

清华大学计算机系

强化学习 2025 秋

11 月 21 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. Once $\gamma < 1$, the expected return of any state is finite, and thus the Monte Carlo prediction method can be applied to any environment.

False, and Explain: MC method can only be applied to episodic environments.

2. Value iteration method guarantees monotonic policy improvement.

True

2 不定项选择题 (2×4 分)

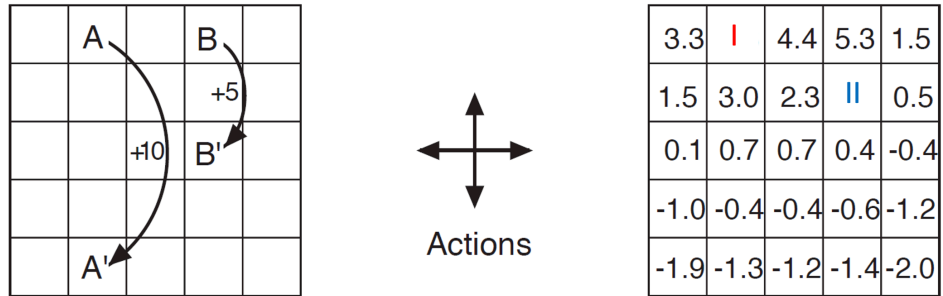
请勾选你认为所有正确的选项。

1. Off-policy MC control often faces much more significant practical challenges than on-policy MC control does not, why is that? (a)
(a) high variance in importance sampling based value estimation can cause unstable policy improvement;
(b) it requires the behavior policy b to be deterministic, which severely limits exploration of the state-action space;
(c) it cannot be used for problems with continuous action spaces;
(d) it can only learn about stochastic policies, making it unsuitable for finding the optimal deterministic policy.
2. Where can policy evaluation be useful: (a)(b)(c)(d)
(a) obtain the value of each state; (b) compare different policies;
(c) obtain the value of each action under a state; (d) improve the policy.

3 简答题 (6 分)

请用尽量简短的语句回答下列问题。

1. In the following grid world environment, at each position, the agent can take one of the four actions: up, down, right and left; and each action moves the agent along the direction by one step. When the move hits the wall, the agent stays at its current position and receives reward -1; when the agent moves onto position A, it will be transit to position A' and receive reward +10; when the agent moves onto position B, it will be transit to position B' and receive reward +5; otherwise, no reward will be given. Given the state values shown in the right figure, please compute the missing values for state I and II under a uniformly random policy with $\gamma = 0.9$.



Write out the Bellman equation for state I and II and solve the corresponding value. $v(s_I)$ depends on your interpretation of the state transition, and we will only assess the correctness of your Bellman equation; but $v(s_{II})$ has to be $v(s_{II}) = 0.25 * 5 + 0.9 * 5.3 + 0.25 * 0.9 * (2.3 + 0.5 + 0.4) = 3.2$.