# 上海科技大学

# 2024-2025 强化学习应用实践

Project1 Part-A

2025 年 2 月 26 日

- **状态集合** $S$: $S = \{s_1, s_2, s_3\}$
- **动作集合** $A$: $A = \{a_1, a_2\}$
- **终止状态**: $s_3$ 为终止状态，不再执行动作。
- **转移概率** $P(s'|s, a)$:
  - 从 $s_1$:
$$P(s_2|s_1, a_1) = 0.5, \quad P(s_3|s_1, a_1) = 0.5$$
$$P(s_2|s_1, a_2) = 0.7, \quad P(s_3|s_1, a_2) = 0.3$$
  - 从 $s_2$:
$$P(s_1|s_2, a_1) = 0.6, \quad P(s_3|s_2, a_1) = 0.4$$
$$P(s_1|s_2, a_2) = 0.8, \quad P(s_3|s_2, a_2) = 0.2$$
  - 从 $s_3$: 无后续转移（所有 $P(s'|s_3, a) = 0$）。
- **奖励函数** $R(s, a)$:
$$R(s_1, a_1) = 1, \quad R(s_1, a_2) = 2$$
$$R(s_2, a_1) = 3, \quad R(s_2, a_2) = 0$$
$$R(s_3, a) = 0 \quad \forall a$$
- **折扣因子** $\gamma$: 0.9
- **初始策略** $\pi(s, a) = \frac{1}{|A|}$: 所有非终止状态的动作选择均匀随机：
$$\pi(s_1, a_1) = \pi(s_1, a_2) = 0.5, \quad \text{其余同理。}$$
- **初始值函数/状态-动作值函数**:
$$V(s) = 0 \quad \forall s, \quad Q(s, a) = 0 \quad \forall s, a$$
- **学习率** $\alpha = 0.1$。

**1. 策略迭代 (20 分)** 根据初始策略 $\pi(s,a)=\frac{1}{2}$，手动推导策略迭代第一轮的迭代步骤，包括策略评估和策略改进，写出：

1. 第一轮策略评估的状态值函数 $V(s)$。(10 分)

2. 改进后的新策略 $\pi'(s)$。(10 分)

**2. 价值迭代 (20 分)** 根据给定的环境和初始值函数 $V(s) = 0 \forall s$，推导价值迭代第一轮的迭代步骤，写出：

1. 每个状态的值函数 $V(s)$。(10 分)
2. 相应的策略 $\pi(s)$。(10 分)