Assignment 5: Exploration and Offline Reinforcement Learning

Yulun Rayn Wu, 3034358565

November 26, 2020

1 Part 1

1. sub-part 1

   Easy:

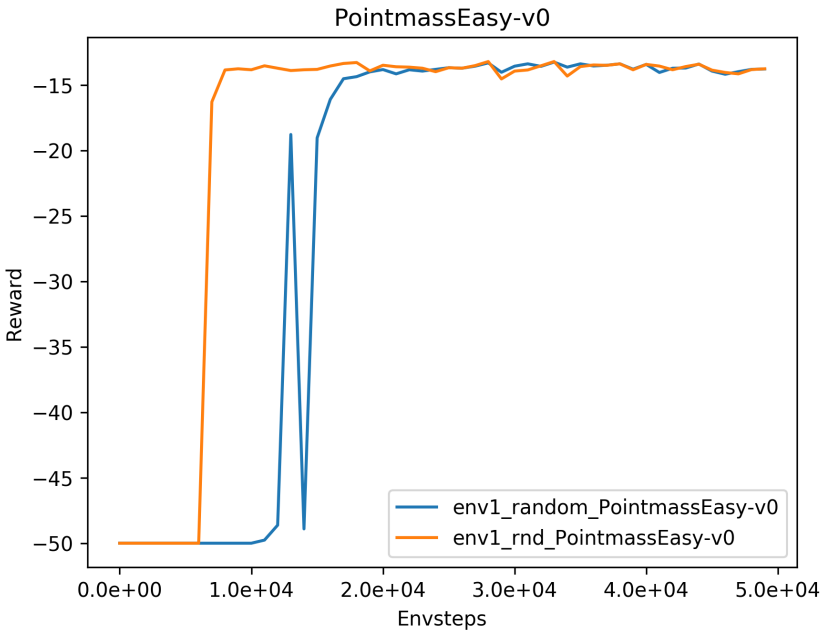

Figure 1: Eval Average Return - Easy

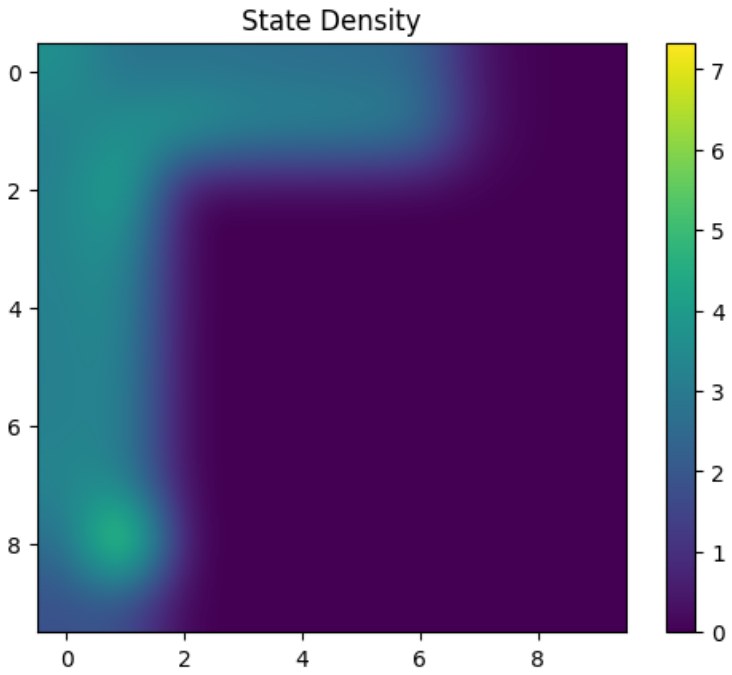

Figure 2: Heatmap-rnd

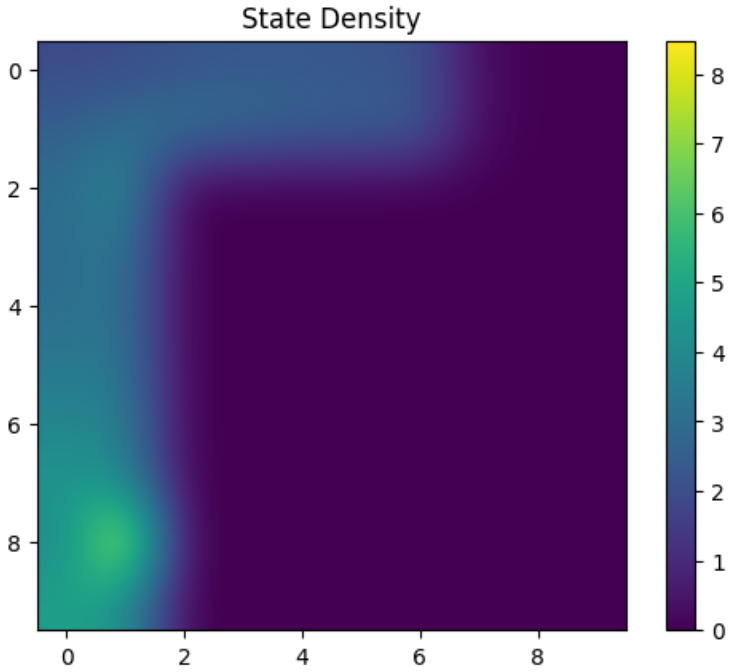

Figure 3: Heatmap-random

Medium:

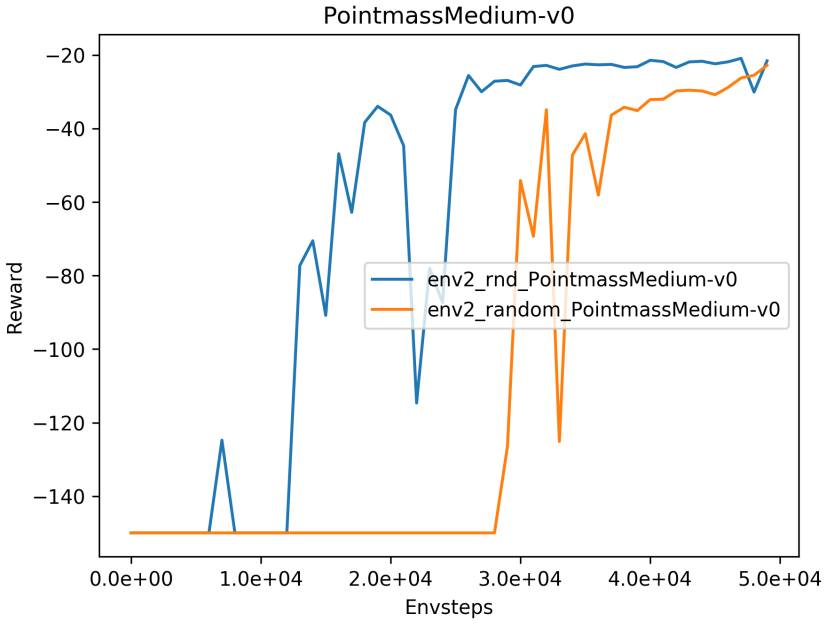

Figure 4: Eval Average Return - Medium

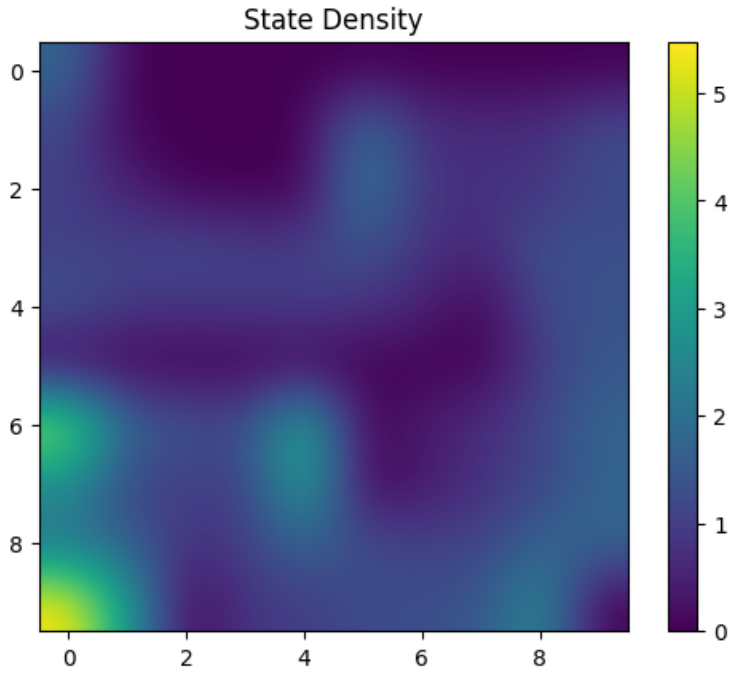

Figure 5: Heatmap-rnd

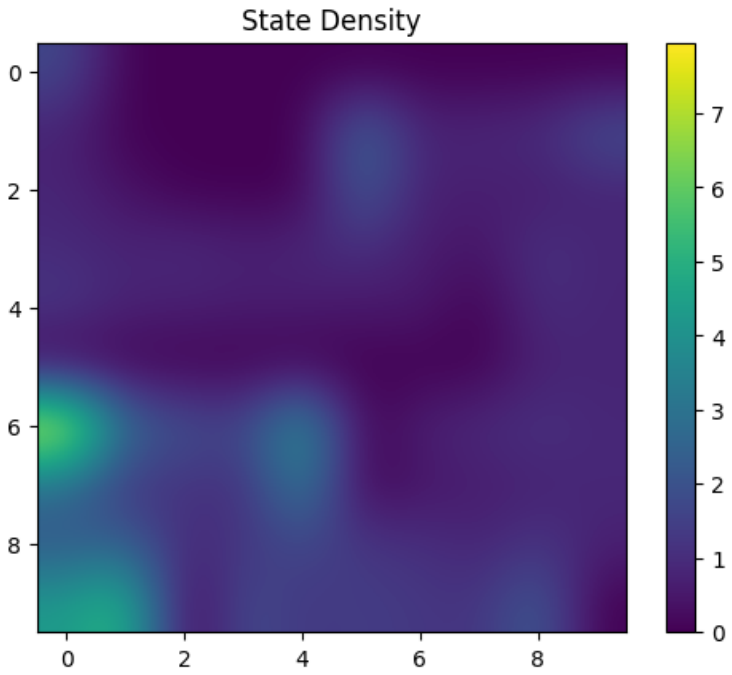

Figure 6: Heatmap-random

2. sub-part 2

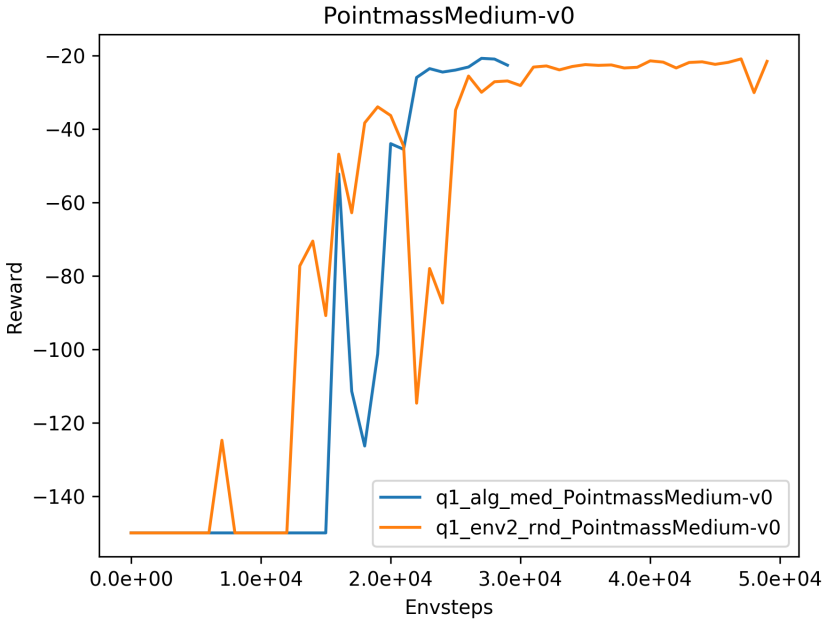

Figure 7: Eval Average Return

The blue line uses RBF for exploration. It seems to be slower to explore to a rewardable result but converges faster in terms of envsteps after the discovery. I think the initial

discovery being slower might just be random or settings related but a count-based method provides more consistent exploration and it might contribute to faster convergence afterwards. Again, this is in terms of envsteps, an NN based approximation would cost less time to converge in terms of time.

2 Part 2

1. sub-part 1

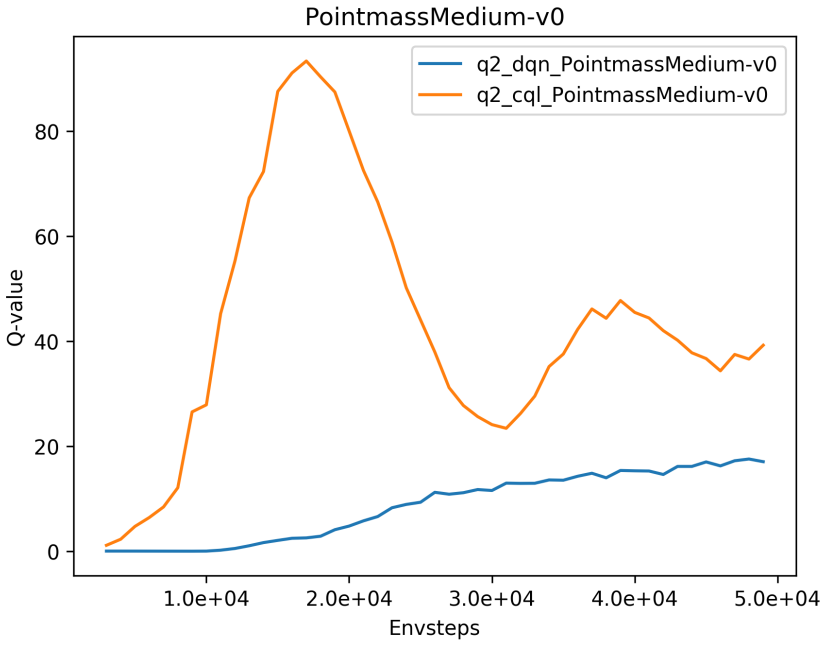

Figure 8: Data Q-values

CQL did give rise to Q-values in the experiment.

2. sub-part 2

Table 1: Last Eval Average Return

	CQL	**DQN**
Num_Step 5000	−23.19	−31.35
Num_Step 20000	−52.6	−41.92

Lower number of exploration steps seem to be adequate for medium environment.

3. sub-part 3

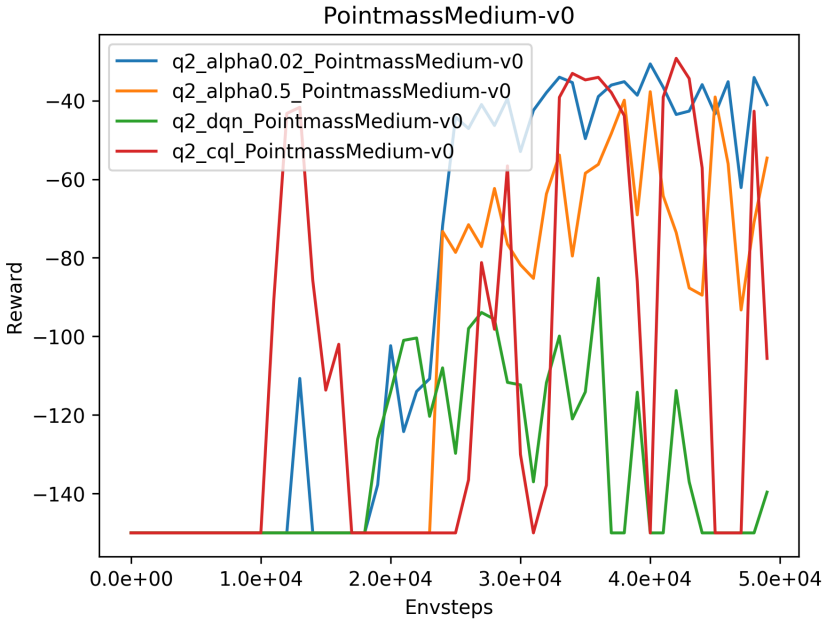

Figure 9: Data Q-values

CQL with $\alpha = 0.02$ performs the best while DQN performs the worst in these experiments.

3 Part 3

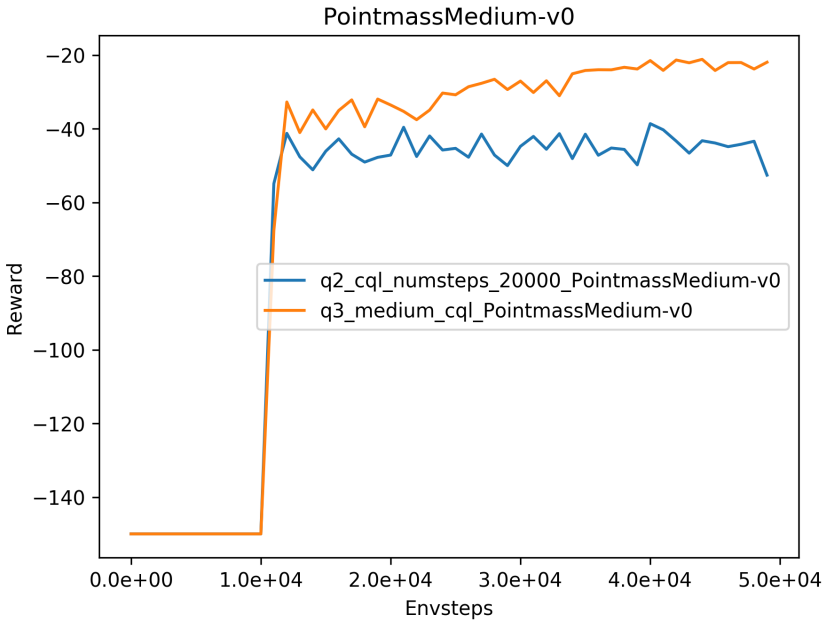

Figure 10: Eval Average Return - CQL

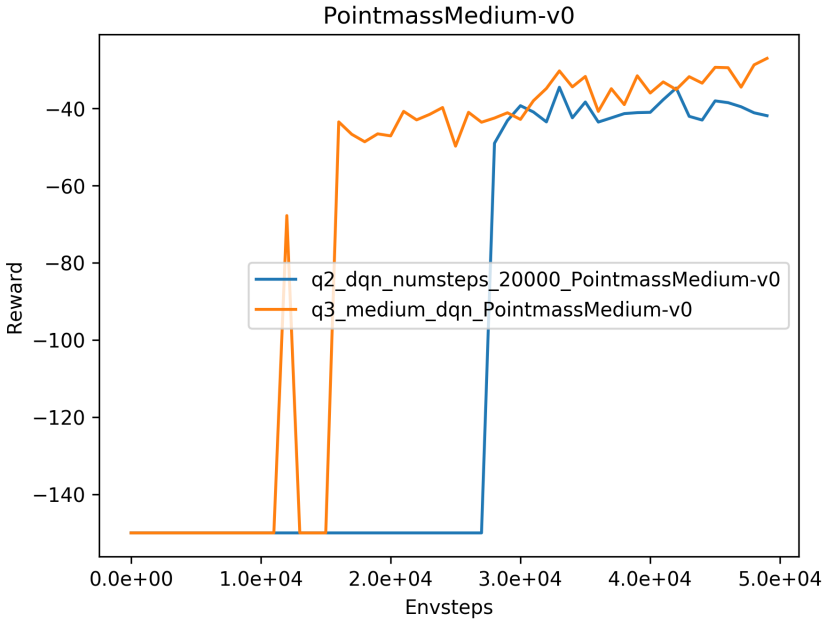

Figure 11: Eval Average Return - DQN

1. "Supervised" exploration seems to be able to converge to higher returns.

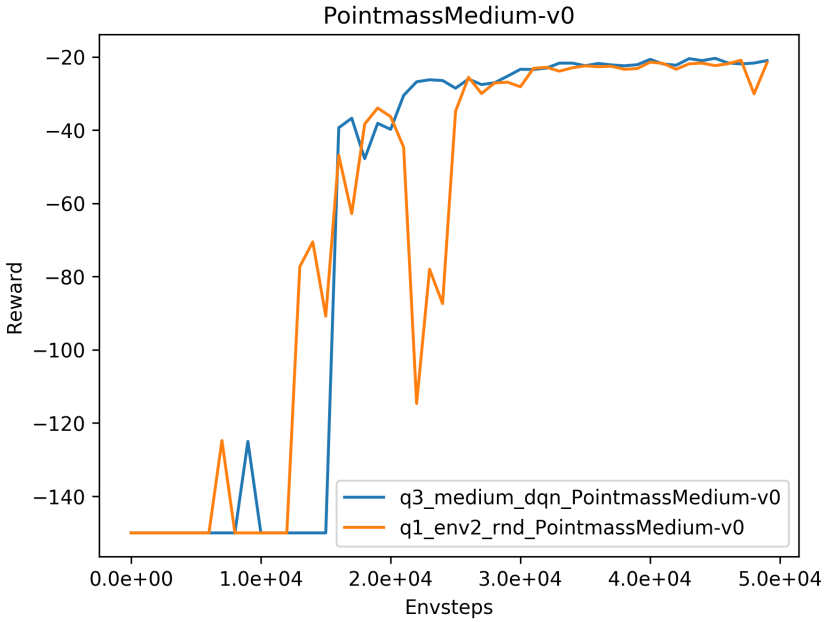

Figure 12: Eval Average Return

2. The results start to converge later than lower exploration steps. Exploration with a combination of both rewards converges more robustly and thus showed more efficiency in the experiment.