CS 285 Homework 4

Sagnik Bhattacharya SID: 3033583960

November 18, 2021

Part 1

Pointmass-Easy Environment

As we can see from Figures 1 and 2, the state density is somewhat more uniform for RND than it is for epsilon-greedy, although the difference is not very pronounced.

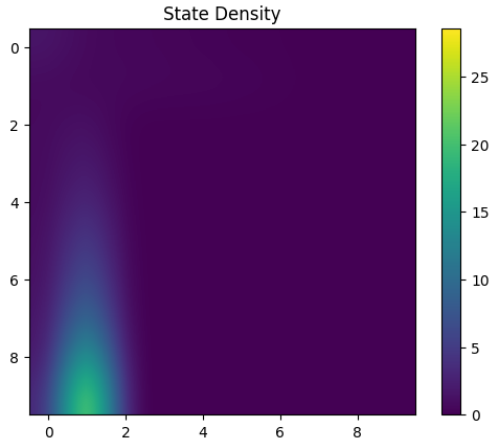

Figure 1: State density diagram for RND exploration.

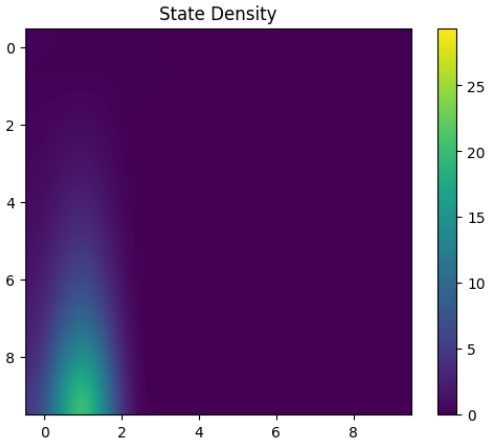

Figure 2: State density diagram for random epsilon-greedy exploration.

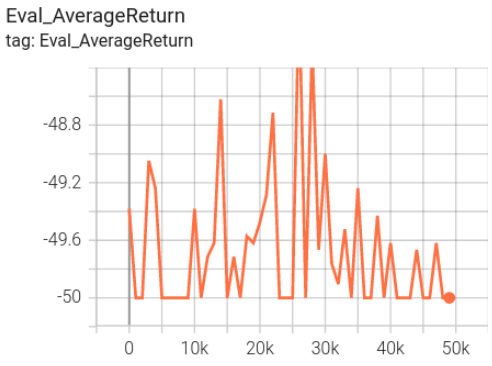

Figure 3: Learning curve for RND exploration.

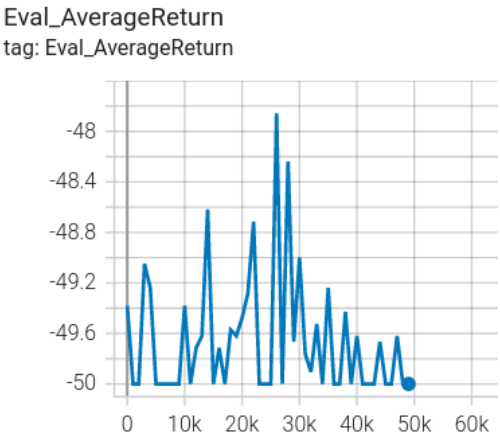

Figure 4: Learning curve for random epsilon-greedy exploration.

Pointmass-Medium Environment

As we can see from Figures 5 and 6, the state density is somewhat more uniform for RND than it is for epsilon-greedy, although the difference is not very pronounced.

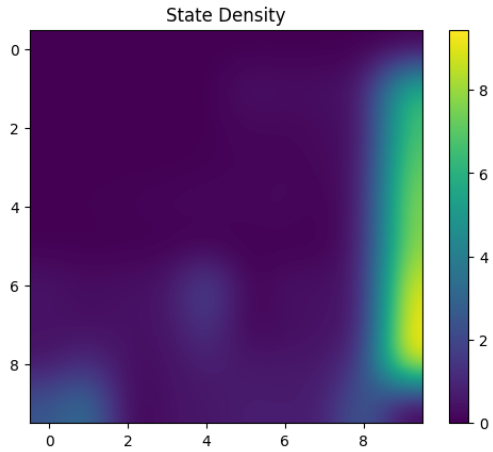

Figure 5: State density diagram for RND exploration.

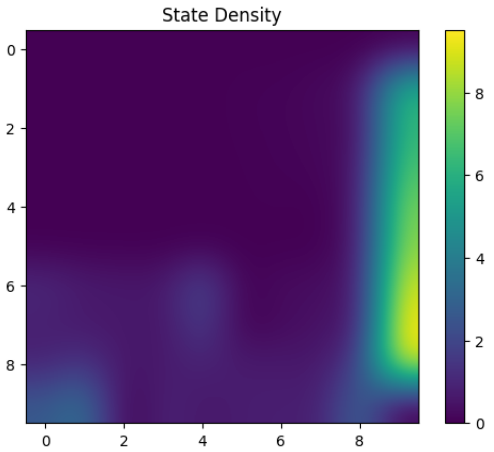

Figure 6: State density diagram for random epsilon-greedy exploration.

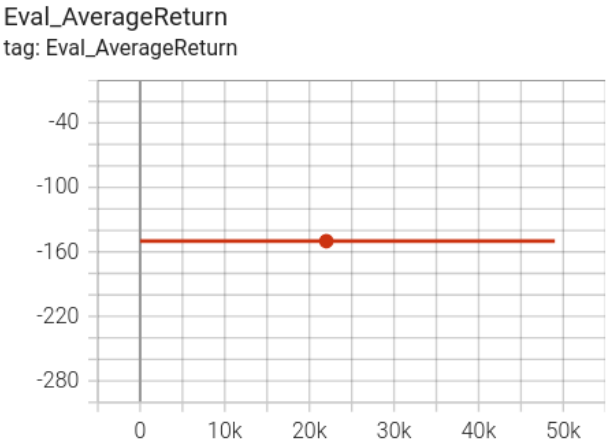

Figure 7: Learning curve for RND exploration.

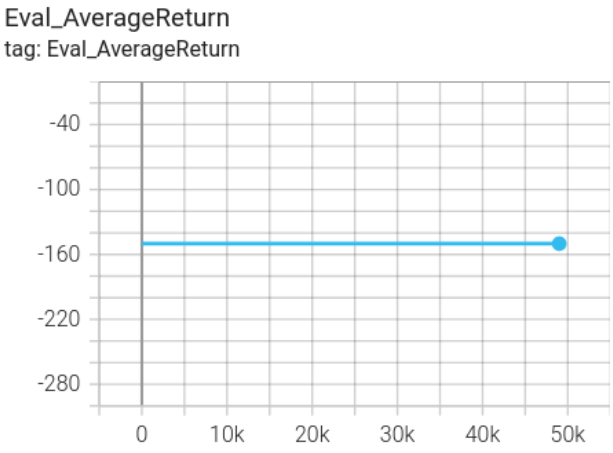

Figure 8: Learning curve for random epsilon-greedy exploration.

Part 2

Subpart 1

Figure 9 shows that CQL performs better than DQN, as it achieves optimal performance much faster. Figure 10 shows that CQL underestimates Q-values compared to DQN. This is most likely a result of the training procedure where the CQL is trained to reduce the Q-values of state-action tuples that are not seen in the training set.

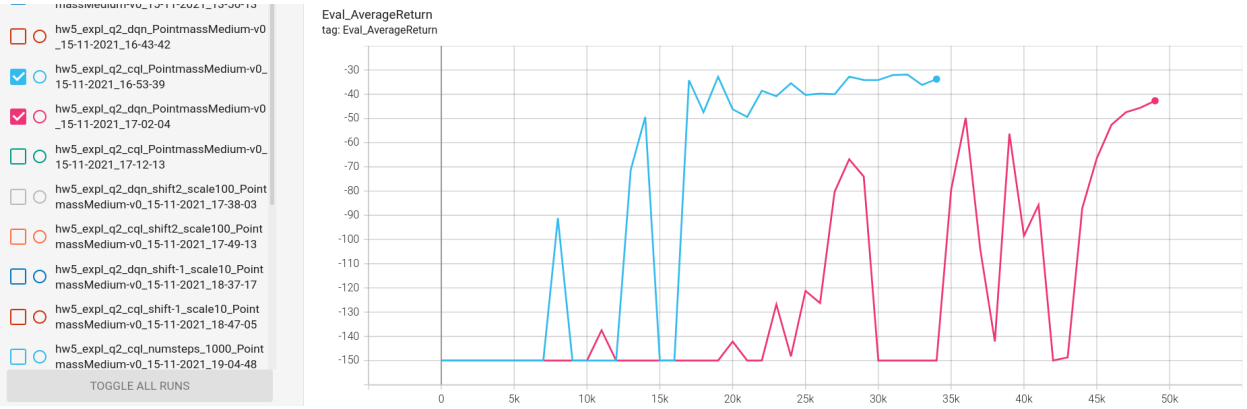

Figure 9: Average return of CQL (blue) vs DQN (pink) on PointmassMedium environment.

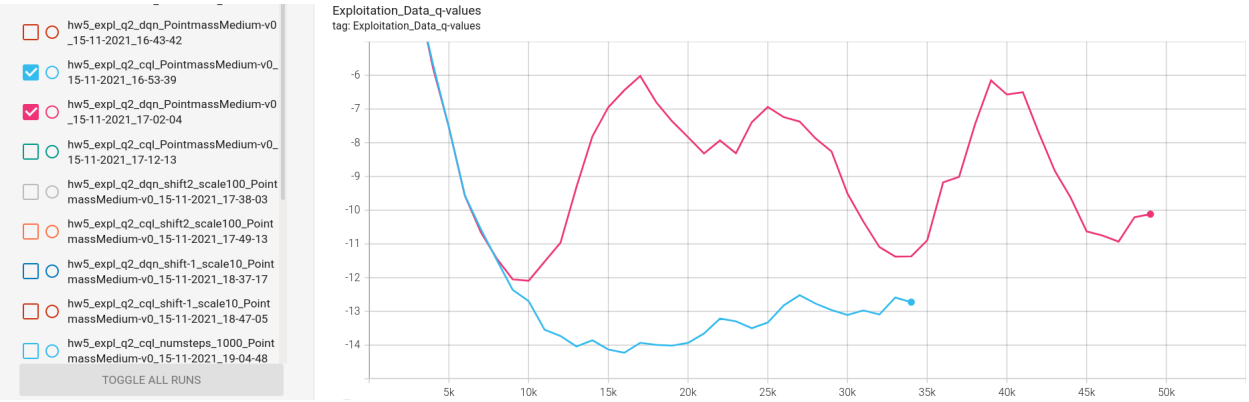

Figure 10: As we can see, CQL (blue) underestimates Q-values compared to DQN, since it is trained to reduce Q-values of samples not seen in the dataset.

Subpart 2

num_exploration_steps	CQL Average Return	DQN Average Return
1000	-150	-150
5000	-27.61	-28.43
15000	-22.63	-97.45

Table 1: A comparison of the performance of CQL and DQN with varying amounts of exploration data. The values in the average return columns are the average returns at convergence of these algorithms. As we can see from the table, having more exploration steps is better for CQL, although for DQN there seems to be a limit to how much exploration is good, as it seems that DQN is unable to filter out bad states.

Subpart 3

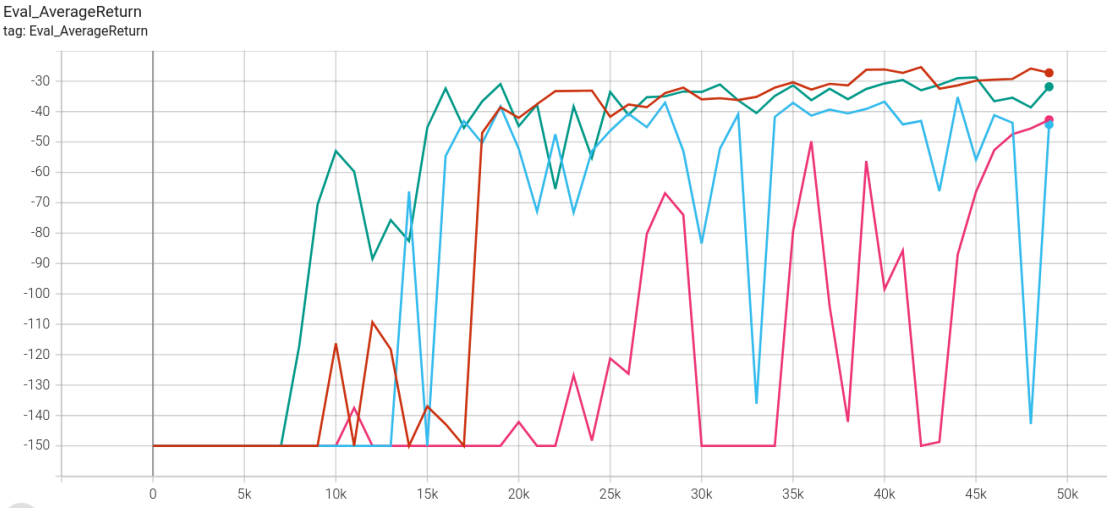

Figure 11: Evaluation average returns for parameter sweep of CQL alpha. Pink is $\alpha = 0.02$, blue is $\alpha = 0.5$, green is $\alpha = 0.1$, and red is DQN. As we can see, the best value for alpha is somewhere in between, as $\alpha = 0.02$ leads to the best performance.

Part 3

As seen in Figure 12 the results in the medium environment are somewhat better than those for purely offline exploration in part 2 for a given number of `num_exploration_steps`. Supervised exploration is better than unsupervised exploration since it is able to also take into account the reward structure in the environment and align its exploration with that, and it is also able to tune itself to the exploitation policy, thus showing it parts of the state-action space that it may otherwise avoid, and it avoids exploration in spaces that the exploiter likes to visit often.

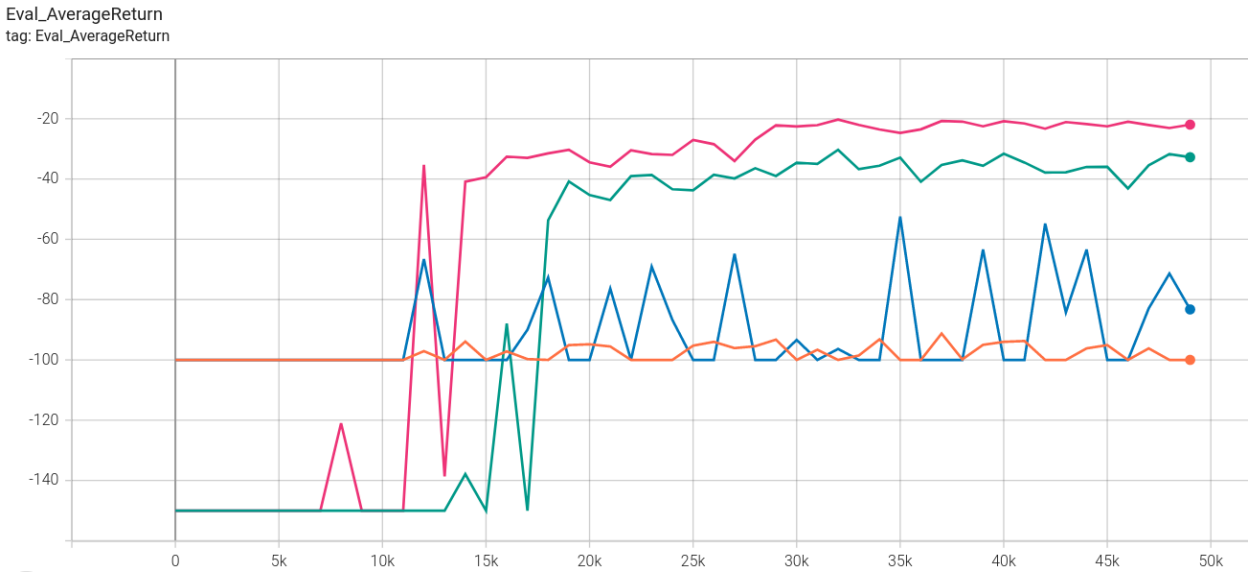

Figure 12: Evaluation average returns learning curve for CQL and DQN. Pink is CQL on PointmassMedium, green is DQN, blue is CQL on PointmassHard, and orange is DQN on PointmassHard.

Part 4

Figure 13 shows a comparison of supervised and unsupervised exploration with AWAC on the PointmassEasy environment, and 14 shows the same on the PointmassMedium environment. As we can see, supervised exploration leads to better performance across all values of λ that were provided in the homework specification. The specific value of λ does not matter that much in the easy environment, however, for optimal performance in the medium environment, supervised exploration along with an in-between value of $\lambda = 2$ seems to be best as seen in the green curve in Figure 14. As we can see, AWAC performs somewhat better than CQL, as evidenced by the graphs presented here.

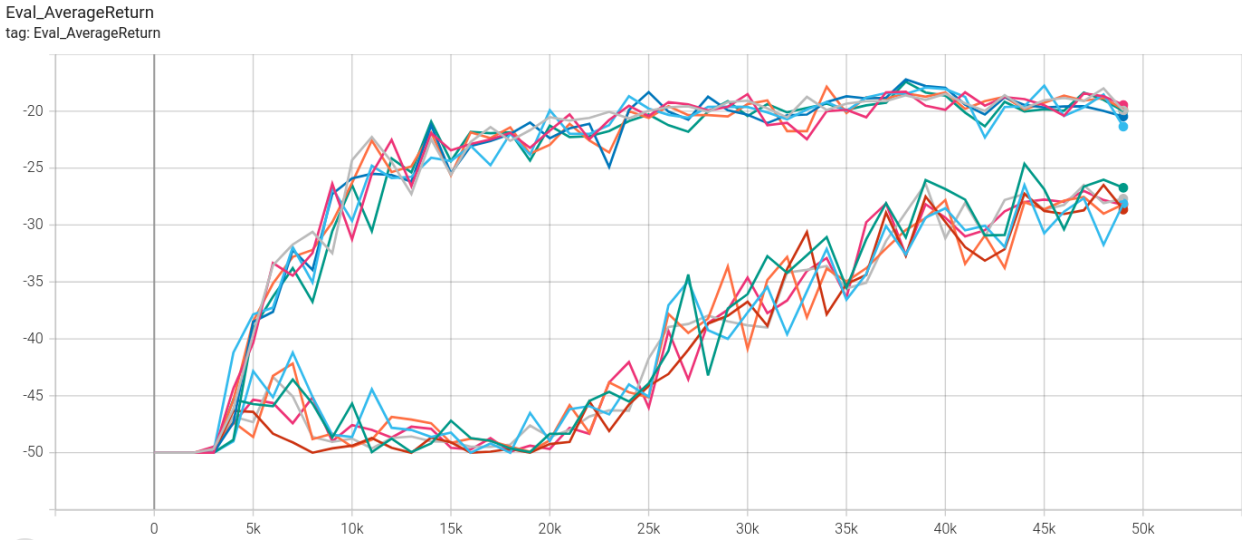

Figure 13: Supervised exploration (top curves) and unsupervised exploration (bottom curves) for all values of λ specified in the homework specification with AWAC on the PointmassEasy environment.

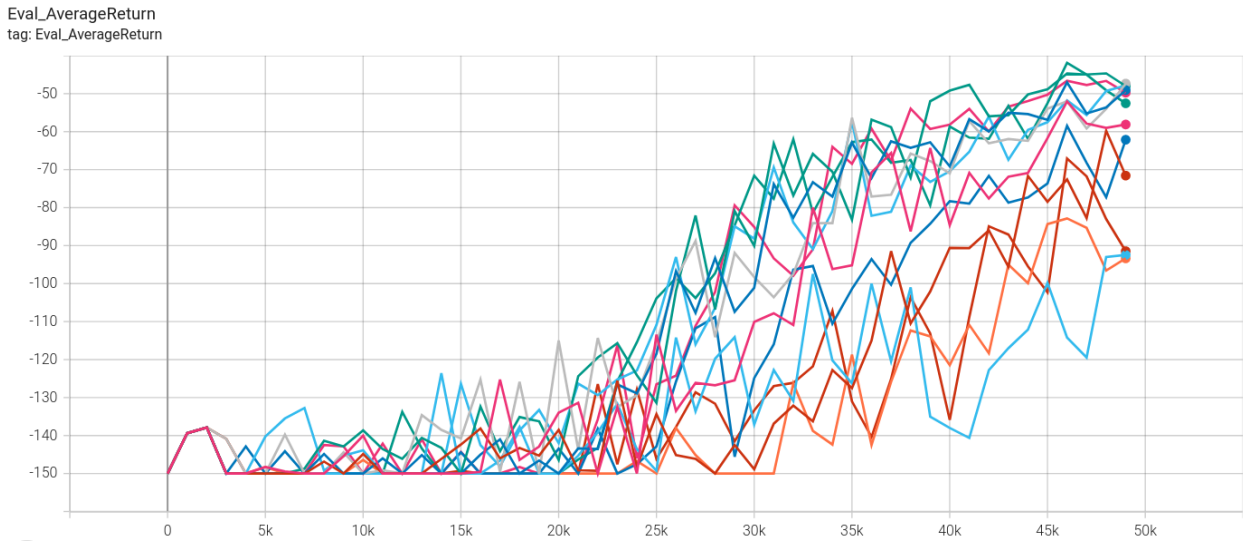

Figure 14: Supervised exploration (top curves) and unsupervised exploration (bottom curves) for all values of λ specified in the homework specification with AWAC on the PointmassMedium environment.