

清华大学计算机系

强化学习 2024 秋

12 月 27 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. We choose to optimize mean square error in value function approximation because it is guaranteed to lead to improved policy.

False, and Explain: We choose to optimize mean square error only because of its mathematical simplicity; instead, Bellman error is the ideal objective for function approximation, but it is computationally infeasible to optimize.

2. Proper design in experience replay helps us break correlation in RL's observation sequence.

False, and Explain: It helps us to reduce variance in parameter estimation, just as batch update in SGD.

2 不定项选择题 (2×4 分)

请勾选你认为所有正确的选项。

1. What constitute the so-called deadly triad? (a), (c), (d)
(a) function approximation; (b) non-stationarity; (c) off-policy learning;
(d) bootstrapping.
2. What are the key design choices for AlphaGo? (a), (b), (c), (d)
(a) Supervised policy learning; (b) Self-play; (c) Monte Carlo tree search;
(d) CNN for state encoding.

3 简答题 (6 分)

请用尽量简短的语句回答下列问题。

1. You have decided to start your business in robotics industry, and your first proposal is to design an RL-based solution for a robot vacuum (扫地机器人). Now how are you going to modernize your robot vacuum with deep RL techniques?

We can take the robot's current position as state (e.g., the position of the base station as the initial state), the action is to move left, right, forward, or backward with a fixed step-size. The reward is +1 if we visited an unvisited state, otherwise -1. We then use neural networks to enhance the state representation, Q-function and/or the policy. Basically, neural networks provide us a flexible way to encode the problem.

For example, we can use CNN to encode imagery sensor input as part of the state, and use an RNN to encode the traveled trajectory to enhance state representation. For the value function and policy, we can use an MLP to map state action pairs into value or probability estimates.