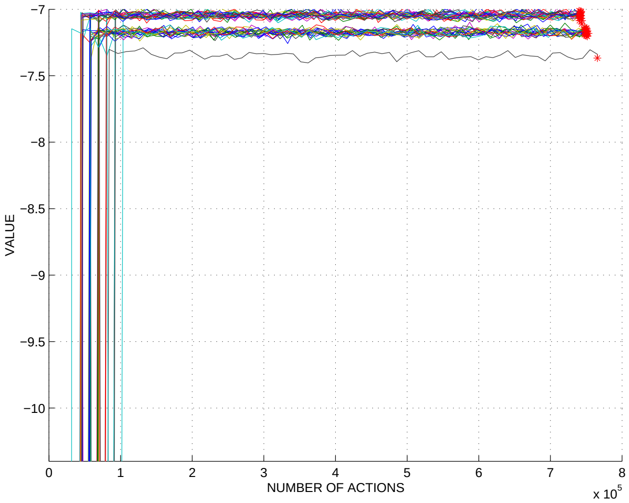

Figure 22: Learning Curves

1.2.8 Gamma=1.0

Twenty four of the runs converged to the optimal policy. The mean over 45 runs of the *median/mean* number of steps to goal of the final policy was 9/9

1.3 *Littman. Cassandra, and Kaelbling's* 89 *State Office World*

Number of distinct policies: 7.6×10^{11}

1.3.1 Algorithm MCESP-SAA

Number of runs=45, Epsilon=0.0001, Gamma=0.95, Number of samples=100000, Maximum length of trajectory=250, Maximum number of trajectories=40000

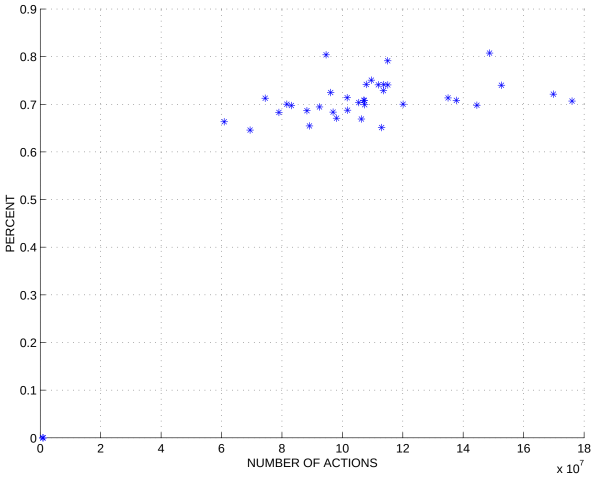

Figure 23: Value of the final policy generated over multiple runs (or the percent of trials that reach the goal state)

The mean over 45 runs of the *median/mean* number of steps to goal of the final policy was 32/62

1.3.2 Gamma=1.0

The mean over 45 runs of the *median/mean* number of steps to goal of the final policy was 60/79

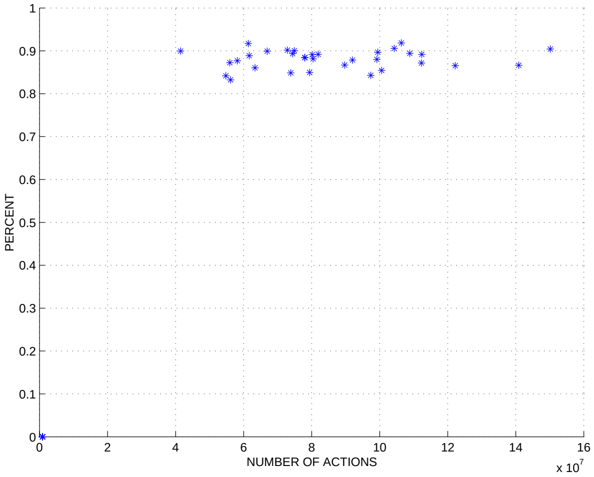

Figure 24: Value of the final policy generated over multiple runs (or the percent of trials that reach the goal state)

1.3.3 Algorithm MCESP-CE

Gamma=0.95, b=0.1, p=0.6

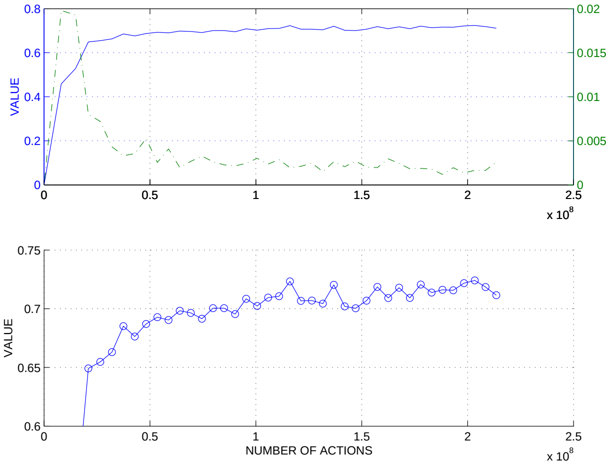

Figure 25: Average learning and variance curve for multiple runs

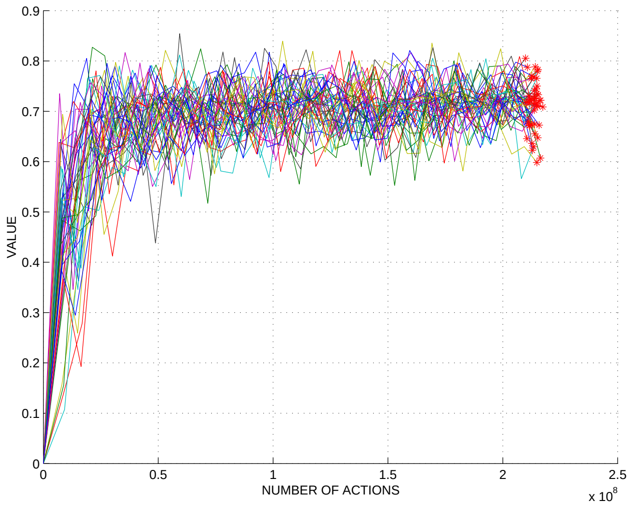

Figure 26: Learning curves (or the percent of trials that reach goal state) for multiple runs

Only 71% of trials reach a terminal state. The mean over 45 runs of the *median/mean* number of steps to goal of the final policy was 34/64

1.3.4 Gamma=1.0, b=0.1, p=0.6

Only 89% of trials reach a terminal state. The mean over 45 runs of the *median/mean* number of steps to goal of the final policy was 62/79

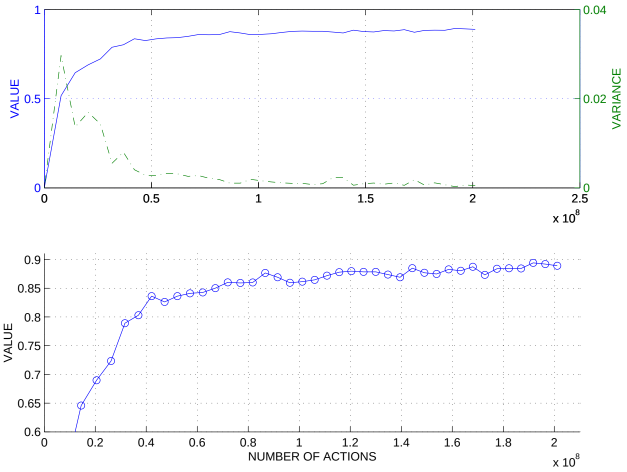

Figure 27: Average learning and variance curve for multiple runs

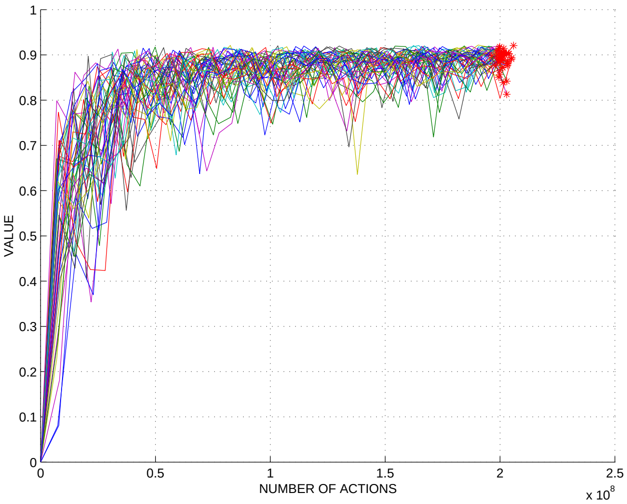

Figure 28: Learning curves (or the percent of trials that reach goal state) for multiple runs

1.3.5 Algorithm Sarsa(0.9)

Number of runs=45, Initial Epsilon=0.2, Final Epsilon=0.0, Rate of epsilon decline=0.0000001/0.0000001, Gamma=1.0, Number of samples=100000, Maximum length of trajectory=250, Maximum number of trajectories=30000

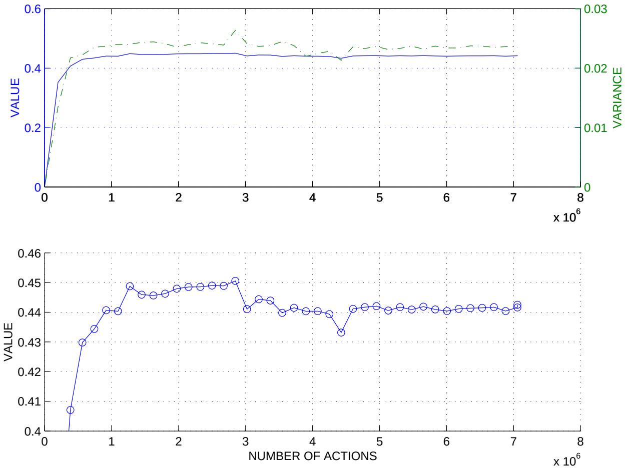

Figure 29: Average learning curve for multiple runs

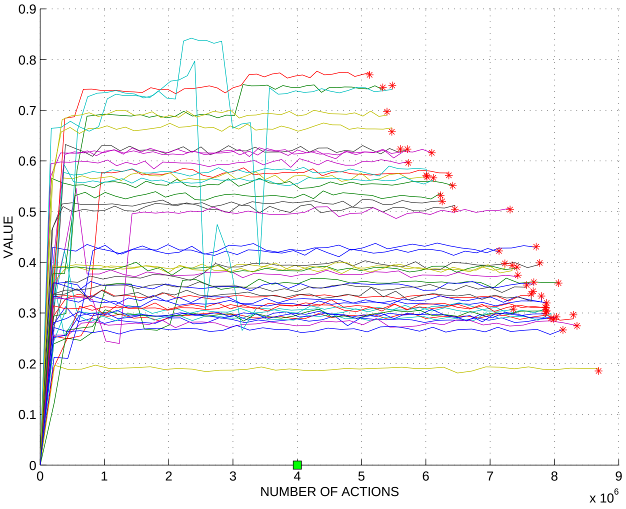

Figure 30: Learning curves for multiple runs (the green square marks the point at which exploration stops)

Only 45% of trials reach a terminal state. The mean over 45 runs of the *median/mean* number of steps to goal of the final policy was 57/80

For both values of epsilon we see that there is convergence much before exploration stops.

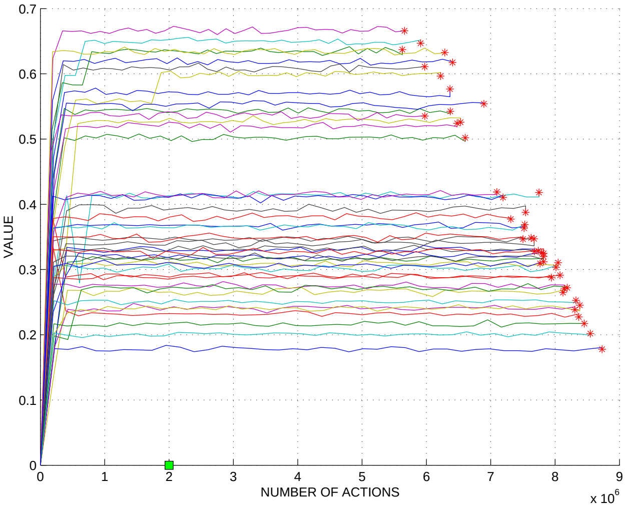

Figure 31: Learning curves for multiple runs. Notice that exploration stops earlier (bigger epsilon decline) than Figure 30.

1.3.6 Gamma=0.95

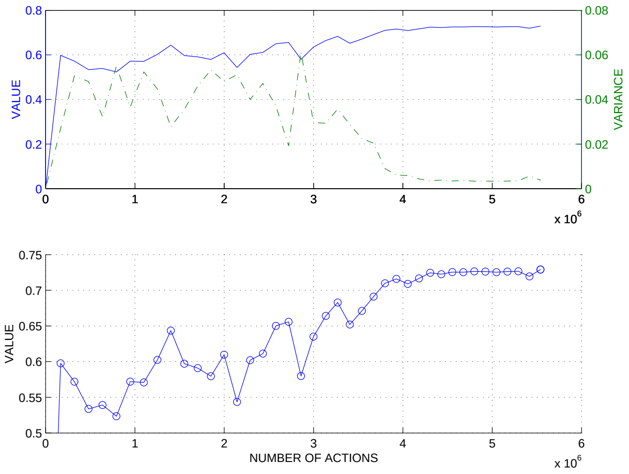

Figure 32: Average learning curve for multiple runs

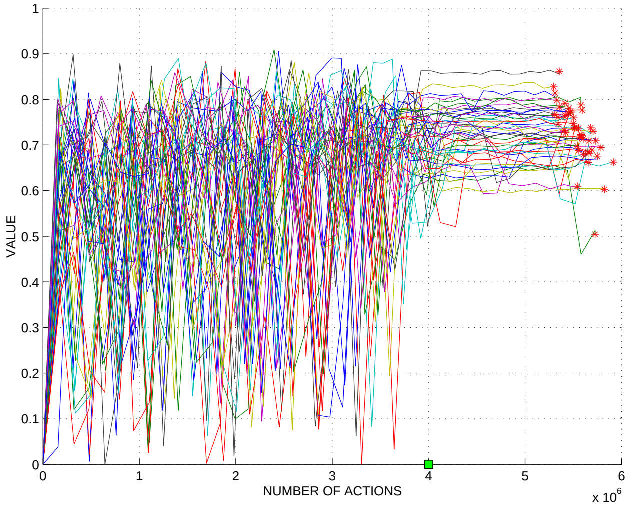

Figure 33: Learning curves for multiple runs (the green square marks the point at which exploration stops)

Only 72% of trials reach a terminal state. The mean over 45 runs of the *median/mean* number of steps to goal of the final policy was 50/75

There is no clear increase or decrease in the values of the final policies generated by Sarsa when epsilon is changed (see Figures 33 and 34).

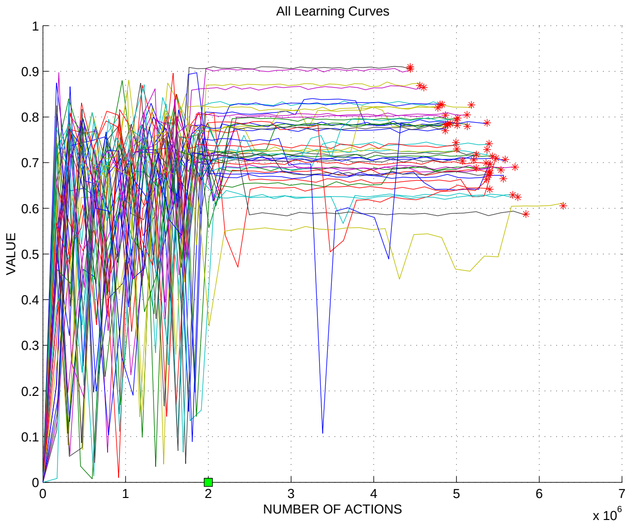

Figure 34: Learning curves for multiple runs. Notice that exploration stops earlier (bigger epsilon decline) than Figure 33.

1.4 *Aircraft Identification*

Number of distinct policies: *i*

1.4.1 Algorithm MCESP-SAA

Number of runs=45, Epsilon=0.0001, Gamma=0.95, Number of samples=10000, Maximum length of trajectory=1500, Maximum number of trajectories=80000

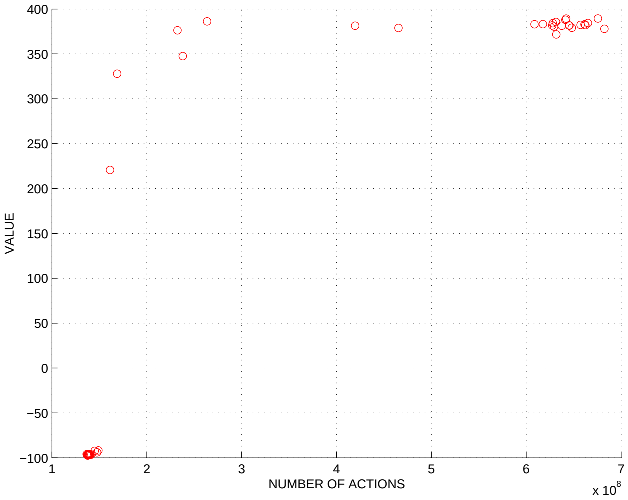

Figure 35: Value of the final policy generated over multiple runs

Only 99% of trials reach a terminal state. The maximum over 45 runs of the *median/mean* number of steps to terminal state of the final policy was 265/360

1.4.2 Gamma=1.0

Only 97.8% of trials reach a terminal state. The maximum over 45 runs of the *median/mean* number of steps to terminal state of the final policy was 255/360

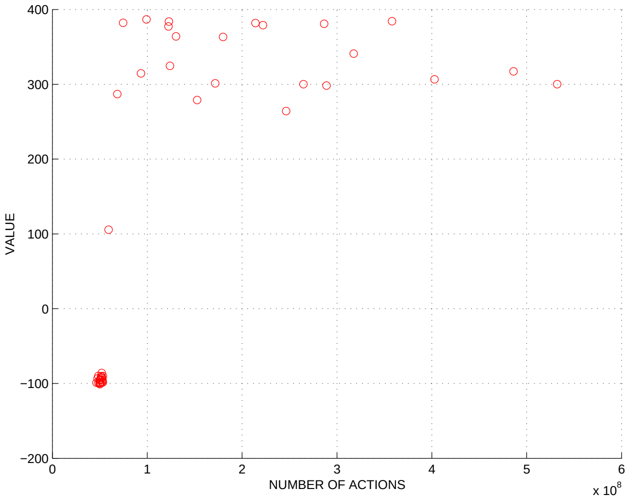

Figure 36: Value of the final policy generated over multiple runs

1.4.3 Algorithm MCESP-SAA (No Averaging Zeroes)

Gamma=0.95

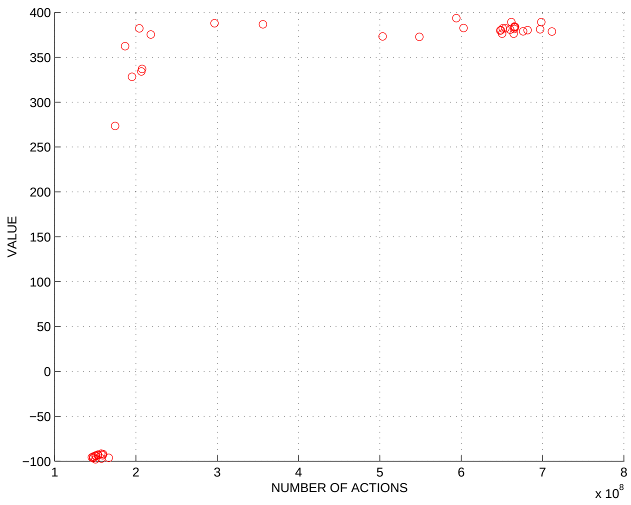

Figure 37: Value of the final policy generated over multiple runs

Only 97.5% of trials reach a terminal state. The maximum over 45 runs of the *median/mean* number of steps to terminal state of the final policy was 265/360

1.4.4 Gamma=1.0

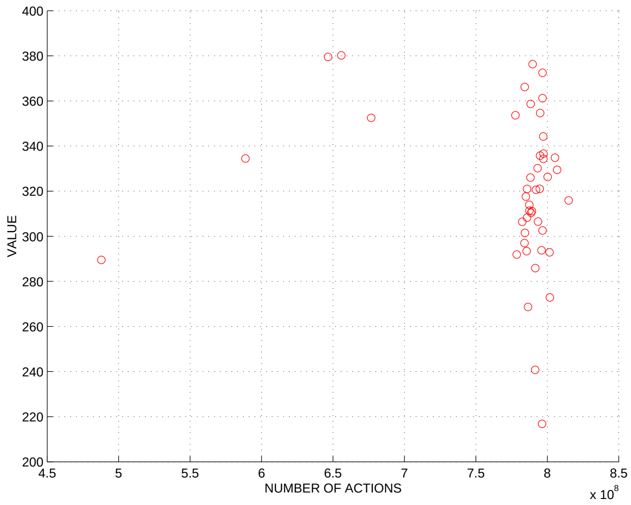

Figure 38: Value of the final policy generated over multiple runs

Only 97.8% of trials reach a terminal state. The maximum over 45 runs of the *median/mean* number of steps to terminal state of the final policy was 255/360

1.4.5 Algorithm MCESP-CE (No Averaging Zeroes)

Gamma=0.95, b=0.1, p=0.6

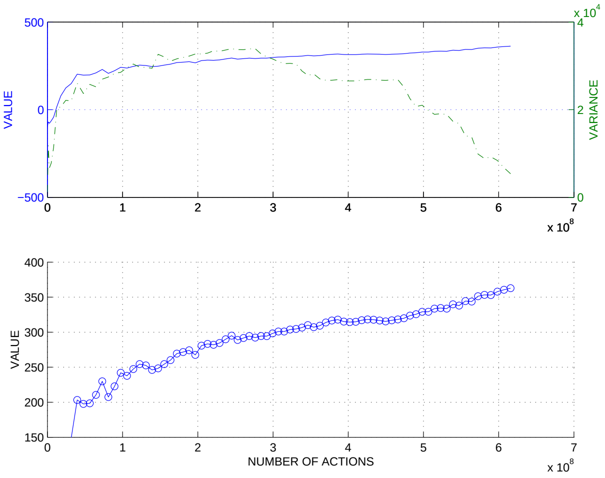

Figure 39: Average learning and variance curve for multiple runs

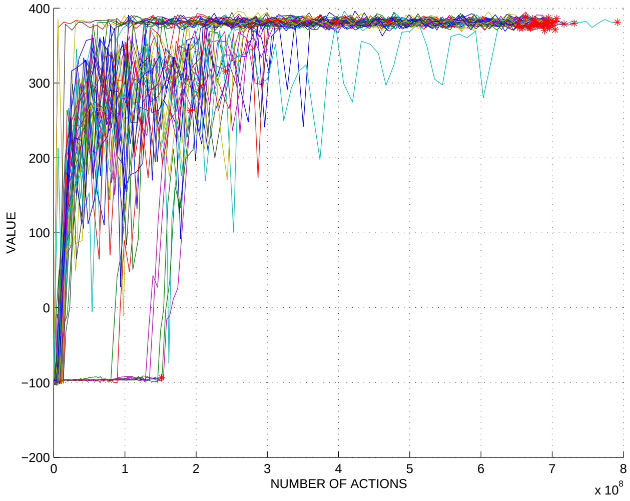

Figure 40: Learning curves for multiple runs

One run converges to a policy with value about -92 and three others to a policy with value 280. The rest converge to policies with values between 373 and 387. Only 97.7% of trials reach a terminal state. The mean over 45 runs of the *median/mean* number of steps to terminal state of the final policy was 257/350

1.4.6 Gamma=1.0, b=0.1, p=0.6

Only 98% of trials reach a terminal state. The mean over 45 runs of the *median/mean* number of steps to terminal state of the final policy was 240/325

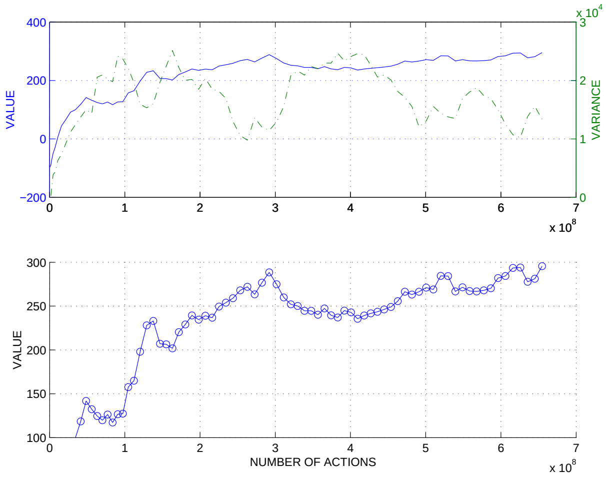

Figure 41: Average learning and variance curve for multiple runs

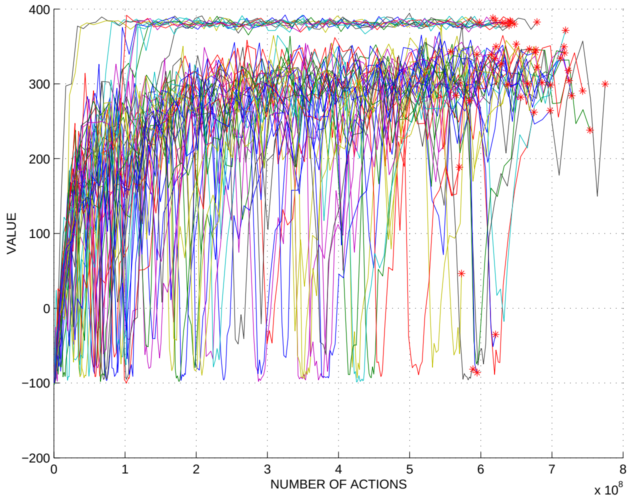

Figure 42: Learning curves for multiple runs

1.4.7 Algorithm MCESP-CE

Gamma=0.95, b=0.1, p=0.6

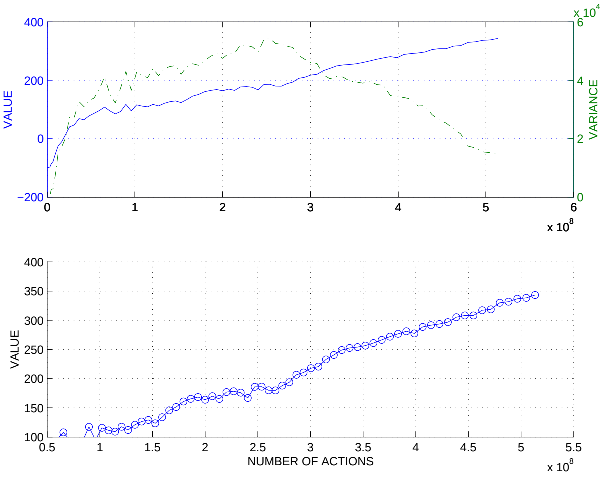

Figure 43: Average learning and variance curve for multiple runs

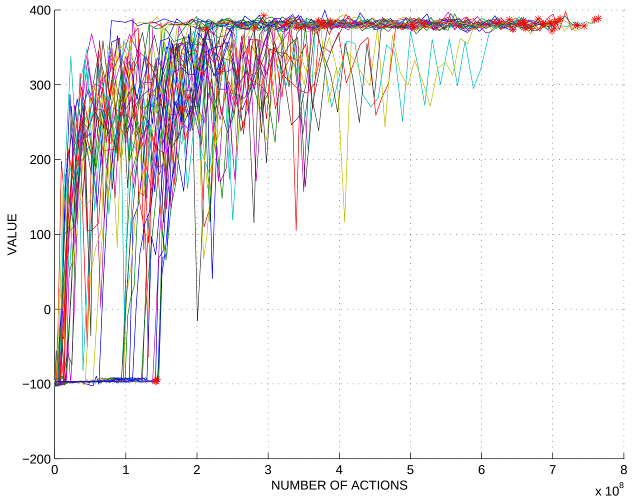

Figure 44: Learning curves for multiple runs

Three runs converge to a policy with value approximately −96, and three others to policies with value about 270. The rest converge to policies of value between 367 to 392. Only 97.7% of trials reach a terminal state. The mean over 45 runs of the *median/mean* number of steps to terminal state of the final policy was 247/338

1.4.8 Gamma=1.0, b=0.1, p=0.6

Only 99.5% of trials reach a terminal state. The mean over 45 runs of the *median/mean* number of steps to terminal state of the final policy was 23/32

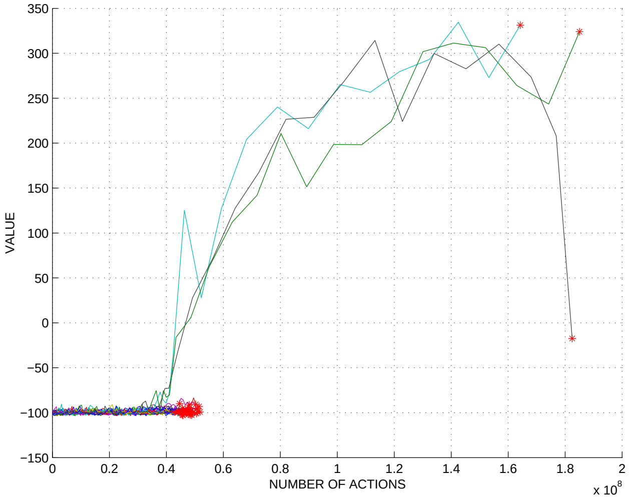

Figure 45: Learning curves for multiple runs

1.4.9 Algorithm Sarsa(0.9)

Number of runs=45, Initial Epsilon=0.2, Final Epsilon=0.0, Rate of epsilon decline=0.00000005, Gamma=1.0, Number of samples=10000, Maximum length of trajectory=1500, Maximum number of trajectories=80000

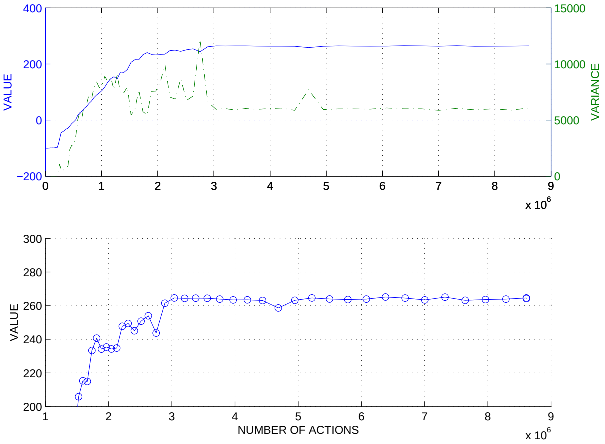

Figure 46: Average learning curve for multiple runs

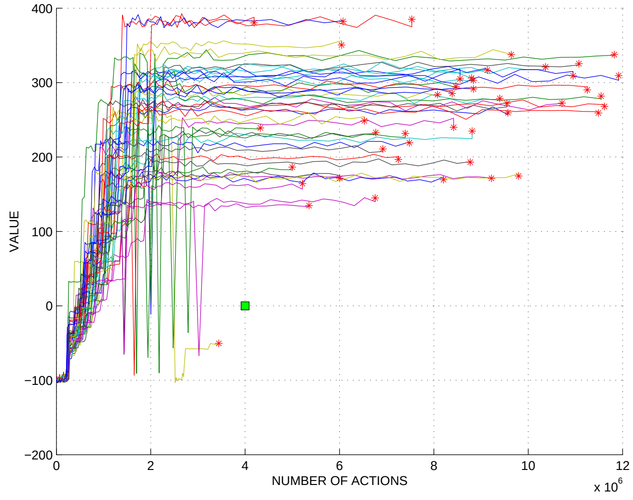

Figure 47: Learning curves for multiple runs (the green square marks the point at which exploration stops)

Only 98.8% of trials reach a terminal state. The mean over 45 runs of the *median/mean* number of steps to terminal state of the final policy was 220/308

1.4.10 Gamma=0.95

Only 98% of trials reach a terminal state. The mean over 45 runs of the *median/mean* number of steps to terminal state of the final policy was 250/344

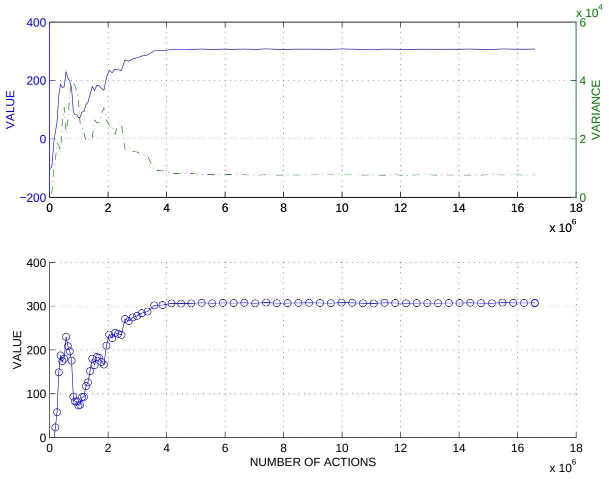

Figure 48: Average learning curve for multiple runs

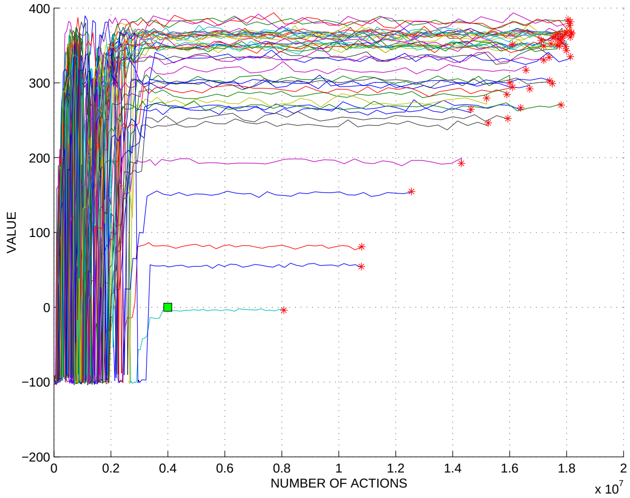

Figure 49: Learning curves for multiple runs (the green square marks the point at which exploration stops)

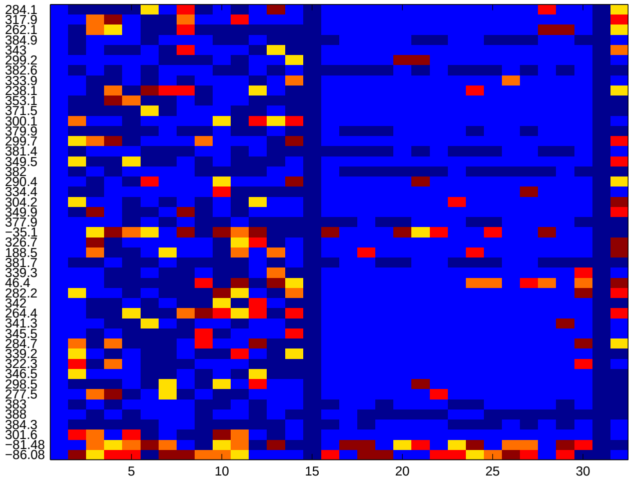

Figure 50: Policy Values for CE (Gamma=1.0 with no averaging zeroes)

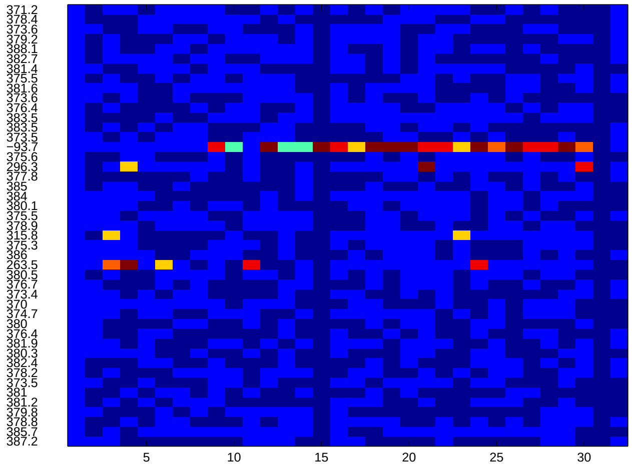

Figure 51: Policy Values for CE (Gamma=0.95 with no averaging zeroes)

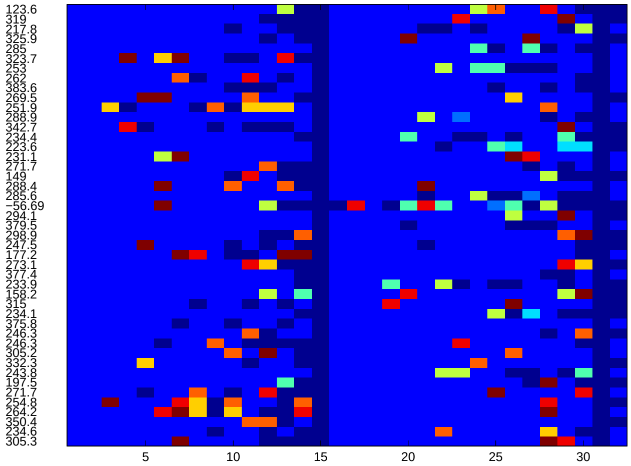

Figure 52: Policy Values for Sarsa (Gamma=1.0 with no averaging zeroes)

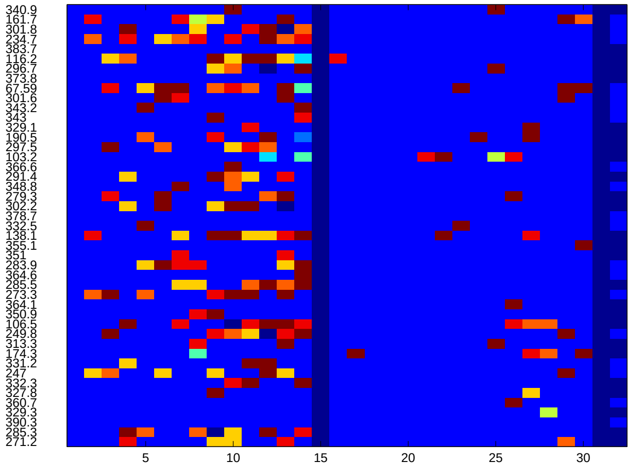

Figure 53: Policy Values for Sarsa (Gamma=0.95 with no averaging zeroes)