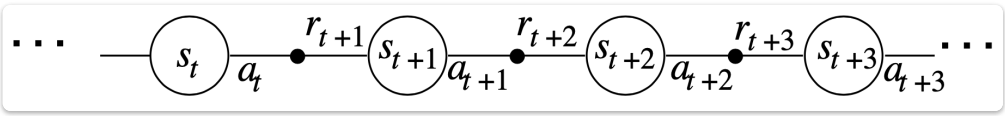
lec06

马尔科夫性

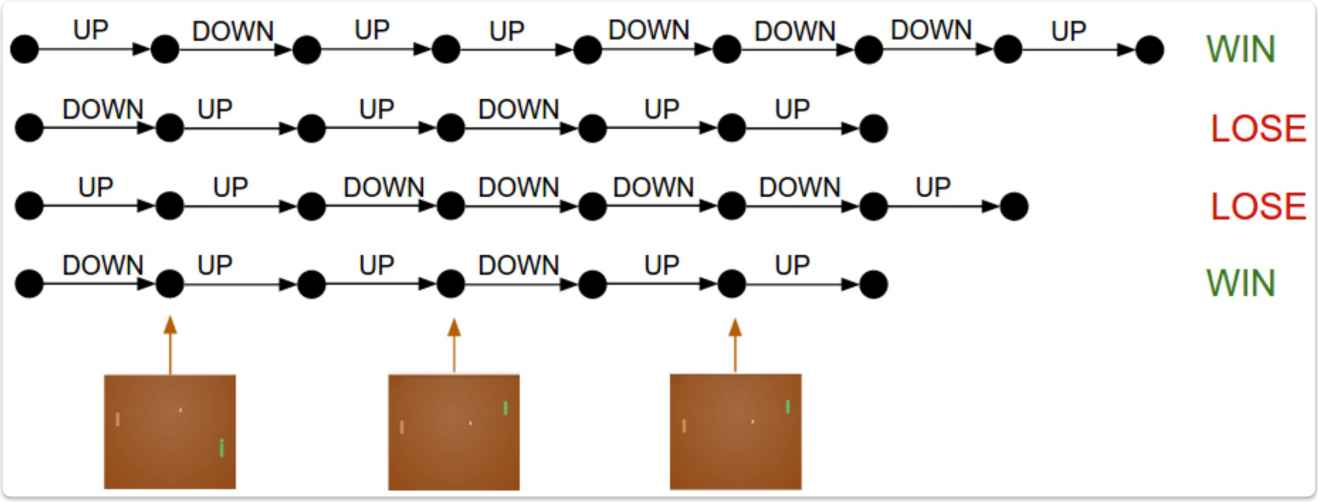
Agent-Environment Interface

- 代理和环境在离散时间步上交互：$t = 0, 1, \ldots, K$
- 代理在 t 步观察状态：$s_t \in S$
  - 在 t 步执行动作：$a_t \in A(s_t)$
  - 得到 reward：$r_t \in \Re$
  - 下一个状态：s_{t+1}



Agent 学习策略

- 在 t 步的策略 π_t：
  - 状态到动作概率的映射
  - $\pi_t(s, a)$ 是 $s_t = s$ 时 $a_t = a$ 的概率
- 强化学习方法规定了代理如何根据经验改变其策略
- 代理的目标大致是尽可能多地获得长期奖励



目标与奖励

- 标量奖励信号是否足够表示目标?
  - 可能不够，但却非常灵活
- 目标应指定我们想要实现的，而不是我们想要如何实现它
- 目标必须在代理的直接控制之外
- 代理必须能够衡量成功：
  - 显式地
  - 经常在其生命周期中

- 奖励假设：所有我们认为的目标和目的都可以被视为最大化累积收到的标量信号（奖励）的总和

Returns

离散

假设在 t 步之后的奖励序列是：$r_{t+1}, r_{t+2}, \ldots$ 一般来说，我们想要最大化每一步 t 的期望回报

对于分段任务：交互自然分成多个阶段，例如玩游戏，走迷宫

$$R_t = r_{t+1} + r_{t+2} + \ldots + r_T$$

其中 T 是是达到终止状态时的最后一步，这意味着结束一个阶段

连续

$$R_t = r_{t+1} + \gamma r_{t+1} + \gamma^2 r_{t+3} + \ldots = \sum_{k=1}^{\infty} \gamma^k r_{t+k+1}$$

其中 γ 是折扣率 discount rate，且

- $\gamma \to 0$：短视
- $\gamma \to 1$：远视

例子

对于平衡杆任务（Pole Balancing）

- 避免失败：杆倾斜超过临界角度或小车撞到轨道末端
  - 作为一个分段任务，其中阶段在失败时结束：
    - 每步奖励为 +1 直到失败
    - 回报 = 失败前的步数
  - 作为一个持续任务，带有折扣回报：
    - 失败时奖励为 -1；否则为 0
    - 对于失败前的 k 步，回报为 $-\gamma^k$
- 无论哪种情况，通过尽可能长时间地避免失败来最大化回报

符号

- 在分段任务中，我们从 0 开始编号每个阶段的时间步
  - 通常我们不必区分各个阶段，因此我们用 s_t 表示第 t 步的状态，而不是 $s_{t,j}$
  - 同样地，我们用 R_t 表示总回报，而不是 $R_{t,j}$
- 因此我们可以通过

  $$R_t = \sum_{k=1}^{\infty} \gamma^k r_{t+k+1}$$

  囊括所有可能的情况，这也包括 $T = \infty$ 或 $\gamma = 1$；但并非 both

马尔可夫性

- 在第 t 步的“状态”指的是代理在第 t 步可以获得的有关环境的所有信息

- 状态可以包括
  - 直接的“感知”
  - 高度处理过的感知
  - 感知序列中积累起来的结构
- 理想情况下，一个状态应总结过去的感知，以保留所有“重要”信息，即应具有马尔可夫性：

$$P\left(s_{t+1}=s', r_{t+1}=r \mid \{s_i\}_{i=0}^{t}, \{a_i\}_{i=0}^{t}, \{r_i\}_{i=0}^{t}\right) = P(s_{t+1}=s', r_{t+1}=r \mid s_t, a_t)$$

  对于所有的 s' 和 r，以及所有的历史信息 $\{s_i\}_{i=0}^{t}, \{a_i\}_{i=0}^{t}, \{r_i\}_{i=0}^{t}$

马尔可夫决策过程 Markov Decision Process, MDP

- 如果一个强化学习任务具有马尔可夫性质，它基本上就是一个马尔可夫决策过程（MDP）
- 如果状态集和动作集是有限的，那么它就是一个有限的MDP
- 要定义一个 MDP，我们需要
  - 定义动作集 $A(s)$
  - 定义状态集 S
  - 定义一步动态 one-step dynamics：

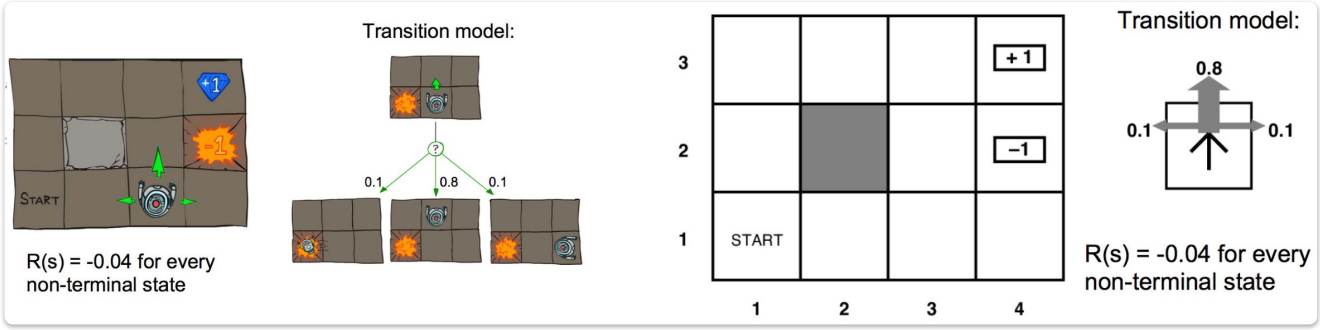
$$P_{ss'}^{a} = \Pr\{s_{t+1}=s' \mid s_t, a_t\}, \quad \forall s, s' \in S, a \in A(s)$$

    这表示状态 s_t 采取动作 a_t 后转移到 s' 的概率
  - 定义奖励期望：

$$R_{ss'}^{a} = \mathbb{E}\{r_{t+1} \mid s_t=s,\ a_t=a,\ s_{t+1}=s'\}, \quad \forall s, s' \in S, a \in A(s)$$

    这表示在状态 s 采取动作 a 并转移到状态 s' 后得到的奖励的期望

MDP 例子



这包括了

- 状态集
- 动作集
- 回报函数
- 转移函数