COS 435/ECE 433: Spring 2024 **Name:** ______________________

March 7, 11:00am - 12:20pm **Princeton email:** ______________________

Questions

1. (5 points) (single choice) What is the classical dilemma in the multi-armed bandit (MAB) problem?

A. **Exploration/Exploitation**
B. Expectation/Maximization
C. Minimization/Maximization
D. Exploration/Expectation

2. (5 points) In bandits, how does the ϵ-greedy algorithm balance between exploration and exploitation?

Your answer here:

The algorithm will explore at a probability of epsilon while exploiting at a probability of $(1 - \epsilon)$. This ensures that the algorithm can do both exploration and exploitation.

3. (10 points) Explain how the MAB and contextual bandits can be viewed as special cases of the MDP.

Your answer here:

MAB: Any chosen action only affects the immediate payoff but does not affect the future of the game. Contextual: Any action only affects the immediate reward distribution and does not affect future dynamics.

4. (5 points) What is the Markov assumption when modeling a stochastic process?

Your answer here:

The idea is that the stochastic process is memory-less. It is the assumption that conditional probability distribution of future states of the process depends only upon the present state, not on the history of events that happened before.

5. (5 points) What is the motivation to introduce a discounted factor γ in infinite-horizon RL?

Your answer here:

It is mathematically convenient because it converges to the finite solution (geometric series).

(5 points) For an infinite-horizon RL problem, using a discount factor of $\gamma \in [0, 1)$ corresponds to reasoning ______ steps into the future in expectation.

It corresponds to $\frac{1}{1-\gamma}$ steps.

6. (5 points) Imagine that you apply behavioral cloning to a large dataset of state-action pairs. The actions are discrete, and your policy's validation accuracy is 99.99%. However, upon deploying the policy, you find that it often takes incorrect actions, much more than 0.01% of the time. What likely explains this seeming inconsistency?

Your answer here:

There might be a significant distributional shift. The policy might not have learned appropriate actions if the dataset lacks diversity or certain scenarios are underrepresented.

(5 points) Building on the previous question, provide one way to mitigate this problem.

Your answer here:

Through data augmentation by synthetic data generation, a more diverse dataset is collected using clever hardware or interactive expert queries (DAGGER). GAIL is also a solution.

7. (5 points) Denote the optimal Q function as Q^* and the optimal value function V^*. When is $V^*(s)$ equal to $Q^*(s,a)$?

Your answer here:

For any state s, we have $Q^*(s,a) = V^*(s)$ if and only if $a = \arg\max_{a'} Q^*(s,a') = \pi^*(s)$.

8. (5 points) In a γ-discounted infinite-horizon MDP, the value iteration algorithm is guaranteed to converge to the optimal value function.

True

9. (5 points) The optimal value function corresponds to a unique optimal policy.

False

10. (10 points) Explain and compare policy evaluation and policy improvement.

Policy evaluation is ...

Policy evaluation involves iterative learning of the value function for one specific policy - in dynamic programming, this is done using the bellman equations, and in Monte Carlo, this is done by sampling returns.

Policy improvement is ...

Policy improvement is the step of updating a policy based on its value function, for example, choosing the greedy action at each step.

The connection between policy evaluation and policy improvement is ...

Policy iteration is the cyclic process of repeatedly running policy evaluation followed by policy iteration - a policy's value function is estimated, the policy is improved, then its value function is estimated again, etc.

11. (10 points) This question will look at how some reinforcement learning ideas apply to other ML areas where we want to optimize a sampling distribution. There's an area of ML known as variational inference where the objective function is

$$E_{q_\theta(z|x)}[\log p(x|z) - \log q_\theta(z|x)].$$

Write down the gradient of this objective w.r.t. θ for a fixed x. Your gradient should be in the form $E_{q_\theta(z|x)}[\,\cdots\,]$.

Your answer here:

$$
\begin{aligned}
&\nabla_\theta E_{q_\theta(z|x)}[\log p(x|z) - \log q_\theta(z|x)] \\
&= E_{q_\theta(z|x)} \nabla_\theta \log q_\theta(z|x)[\log p(x|z) - \log q_\theta(z|x)] - E_{q_\theta(z|x)}[\nabla_\theta \log q_\theta(z|x)] \\
&= E_{q_\theta(z|x)} \nabla_\theta \log q_\theta(z|x)[\log p(x|z) - \log q_\theta(z|x) - 1]
\end{aligned}
$$

12. (10 points) What is one advantage of using experience replay?

Your answer here:

First, make the training data more IID-like and enable better convergence. Second, make efficient use of previous experiences over multiple iterations.

13. (10 points) Write down the expected Bellman operators for arbitrary policy π, state s and action a in a γ-discounted infinite-horizon MDP. Fill in the blanks.

Your answer here:

$$V^\pi(s) = \underline{E_{a \sim \pi(\cdot|s)} Q^\pi(s,a)}$$

$$Q^\pi(s,a) = r(s,a) + \gamma \underline{E_{s' \sim P(\cdot|s,a)} V^\pi(s')}$$