University of Virginia
Department of Computer Science

CS 4501: Introduction to Reinforcement Learning
Fall 2022

Tuesday 10:30am-10:45am EDT, November 29th, 2022

Name:
ComputingID:

- This is a **closed book** and **closed notes** quiz. No electronic aids or cheat sheets or discussing the questions with anyone else are allowed.
- You are expected to finish this quiz within 15 minutes.
- There are 2 pages, 3 parts of questions, and 20 total points in this quiz.
- The questions are printed on the back of this page!
- Please carefully read the instructions and questions before you answer them.
- If you need any clarification of the quiz questions, please raise your hand and discuss with the instructor within the quiz period.
- Try to keep your answers as concise as possible; our grading is *NOT* by keyword matching.

Total	/20

1 True/False Questions (2×3 pts)

Please choose either True or False for each of the following statements. For the statement you believe it is False, please give your brief explanation of it (you do **NOT** need to explain when you believe it is True). Three point for each question. *Note: the credit can only be granted if your explanation for the false statement is correct.*

1. Temporal difference method introduces variance but reduces bias.
   ***False**, and Explain*: TD methods introduce bias.

2. In an episodic environment, the goal of policy-based RL algorithms is to maximize the expected return of initial states.
   True

2 Multiple Choice Questions (2×4 pts pts)

Please choose ALL the answers that you believe are correct for each question.

1. Which of the follow is/are off-policy RL method(s)? (c)
   (a) Sarsa; (b) REINFORCE; (c) Q-learning; (d) Actor-Critics.

2. What are the general principles for designing the policy in policy-based RL methods: (a) (b) (c)
   (a) differentiable; (b) non-deterministic; (c) easy to sample from; (d) additive.

3 Short Answer Question (6 pts)

The question can be answered by one or two sentences; so please make your answer concise and to the point.

1. What is the "maximization bias" in Q-learning, and how do we address it?

   In Q-learning, we use maximum (i.e., $\arg\max_{a\in A_t} Q(s_t, a)$) to choose an action and also treat it as an estimate of the maximum of the true value (i.e., $r_{t+1}+\gamma \max_{a\in A_{t+1}} Q(s_{t+1}, a)$). This leads to an exaggeration of the true value of the next state and is often referred to as the "maximization bias".
   A common way to handle "maximization bias" is to use double Q-learning, where we use separately estimated Q-values for action selection and value estimation.