CS 277 (W24): Control and Reinforcement Learning
Quiz 7: Exploration and Partial Observability

Due date: Monday, March 4, 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 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 ϵ-greedy exploration, if and only if ϵ converges to 0.

Question 2 In a Partially Observable Markov Decision Process (POMDP) (check all that hold):

- ☐ If the rewards are provided by an external mechanism, they can potentially carry useful information, and should therefore be included as part of the observations.
- ☐ The agent's memory state m_t is Markov (i.e. $m_{<t}$ and $m_{>t}$ are independent given m_t) if and only if it is equivalent (maps bijectively) to the Bayesian belief $b_t = p(s_t|o_{\leq t}, a_{<t})$.
- ☐ The optimal belief-value function $V^*(b_t)$ is (weakly) convex in the Bayesian belief b_t.
- ☐ 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 3 Using RNNs in deep RL (check all that hold):

- ☐ REINFORCE with an RNN policy $\pi_\theta(a_t|m_t)$, with $m_t = f_\theta(m_{t-1}, o_t)$ (f_θ is called an *RNN cell*), 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 ξ) 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 π_θ as above, the true value function for the critic to learn is the expected future return given the memory state, $V^{\pi_\theta}(m_t) = \mathbb{E}_{\xi\sim p_\theta}[R_{\geq t}(\xi)|m_t]$.
- ☐ In value-based algorithms, a value network $Q_\theta(m_t, a_t)$ that pre-processes observations with an RNN $m_t = f_\theta(m_{t-1}, o_t)$ is optimal when it has no Bellman error for any state m_t and action a_t at time t, i.e. $Q_\theta(m_t, a_t) = \mathbb{E}[r + \max_{a_{t+1}} Q_\theta(m_{t+1}, a_{t+1})|m_t, a_t]$, where the expectation is over the observation o_t that determines m_{t+1}.