Supporting document

This document consists of answers with figures that assist the rebuttal of Paper711.

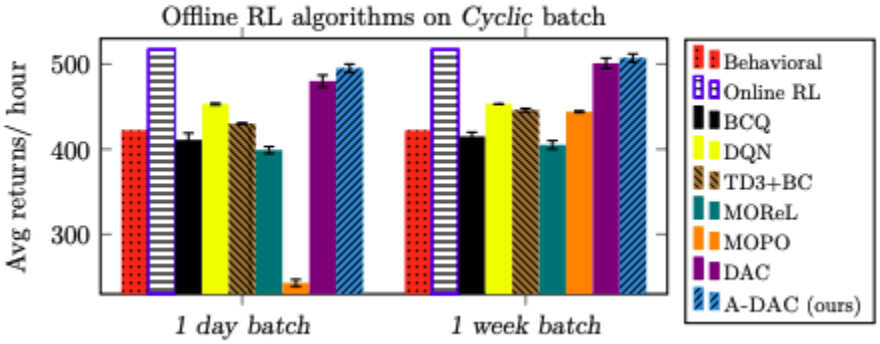
Reviewer 6AK5

1. **The author constructed the transit model by using the KNN method, can the author explain why they use this method, and did they try other supervised learning methods to learn the transition model?**

We evaluate two MBRL algorithms that first build an approximate MDP dynamics from the offline data set and then do sample rollouts from the derived MDP to optimize a policy network.

The first, MOReL [Kidambi et.al., 2020], derives an ensemble of learned dynamics models which allows tracking of uncertainty in estimation of next state. This uncertainty quantification is used in creating a pessimistic(P-) MDP model where transitions to uncertain regions are restricted by a threshold parameter.

The second, MOPO [Yu et.al., 2020], starts by building an ensemble of learned dynamics models similar to MOReL. Further, it learns a reward model from the data set as well. It then derives a new uncertainty-penalized MDP wherein the uncertainty quantification given by the ensemble of models is used to penalize reward estimates.



Supporting document

Both these methods fail to match the performance provided by our KNN-based transition model as can be seen from the picture above. On discrete action settings like ours, our method proves to be much superior.