# CS 277 (W22): Control and Reinforcement Learning

# Quiz 5: Model-Based and Partial Observability

## Due date: Friday, February 25, 2022 (Pacific Time)

Roy Fox
https://royf.org/crs/W22/CS277

**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.

**Question 2** In Iterative LQR (iLQR) (check all that hold):

- ☐ If the dynamics is globally linear and the cost globally quadratic, the algorithm converges in one step.
- ☐ The cost Hessians are guaranteed to be positive (semi)-definite $\nabla_x^2 \hat{c}_t \succeq 0$, $\nabla_u^2 \hat{c}_t \succ 0$.
- ☐ The algorithm always converges, but possibly to a local optimum.
- ☐ An agent that can interact with a deterministic environment can run iLQR without a known (or learned) model: after updating the policy to the LQR optimum, the new trajectory can be found by rolling it out in the environment.

**Question 3** 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 $\pi$ in $\hat{M}$ and the optimal policy $\pi'$ in $\hat{M}'$:

- ☐ If $\pi$ has low probability to reach an unknown state, than it is near-optimal in $M$.
- ☐ If $\pi'$ has low probability to reach an unknown state, than it is near-optimal in $M$.
- ☐ $\pi$ tends to have a higher probability than does $\pi'$ to reach an unknown state.
- ☐ $E^3$ uses $\pi'$ rather than $\pi$ for exploration, because $\pi'$ is optimistic under uncertainty and thus explores more.

**Question 4** 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 can be beneficially used in unobservable environments. **True / False**.

| **Briefly justify:** |
|---|
| |

**Question 5** Using RNNs in deep RL (check all that hold):

- ☐ REINFORCE with an RNN policy $\pi_\theta(a_t|m_t)$, with $m_t = f_\theta(m_{t-1}, o_t)$, can compute an unbiased policy gradient $\sum_t R(\xi)\, \nabla_\theta \log \pi_\theta(a_t|m_t)$.
- ☐ A2C (on an entire sampled trajectory $\xi$) with an RNN actor as above and a critic $V_\phi(m_t)$ can compute an unbiased policy gradient $\sum_t (R_{\geq t}(\xi) - V_\phi(m_t))\, \nabla_\theta \log \pi_\theta(a_t|m_t)$.
- ☐ In actor–critic algorithms with an RNN actor as above, the critic has the correct policy value when $V_\phi(m) = \mathbb{E}[R_{\geq t}(\xi)|m_t = m]$ for each RNN state $m$.
- ☐ In value-based algorithms, a value network $Q_\theta(m_t, a_t)$ that pre-processes observations with $m_t = f_\theta(m_{t-1}, o_t)$ is at the optimal value when it satisfies the Bellman recursion, $Q_\theta(m, a) = \mathbb{E}[r + \max_{a'} Q_\theta(m', a')|m, a]$, for each RNN state $m$ and action $a$.