Project Milestone Report

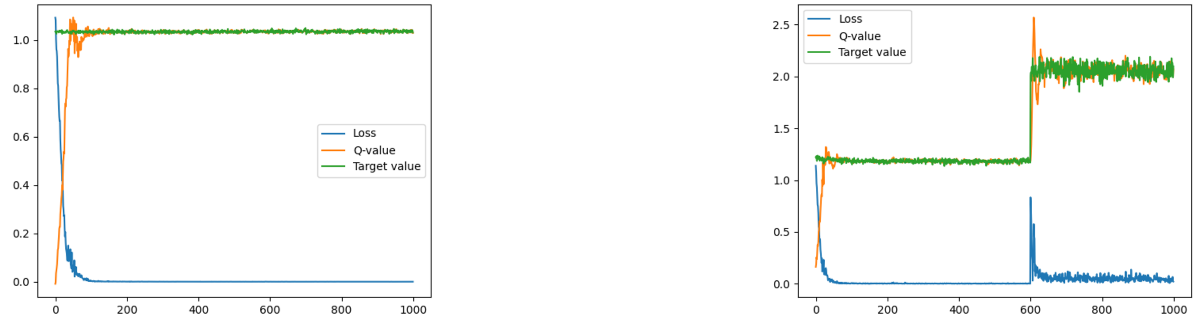
Alina Trinh, Samarth Goel

1 What experiments have you conducted so far?

We have created a framework to train a Q-value estimator using expert data that we have taken from course assignments, then analyzing the trajectories taken by the expert alongside our training process to determine states with high uncertainty in expert decision-making, quantified by some measure of variance on either agent actions or actions taken by the expert. We used the gymnasium environment instead of the OpenAI gym package that is used in the course assignments, and recreated our own custom networks and agents, along with custom state and trajectory class representations to support the operations we needed on these concepts.

For the variance of a set of states, we decided to take the average variance of all the states in the cluster. To determine the variance of a state, we ran a trained agent on a state some amount of times and took the variance of its outputted actions.

Below we plot the training loss, q-values, and target values during training the agent on the expert data, and then again on pruned trajectories which we simulate using the trained agent. We see that with the pruning of “bad" data consisting of states that have too high of action variance, the agent is able to achieve twice as much return. We hypothesize that the jump in return is due to the epsilon schedule which leads the agent to exploit the good, pruned data more than explore.



(a) Expert Trajectories

(b) Pruned Trajectories

Figure 1: Training Metrics with Non-Pruned and Pruned Trajectories

2 Are there any changes to the research hypothesis or problem statement from the proposal?

Our problem statement is generally similar to what we proposed earlier. We are still focused on determining sub-optimal trajectories and have focused on the research hypothesis of what effect trajectory pruning based on various measures of state similarity and action variance will have on model training metrics. This can look like time till reaching certain levels of rewards, maximum reward attained over all timesteps, and changes in the return, loss, and period curves. We expect a lot of fine-tuning and refinement for this aspect.

In addition, we are also planning to look at doing this pruning to a smaller extent but also online, so pruning will happen while the Q-function estimator is being trained instead of after, as we are currently doing. This might lead to worse results but is more useful in practice.

Lastly, we are still planning to determine differences between high-variance states where there is indifference vs. uncertainty. We haven’t implemented that for the checkpoint, however, but once this is in place, we expect to see better results in the long-run setting due to being able to keep more data and reduce unnecessary pruning.