# CS 277 (W26): Control and Reinforcement Learning

# Quiz 5: Exploration and Partial Observability

## Due date: Monday, February 9, 2026 (Pacific Time)

Roy Fox
https://royf.org/crs/CS277/W26

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

**Question 1** In Multi-Armed Bandits, the regret grows (asymptotically) sub-linearly (check all that hold):

- ☐ If and only if the probability of taking an optimal action converges to 1.
- ☐ If and only if every action is almost surely (with probability 1) taken infinitely many times.
- ☐ If in each time step we take the action that is most likely to be optimal.
- ☐ With $\epsilon$-greedy exploration, if and only if $\epsilon$ converges to 0.

**Question 2** The following variables always satisfy the Markov property, i.e. $x_{<t}$ and $x_{>t}$ are independent given $x_t$ (and nothing else), where $x_t$ is:

- ☐ The world state $s_t$ in an MDP controlled by a history-based policy $\pi(a_t|s_{\leq t}, a_{<t})$.
- ☐ The observable history $h_t = (o_{\leq t}, a_{<t})$ in a POMDP controlled by a history-based policy $\pi(a_t|h_t)$.
- ☐ The Bayesian belief $b_t = p(s_t|o_{\leq t}, a_{<t})$ in a POMDP controlled by a belief-based policy $\pi(a_t|b_t)$.
- ☐ The agent's memory state $m_t = f(m_{t-1}, a_{t-1}, o_t)$ in a POMDP controlled by a memory-based policy $\pi(a_t|m_t)$.
- ☐ The full system state $x_t = (s_t, m_t)$ in a POMDP controlled by a memory-based policy $\pi(a_t|m_t)$ as above.

**Question 3** In a Partially Observable Markov Decision Process (POMDP) (check all that hold):

- ☐ If the rewards are provided by an external mechanism available in deployment, they can potentially carry useful information, and should therefore be included as part of the observations.
- ☐ The optimal belief-value function $V^*(b_t)$ is weakly convex in the Bayesian belief $b_t$, i.e. if $b_t = \alpha b + (1-\alpha)b'$ then $V^*(b_t) \geq \alpha V^*(b) + (1-\alpha)V^*(b')$.
- ☐ The difficulty of policy optimization is due to the rewards depending on a hidden state: if rewards only depended on the observations $r(o_t, a_t)$, it would be as easy as in an MDP.

**Question 4** Using RNNs in deep RL (check all that hold):

- ☐ The REINFORCE policy gradient with an RNN policy $\pi_\theta(a_t|m_t)$ using an RNN with cell $m_t = f_\phi(m_{t-1}, a_{t-1}, o_t)$, namely $g_\theta = \sum_t R(\xi)\, \nabla_\theta \log \pi_\theta(a_t|m_t)$, is unbiased regardless of $\phi$.
- ☐ A2C with an actor as above, i.e. $\pi_\theta(a_t|m_t)$ with RNN $f_\phi$, and a critic $V_\psi(m_t)$ has a policy gradient $g_\theta = \sum_t (R_{\geq t}(\xi) - V_\psi(m_t))\, \nabla_\theta \log \pi_\theta(a_t|m_t)$ that is unbiased regardless of $\phi$ or $\psi$.
- ☐ In actor–critic algorithms as above, the optimal baseline $V_{\pi_\theta}(m_t) = \mathbb{E}_{\xi \sim p_\theta}[R_{\geq t}(\xi)|m_t]$ satisfies the recursion $V_{\pi_\theta}(m_t) = \mathbb{E}_{p_{\theta,\phi}}[r_t + \gamma V_{\pi_\theta}(m_{t+1})|m_t]$ regardless of $\theta$ and $\phi$.
- ☐ In value-based algorithms, the optimal value network $Q_\theta(m_t, a_t)$ with RNN $f_\phi$ satisfies the Bellman recursion $Q_\theta(m_t, a_t) = \mathbb{E}[r_t + \gamma \max_{a_{t+1}} Q_\theta(m_{t+1}, a_{t+1})|m_t, a_t]$, regardless of $\phi$.