

清华大学计算机系

强化学习 2024 秋

10 月 18 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. Reward is only a relative concept, and thus simply scaling or shifting it will not affect the learning of optimal policy.

False, and Explain: 正如我们课上最短路径的例子，reward 上的操作将直接影响 optimal policy 的结果。

2. All goals can be described by the maximization of expected cumulative reward.

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. What make(s) reinforcement learning different from supervised learning? **c, d**
(a) observations arrive sequentially; (b) complexity of output space;
(c) algorithm's output influences future data distribution;
(d) the need of exploration.
2. What make(s) the UCB1 algorithm better than the Explore-then-Commit algorithm?
a, b, c
(a) more accurate estimation of rewards; (b) more adaptive exploration;
(c) more flexible hyper-parameter setting; (d) easier to implement.

3 简答题 (6 分)

请用尽量简短的语句回答下列问题。

1. Why is there a trade-off between exploration and exploitation in reinforcement learning?

由于环境未知，agent 需要尝试不同 action 以获得对于它们 value 的更好估计，方便更好地进行 exploitation，从而减小未来的 regret。而探索往往意味着当下需执行次优 action，从而积累 regret。