CSCI 2951-F Final Project

Enrique and Ellis and Kavosh and Dave

1 Paper Overview

Note: Introduce the problem that they're solving, i.e. identify the main research contribution made (a natural language version of theorem 1 ish? that there is this relationship between model accuracy and planning depth, and that gamma lets you control that).

2 Domains

2.1 RockSample

The RockSample domain consists of an agent acting in a partially observable GridWorld bounded by walls on the west, south and north sides. There are k rocks that occupy unique cells of the grid world some of which are good and some of which are bad. There are $5 + k$ actions available to the agent, $\{North, East, South, West, Sample, Check_1 \dots Check_k\}$.

If the agent calls *Sample* while it is on top of a good rock it receives +10 reward and if it does so while on top of a bad rock it receives −10 reward. The agent also receives +10 reward if it runs of the east edge of the GridWorld, thereby terminating the episode.

The state space is fully observable to the agent except the goodness of rocks – that is, it always knows its own position and that of all rocks, but not the goodness of the rocks. The $Check_i$ action provides the agent with noisy knowledge of the ith rock. If the agent is directly on top of $rock_i$ the $Check$ action returns the true goodness of the rock. As the agent gets further and further from the rock that it is $Check$ing, the fidelity of its sensor falls off exponentially, bottoming out at a 50% probability of returning the true goodness of the rock it is $Check$ing. The agent's initial belief state is that each rock has a 50% chance of being good.

2.2 Randomized MDPs

3 Hypotheses

3.1 Figure 6

3.2 Figure 7

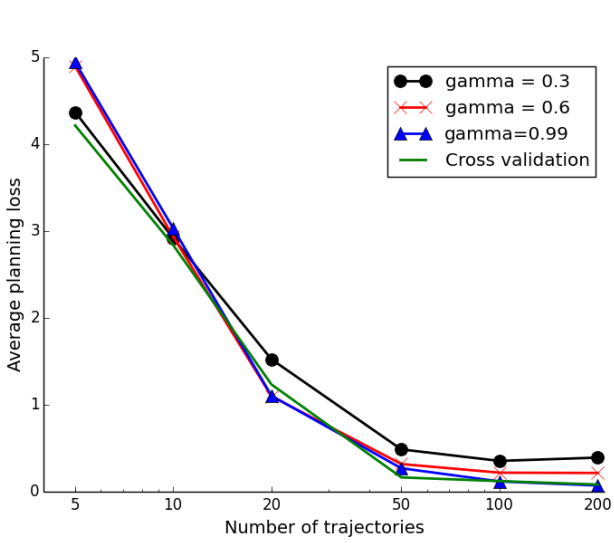
4 Experiments

Idea: Discuss how we chose to implement our version of the experiments?

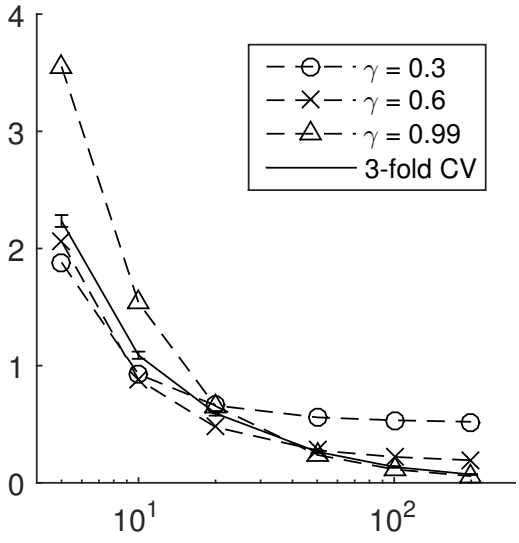
5 Results

Summary of their results, summary of our results.

Cross validation with RandomMDP



our results



their results

6 Reproducibility Discussion

Discussion of how reproducing results was difficult, what assumptions we made, what parameters and other bits of info were left from the paper.

6.1 Ambiguity about RockSample

The initial state of the agent was underspecified. It was not stated whether the same initial state was used in each episode or if the state was initially randomized. In the case of the former, it was not clear what the initial state would be and in the case of the latter the distribution over states and number of rocks was indeterminate. We consulted the authors on this and were able to to determine that the same initial state with 6 rocks at fixed locations was used (this results in 16,000 underlying states which is huge, see later Section **??** on why this is problematic). Similarly, the initial goodness of rocks was not specified – we assumed all rocks are initially good.

There was additional ambiguity regarding the the fidelity falloff of the *Check* actions. We assumed the distance between the agent and a rock (which determines how unreliable the sensor is) is measured using Euclidean distance. We could have also reasonably used Manhattan distance since the agent cannot move diagonally. It was also unclear what the decay rate for exponential falloff was; we used $\frac{1}{2}$.

6.2 Ambiguity about UCT

Running UCT involves assigning several parameters that can dramatically alter results. For instance, the assignment of the exploration bias (i.e. the UCB parameter) will affect how UCT plans. The paper reports that their results were optimized over the set $10 * \exp -2, -1, 0, 1, 2$, for each data point on the graph. We fixed this parameter to be the maximum, $10 * exp(2) \sim 74$ for all experiments. Additionally, the value of the discount factor γ is critical in determining behavior,

especially in a domain like rock sample where there are large bursts of reward in the distance. We set γ to be 0.9 for our experiments, but after communicating with the authors, discovered there setting of γ was 0.99 (although they were unsure). Lastly, there are several different ways one could imagine running UCT on a POMDP. We chose to convert the POMDP into a BeliefMDP and run UCT on the BeliefMDP, both during simulation and evaluation. The authors chose to sample ground states from their belief state during simulation but evaluate the performance of UCT by explicitly tracking the belief state using a generative model of the ground MDP. Several other components of the experimental setup were unclear in the paper, but the authors were kind enough to clarify these points over email.

6.3 General Computational Limits in RockSample

Their computer big and fast our computer small and slow.

7 Conclusion