1 Problem Setting

1. The paper is interested in learning purely from off-policy collected data without further possibility for data collection.

2 Extrapolation Errors

1. They define extrapolation error as an error in off-policy value learning.
2. This error happens because of the mismatch between the dataset and the true state-action visitation of the current policy.
3. The value estimate $Q(s, a)$ is affected by extrapolation error during a value update where the target policy selects an unfamiliar action a' at the next state s'.
4. Unfamiliar here can mean that either (s', a') is unlikely or not contained in the replay buffer.
5. They attribute extrapolation error to 3 reason: model bias, absent data, and training mismatch.
6. Model bias means that the target Q value is over the next state distribution in the replay buffer, rather than the true MDP.
7. Absent data is claimed to be a more serious problem in continuous action space setting, where the estimate of $Q_\theta\left(s', \pi\left(s'\right)\right)$ may be arbitrarily bad without data near the corresponding action-state tuple.
8. Training mismatch refers to the mismatch between the distribution of data in the training batch and the distribution under the current policy. If there is mismatch, then the value function may be a poor estimate of actions selected by the current policy, leading to poor target estimate.

2.1 Demonstration of extrapolation error in DRL

1. The dataset generated by and is used to train DQN tends to be heavily correlated to the current policy.
2. This is due to the choice of near on-policy exploration policy.
3. They experimentally demonstrate that the performance of DDPQ deteriorates rapidly when the data is very uncorrelated from the current policy and the value estimate produced by the deep Q-network diverges.
4. They then argue that current off-policy DRL algorithms are ineffective when learning truly off-policy.

3 Batch-Constrained RL

1. They claim that the value of an off-policy agent can be reliably estimated in region with available data.
2. They thus propose an idea to combat extrapolation error: the policy should produce a state-action visitation frequency similar to the batch.
3. They denote policies which satisfy this notion as batch-constrained.
4. The policies they train thus need to satisfy 3 objectives: minimize the distance of selected action to the data in the batch, leads to familiar states in the future, and maximize the value function.

3.1 Addressing extrapolation error in finite MDP

They managed to prove interesting theoretical results for finite MDP.

3.2 Batch-Constrained RL

1. Their approach is called batch-constrained Q-learning (BCQ).
2. There are 3 main components in their algorithms: a generative model which, for a given state, generates action close to action in the batch, a Q function to evaluate the generated actions and a perturbation model which add noises to the generated action.

4 Questions

1. Does the code work on newer version of OpenAI gym?
2. What insight can lead to improvement in model-free RL? The buffer in model-free RL can be large, so there is a large degree of off-policy ness.
3. How is the performance compared to "Model-Predictive Policy Learning with Uncertainty Regularization for Driving in Dense Traffic"?
4. Is model bias only an issue in stochastic MDP?
5. They have mentioned 3 separate sources of extrapolation error. Which source does their approach mitigate?
6. How sensitive is the performance to the weight λ in the value estimate ? Is there an ablation studies which set λ to 1?
7. Is there an ablation study for the perturbation model?
8. Does the approach work on more complex environments like Ant or Humanoid?