CS 285 HW 5

**Part 1: Unsupervised" RND and exploration performance**

**State density plots and learning curves for easy and hard env:**

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Caption: State density for the easy env. On the left: random. On the right: RND. RND state density is more uniform than that without RND.

Chart, line chart

Description automatically generated

Caption: The learning curve for the easy env. Orange: random. Blue: RND.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Caption: State density for the hard env. On the left: random. On the right: RND. RND state density is more uniform than that without RND.

Chart, line chart, histogram

Description automatically generated

Caption: The learning curve for the hard env. Blue: random. Red: RND.

**Part 2**: **Offine learning on exploration data**

**First sub-part:**

Chart, line chart

Description automatically generated

Caption: Green: CQL with shifted/scaled rewards. Red: DQN. Gray: CQL.

The shifted/scaled rewards improved the performance of CQL, i.e. it is better than just CQL without shifting and scaling. I think the reason behind this difference is that the shift and scale makes the large reward much more significant. So the advantage of a better policy is more obvious.

CQL does not give rise to Q-values that underestimate the Q-values learned via a standard DQN.

**Second sub-part:**

I expect learning with more number of exploration steps works better. The CQL result shows the two parameters give rise to similar performance. The DQN results shows smaller number of exploration steps works better.

Chart, line chart

Description automatically generatedCQL:

Caption: CQL. Orange: num\_exploration\_steps=5000. Blue: num\_exploration\_steps=15000

Chart, line chart

Description automatically generatedDQN:

Caption: DQN. Red: num\_exploration\_steps=5000. Blue: num\_exploration\_steps=15000

Chart, line chart

Description automatically generated**Third sub-part:**

Caption: CQL. Red: cql\_alpha=0.02. Green: cql\_alpha=0.5.

**Part 3: “Supervised” exploration with mixed reward bonuses**

Medium:

Chart, line chart

Description automatically generated

Caption: Medium env. Red: CQL. Gray: DQN.

They have similar performance to the offline learning in the subpart of Part 2.

Hard env:Chart, line chart

Description automatically generated

Caption: Hard env. Red: CQL. Blue: DQN.

The CQL has a much better performance compared to that in Part 1. Exploration with a combination of both rewards is more effective than purely RND based exploration.