

清华大学计算机系

强化学习 2024 秋

12 月 6 日 周五 15:20-15:35

姓名:
学号:

- 这是一次闭卷随堂小测，总时长 15 分钟。
- 小测总共有两页，三种不同类型的题目，总共 20 分。
- 请在作答前仔细阅读题目要求。
- 如果对题目描述有疑义或是觉得不清楚的地方，请随时举手询问任课教师。
- 请使用简短的语言回答每道题目。
- 作答时可以使用中文、英文、中英混杂。但请保证字迹清晰、以免造成不必要的误解。

Total	/20
-------	-----

1 判断题 (2×3 分)

如果认为题干的描述并不正确，请简要描述你的理由；如果认为题干正确，则不需要进一步解释。如果做出了正确的判断，但是解释错误，无法得分。（例如，题干的描述确实是错误的，但是你提供的答案中的解释同样是错误的，将无法得分。）

1. Temporal difference methods introduce variance but reduce bias.

False, and Explain: It introduces bias and reduces variance.

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

True

2 不定项选择题 (2×4 分)

请勾选你认为所有正确的选项。

1. Which of the follow is/are off-policy RL method(s)? (b) (d)
(a) Sarsa; (b) Q-learning; (c) REINFORCE; (d) Tree Backup.
2. What are the general principles for designing the policy in policy gradient type RL methods: (a) (b) (c) (d)
(a) differentiable; (b) non-deterministic; (c) easy to sample from; (d) full coverage of action space.

3 简答题 (6 分)

请用尽量简短的语句回答下列问题。

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

“maximization bias” in Q-learning happens as we used the maximum of the empirically estimated Q-value as the maximum of its true value, which is vulnerable to the outliers in the empirical observations. We can use double Q-learning to remove this bias, in which two Q-functions are learnt using isolated observations.