CS 277 (W24): Control and Reinforcement Learning

Quiz 6: Planning and MBRL

Due date: Wednesday, February 21, 2024 (Pacific Time)

Roy Fox
https://royf.org/crs/CS277/W24

Instructions: please solve the quiz in the marked spaces and submit this PDF to Gradescope.

Question 1 When sampling experience (s, a, r, s') for RL, an arbitrary-reset simulator $\hat{p}(s'|s, a)$, which can be reset to any state s, is more useful than a simulator that cannot, in the following ways (check all that hold):

- ☐ s can be sampled from an arbitrary distribution.
- ☐ a can be sampled on-policy $(a|s) \sim \pi$.
- ☐ $(r, s'|s, a)$ can be sampled multiple times.
- ☐ s can be set to s' after every sample (except when s' is terminal), to get entire trajectories.
- ☐ None of the above

Question 2 In model-based exploration algorithms, let $\hat{M}$ be a good approximation of the real MDP in a subset S of states (*known* states). $\hat{M}'$ is similar to $\hat{M}$, except that $\hat{M}$ gives reward 0 in unknown states, while $\hat{M}'$ gives the maximum reward $r_{\max}$. Check all that hold for the optimal policy π in $\hat{M}$ and the optimal policy π' in $\hat{M}'$:

- ☐ If π has low probability to reach an unknown state, than it is near-optimal in M.
- ☐ If π' has low probability to reach an unknown state, than it is near-optimal in M.
- ☐ π tends to have a higher probability than does π' to reach an unknown state.
- ☐ E^3 uses π' rather than π for exploration, because π' is optimistic under uncertainty and thus explores more.
- ☐ None of the above

Question 3 Model Predictive Control (MPC) uses an approximate model for planning, but then only executes each plan for a single step, and re-plans after every action. This scheme partly mitigates the accumulation of model error. This is true regardless of observability, and is equally beneficially in unobservable environments. **Yes / No**.

Briefly justify: