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.

i) Temporal difference method introduces variance but reduces bias.

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

2. Multiple Choice Questions (2×4 pts pts) Please choose ALL the answers that you believe are correct for each question. Provide Explanations

i) Which of the follow is/are off-policy RL method(s)?

(a) Sarsa; (b) REINFORCE; (c) Q-learning; (d) Actor-Critics.

ii) What are the general principles for designing the policy in policy-based RL methods:

(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.

i) What is the “maximization bias” in Q-learning, and how do we address it?