本章目录

智能博弈与强化学习基础

第十章 集群协同与博弈对抗

10.1 集群协同基础与控制架构

10.2 一致性与编队控制

10.3 智能博弈与强化学习基础

10.4 单智能体强化学习经典算法

10.5 多智能体强化学习与规模扩展

10.6 集群关键技术与挑战

10.7 RflySim集群框架与接口

10.8 本章实验案例与综合验证

10.9 习题与课堂讨论

本节 4 个教学单元

10.3.1 集群智能控制框架

教材对应：第10.1.3.1节

10.3.2 强化学习与MDP

教材对应：第10.1.3.2–.3节

10.3.3 回报与价值函数

教材对应：第10.1.3.3节第3–4项

10.3.4 Bellman与策略优化

教材对应：第10.1.3.3节第5项

10.3.1 集群智能控制框架：任务与求解方法

博弈任务与策略学习的关系

- **博弈：**一个决策者的收益会受其他决策者的选择影响，可合作、竞争或兼有。
- **强化学习：**智能体通过交互取得奖励，用经验调整动作选择策略。
- **本节对象：**先固定环境规则，研究单个决策者如何比较当前动作的长期后果。

先定义谁获得什么收益，才有可优化的策略。

10.3.1 集群智能控制框架：图10.4控制框架

智能策略位于任务与控制之间

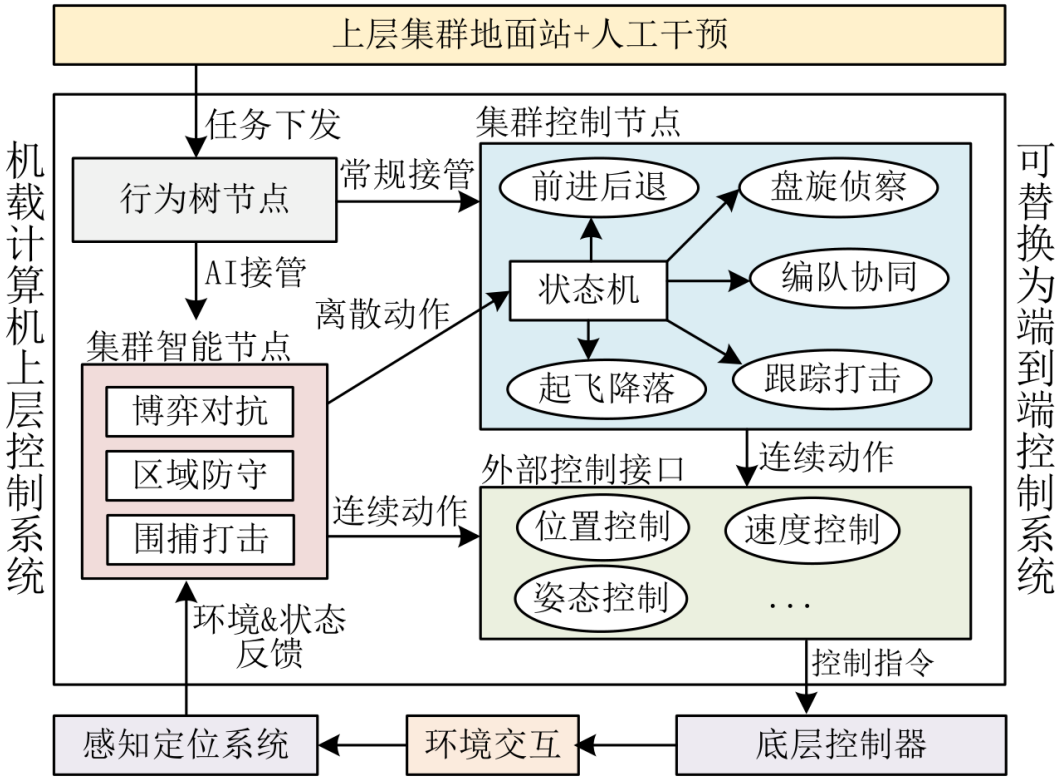


图10.4 典型智能集群控制框架

- **任务层：** 行为树依条件激活节点。
- **决策层：** 智能节点选动作或参考。
- **控制层：** 跟踪参考并反馈状态。

10.3.1 集群智能控制框架：图10.4的任务分支

行为树决定何时启用智能节点

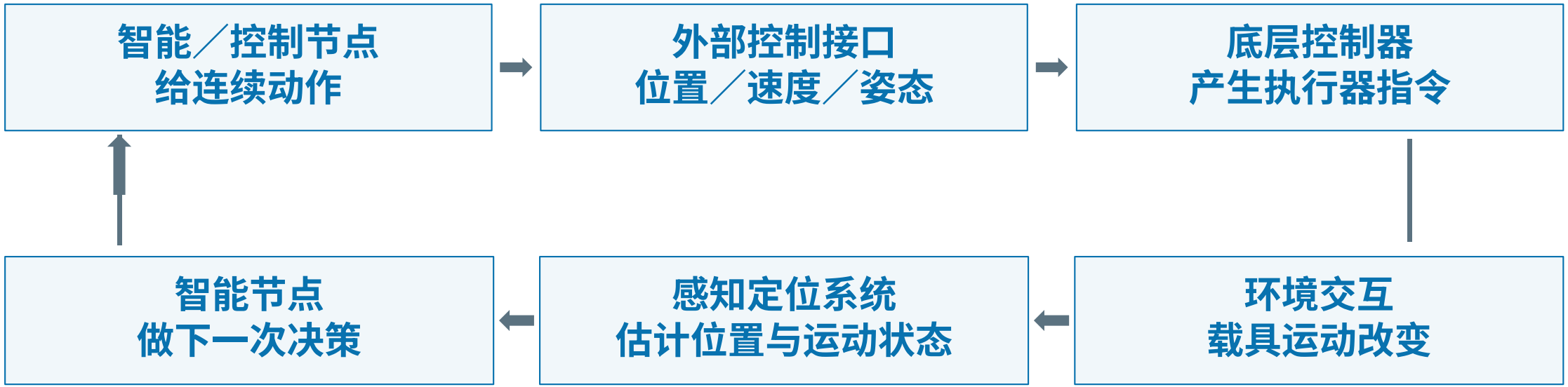
地面站与人工配置下发任务 → 行为树按条件选择运行分支 → 控制／智能节点产生当前行为

行为树分支	决策内容	如何执行
常规接管	选择已定义的控制行为	控制节点生成运动参考
智能接管	学习策略输出离散动作	状态机判定允许转换并激活行为

状态机可组织起飞降落、前进后退、盘旋侦察、编队协同和目标跟踪等行为。

10.3.1 集群智能控制框架：图10.4的控制反馈

运动参考经接口进入反馈闭环



位置、速度、姿态参考对应不同控制模式；反馈使当前运动影响下一次决策。

10.3.1 集群智能控制框架：动作落地

离散行为与连续控制的接口

智能策略
选行为／运动参考 → 状态机与接口
模式判定／限幅 → 底层控制器
跟踪位置或速度

动作形式	例子	交给接口前的约定
离散行为	前进、等待、返回	允许模式、结束条件、参考生成规则
连续参考	位置、速度、姿态	坐标系、单位、范围与发送周期

例如：策略选择“前进”，接口生成前向速度，控制器依据速度误差调节执行器。

10.3.1 集群智能控制框架：模块化与端到端

端到端方法的作用与代价

设计维度	模块化方案	端到端方案
中间语义	行为、模式、运动参考可逐层检查	模型直接学习输入到控制的映射
算力与能耗	可按模块分配计算预算	模型规模、时延与机载功耗耦合
可靠性与解释	可定位到模块和接口	内部决策原因与异常定位较困难
任务复杂性	三维控制、多执行器及动态环境	同样受动力学和环境不确定性约束

本章目录

智能博弈与强化学习基础

第十章 集群协同与博弈对抗

10.1 集群协同基础与控制架构

10.2 一致性与编队控制

10.3 智能博弈与强化学习基础

10.4 单智能体强化学习经典算法

10.5 多智能体强化学习与规模扩展

10.6 集群关键技术与挑战

10.7 RflySim集群框架与接口

10.8 本章实验案例与综合验证

10.9 习题与课堂讨论

本节 4 个教学单元

10.3.1 集群智能控制框架

教材对应：第10.1.3.1节

10.3.2 强化学习与MDP

教材对应：第10.1.3.2–.3节

10.3.3 回报与价值函数

教材对应：第10.1.3.3节第3–4项

10.3.4 Bellman与策略优化

教材对应：第10.1.3.3节第5项

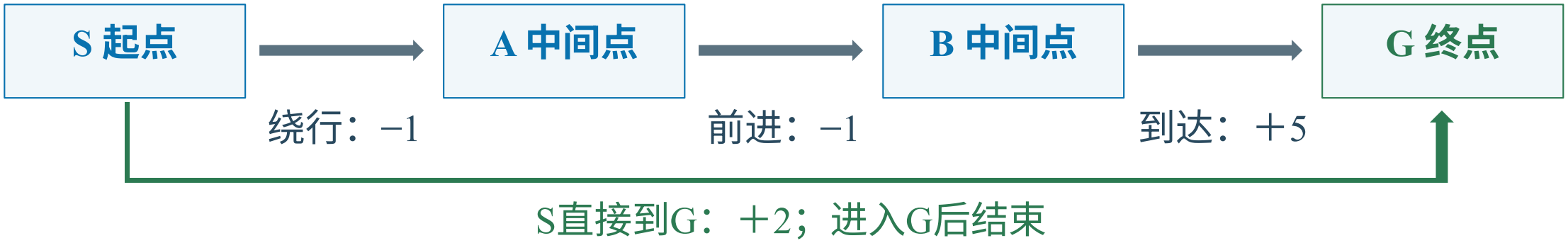
10.3.2 强化学习与MDP：概念总览

强化学习的五个基本要素

要素	含义	导航任务中的例子
状态	决定后续转移所需的信息	当前位置及影响运动的任务变量
动作	当前允许执行的选择	直接到达或经中间点绕行
奖励	每次转移后的数值反馈	途中代价、到达收益
策略	按状态选择动作的规则	S点以某概率选择绕行
价值	遵循某策略的期望长期回报	从S出发平均可得多少分

10.3.2 强化学习与MDP：贯穿算例

四状态导航任务



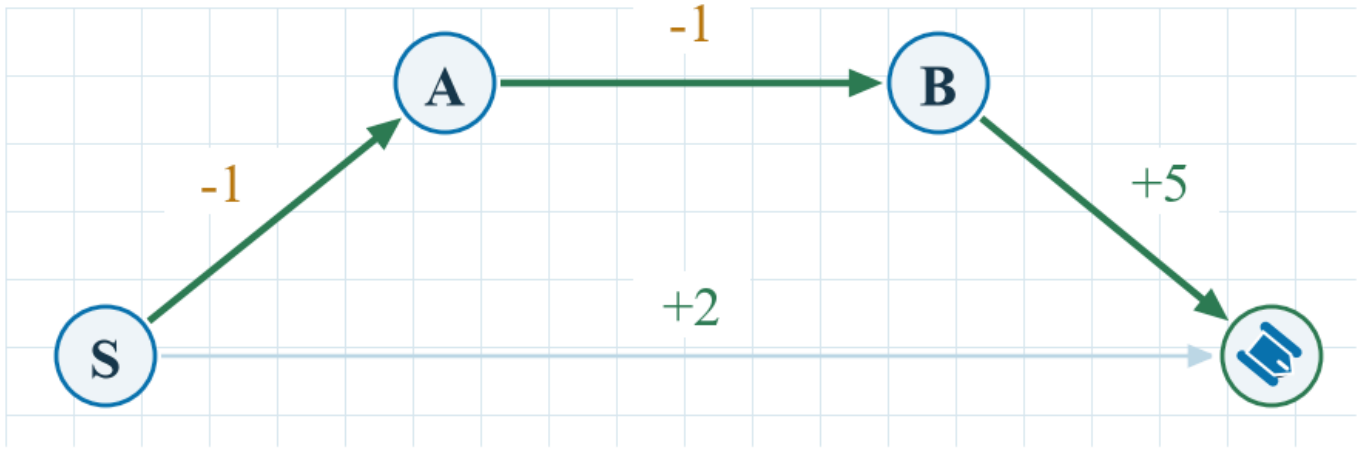
- **教学假设：**位置构成完整状态，转移确定，无障碍、无惯性、无测量误差。
- **动作约束：**S可直接到达或绕行；A、B仅有前进；G为终止状态。

奖励单位为“分”，每条箭头代表一步；折扣因子 $\gamma = 0.9$。

10.3.2 强化学习与MDP：动作与经验

沿着同一张地图，把动作变成一条经验

一次移动，一条经验



每段路径代表一次动作

即时奖励

$B \rightarrow G \quad r = +5$

到达即终止

折扣回报 $\gamma = 0.9$

直达 2.00

绕行 2.15

图：直达与绕行的状态转移；路径上的数值为该次动作的奖励

每到一个状态重新选动作；到达终点结束回合，再从起点开始。

10.3.2 强化学习与MDP：式10.17

马尔可夫决策过程（MDP）

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma) \tag{10.17}$$

对象	数学含义	本例
状态／动作	$\mathcal{S}$：状态集合；$\mathcal{A}$：动作集合	四个状态；S有两种选择
转移概率	P：给定状态和动作的下一状态分布	绕行时：S到A的概率为1
奖励函数	R：一次转移奖励的数值或条件期望	S到A为−1分；B到G为5分
折扣因子	γ：每延后一步多乘一次的因子	本例为0.9

10.3.2 强化学习与MDP：式10.18与状态扩展

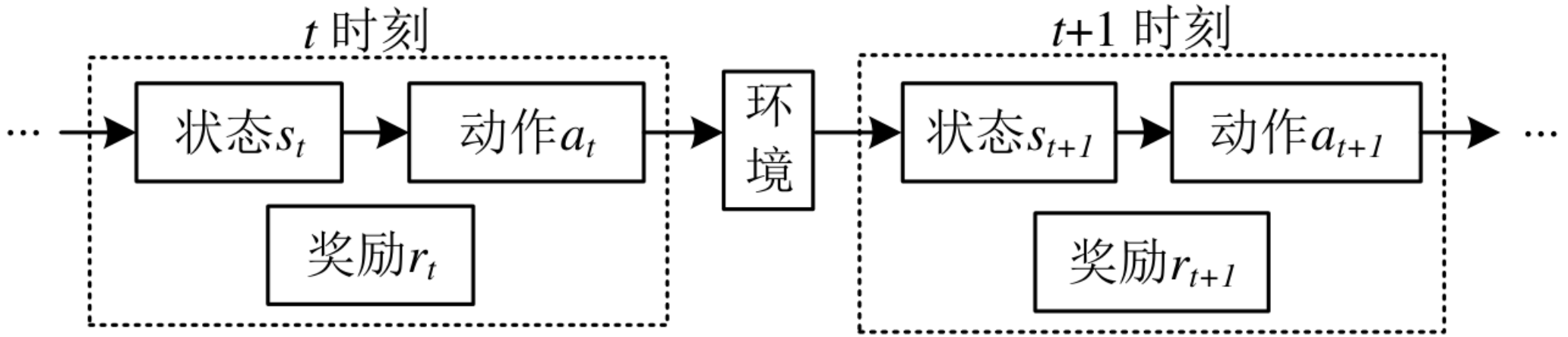
马尔可夫性要求状态信息足够

$$P(s_{t+1} \mid s_t, a_t, s_{t-1}, a_{t-1}, \dots) = P(s_{t+1} \mid s_t, a_t) \tag{10.18}$$

- **位置相同、速度不同：**下一步运动可能不同；只用位置可能遗漏必要信息。
- **状态扩展：**加入速度、任务进度或过去观测，让当前表示包含更多决策依据。
- **历史编码：**RNN为循环神经网络；LSTM用门控记忆处理时序信息，但仍需验证表示。

10.3.2 强化学习与MDP：图10.5交互闭环

一次交互产生一条经验



(a) 马尔科夫决策过程模型

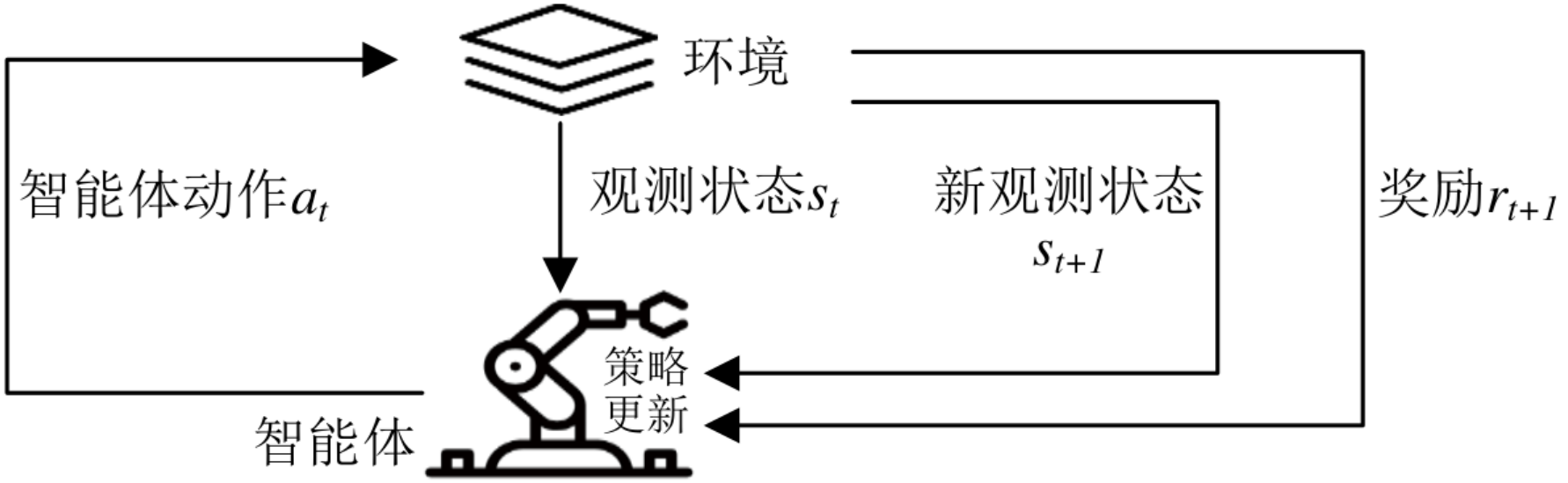
图10.5(a) 马尔可夫决策过程模型

$$(s_t, a_t, r_{t+1}, s_{t+1}, d_{t+1})$$

终止标志为1表示任务真正结束；固定策略执行时仍有反馈，但可以不更新参数。

10.3.2 强化学习与MDP：图10.5强化学习框架

训练把环境反馈转为策略改进



(b) 强化学习框架

图10.5(b) 强化学习框架

环境反馈每一步都存在；训练阶段利用反馈更新，固定策略执行时只计算动作。

10.3.2 强化学习与MDP：随机环境

概率转移怎样形成期望值

假设的结果	概率	给定下一状态价值	即时奖励
A前进后到B	0.8	5分	−1分
A前进后留在A	0.2	3.5分	−1分

$$0.8[-1+0.9\times 5]+0.2[-1+0.9\times 3.5]=3.23$$

一步估计：按概率平均“即时奖励加折扣后续价值”，得3.23分。
仅改变A的一步转移且给定价值估计；后续仍回到确定性导航任务。

本章目录

智能博弈与强化学习基础

第十章 集群协同与博弈对抗

10.1 集群协同基础与控制架构

10.2 一致性与编队控制

10.3 智能博弈与强化学习基础

10.4 单智能体强化学习经典算法

10.5 多智能体强化学习与规模扩展

10.6 集群关键技术与挑战

10.7 RflySim集群框架与接口

10.8 本章实验案例与综合验证

10.9 习题与课堂讨论

本节 4 个教学单元

10.3.1 集群智能控制框架

教材对应：第10.1.3.1节

10.3.2 强化学习与MDP

教材对应：第10.1.3.2–.3节

10.3.3 回报与价值函数

教材对应：第10.1.3.3节第3–4项

10.3.4 Bellman与策略优化

教材对应：第10.1.3.3节第5项

10.3.3 回报与价值函数：式10.19–10.20

即时奖励与折扣回报

$$R_{\text{total}} = r_{t+1} + r_{t+2} + \cdots + r_n \tag{10.19}$$

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+1+k} \tag{10.20}$$

- **有限回合：**在时刻 n 终止；终止后奖励补0，可写为无穷和。

R_t 为折扣回报；k 从0起计，第一项不折扣。

- **持续任务：**有界奖励且折扣小于1时，折扣回报有限；折扣决定时间权重。

10.3.3 回报与价值函数：导航回报

完整路径的回报可以直接复算

路径	奖励序列	步数	未折扣回报
S直接到G	2	1	2分
S经A、B到G	−1，−1，5	3	3分

$$R_0(\gamma = 0.9) = -1 - 0.9 + 0.9^2 \times 5 = 2.15$$

$$R_0(\gamma = 0.5) = -1 - 0.5 + 0.5^2 \times 5 = -0.25$$

10.3.3 回报与价值函数：式10.21

动作价值回答先选这一步值多少

$$Q_\mu(s_t, a_t) = \mathbb{E}_\mu\left[\sum_{k=0}^{\infty} \gamma^k r_{t+1+k} \mid s_t, a_t\right] \tag{10.21}$$

μ 为策略，$\mathbb{E}$ 为概率加权平均；竖线右侧为给定条件。

固定起点S
固定第一步动作
→
后续按策略执行
环境可能随机
→
多种回报取期望
得到该动作的Q值

$$Q_\mu(S, \text{direct}) = 2, \quad Q_\mu(S, \text{via}) = 2.15$$

本例中direct表示直接到G，via表示经A、B绕行；A、B没有其他动作。

10.3.3 回报与价值函数：式10.23与10.25

状态价值还要平均第一步选择

$$V_\mu(s_t) = \mathbb{E}_\mu\left[\sum_{k=0}^{\infty}\gamma^k r_{t+1+k} \mid s_t\right] \tag{10.23}$$

$$V_\mu(s) = \mathbb{E}_{a\sim\mu(\cdot|s)}[Q_\mu(s,a)] \tag{10.25}$$

$$V_\mu(S) = 0.5\times 2 + 0.5\times 2.15 = 2.075$$

随机策略：同一状态按概率选择动作。确定性策略：同一状态总选同一动作。

10.3.3 回报与价值函数：式10.22

Bellman关系把未来拆成一步

$$R_t = r_{t+1} + \gamma R_{t+1}$$

s', a' 是下一状态与下一动作，$r = r_{t+1}$ 为本次转移奖励。

$$Q_\mu(s,a) = \mathbb{E}_{r,s'|s,a}\left[r + \gamma \mathbb{E}_{a'\sim\mu(\cdot|s')} Q_\mu(s',a')\right] \tag{10.22}$$

- **外层期望：** 平均环境可能产生的下一状态与奖励。
- **内层期望：** 平均策略在下一状态可能选择的动作；后续仍遵循同一策略。

10.3.3 回报与价值函数：式10.24与数值递推

从终点倒推各状态价值

$$V_\mu(s) = \mathbb{E}_{a\sim\mu(\cdot|s)}\mathbb{E}_{r,s'|s,a}[r + \gamma V_\mu(s')] \tag{10.24}$$

状态	给定绕行策略下的计算	价值
G	终止后没有后续奖励	0分
B	$5 + 0.9 \times 0$	5分
A	$-1 + 0.9 \times 5$	3.5分
S	$-1 + 0.9 \times 3.5$	2.15分

10.3.3 回报与价值函数：建模选择

奖励怎样改变最优行为

- **到达收益：**定义完成任务的价值；奖励过于稀疏时，探索可能很久看不到正反馈。
- **过程代价：**时间、能耗、越界等指标影响策略取舍，权重决定优先级。
- **终止与截断：**真正终止后价值为0；仅因采样截断时，按原任务保留后续价值。
- **有限时间任务：**若期限本身决定任务结束，剩余时间属于决策状态的一部分。

本章目录

智能博弈与强化学习基础

第十章 集群协同与博弈对抗

10.1 集群协同基础与控制架构

10.2 一致性与编队控制

10.3 智能博弈与强化学习基础

10.4 单智能体强化学习经典算法

10.5 多智能体强化学习与规模扩展

10.6 集群关键技术与挑战

10.7 RflySim集群框架与接口

10.8 本章实验案例与综合验证

10.9 习题与课堂讨论

本节 4 个教学单元

10.3.1 集群智能控制框架

教材对应：第10.1.3.1节

10.3.2 强化学习与MDP

教材对应：第10.1.3.2–.3节

10.3.3 回报与价值函数

教材对应：第10.1.3.3节第3–4项

10.3.4 Bellman与策略优化

教材对应：第10.1.3.3节第5项

10.3.4 Bellman与策略优化：式10.26与10.29

最优价值与最优动作

$$Q^*(s,a) = \max_{\mu} Q_\mu(s,a), \qquad V^*(s) = \max_{\mu} V_\mu(s) \tag{10.26}$$

$$\mu^*(s) = \operatorname{argmax}_{a} Q^*(s,a) \tag{10.29}$$

计算	本例结果	含义
最大值max	2.15分	状态或动作的最优期望回报
最优动作arg max	via：绕行	实际选用的动作

10.3.4 Bellman与策略优化：式10.27

最优Q递推在下一步选择最佳动作

$$Q^*(s,a) = \mathbb{E}_{r,s'|s,a}\left[r + \gamma \max_{a'} Q^*(s',a')\right] \tag{10.27}$$

$$Q^*(S,\text{via}) = -1 + 0.9 \times 3.5 = 2.15$$

- **可控部分：**在下一状态选择动作，用最大值表示最优决策。
- **随机部分：**环境转移和奖励按分布取期望，不能只挑最幸运的结果。

10.3.4 Bellman与策略优化：式10.28

最优V递推比较当前全部动作

$$V^*(s) = \max_a \mathbb{E}_{r,s'|s,a}\left[r + \gamma V^*(s')\right] \tag{10.28}$$

$$V^*(S) = \max\{2 + 0.9 \times 0,\ -1 + 0.9 \times 3.5\} = 2.15$$

动作在环境随机结果出现前选择；每个动作先算期望，再比较大小。

长期价值把当前代价与后续收益放到同一尺度。

10.3.4 Bellman与策略优化：完整数值迭代

价值迭代让终点收益逐步传播

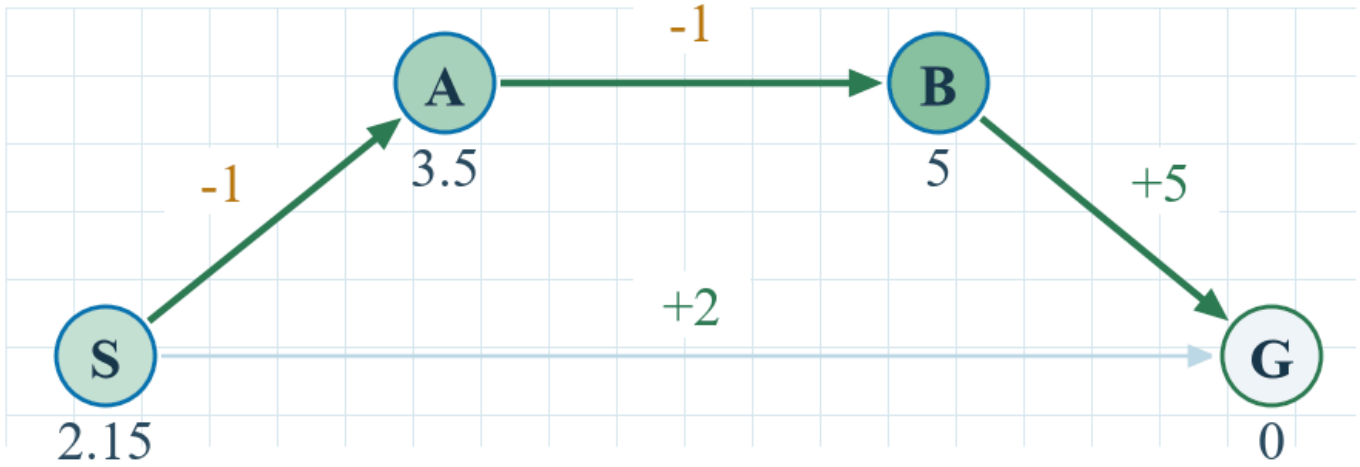
同步轮次	S价值	A价值	B价值	S的贪婪动作
0：全零初值	0	0	0	尚未选择
1	2	−1	5	直接到G
2	2	3.5	5	直接到G
3	2.15	3.5	5	经A绕行
4：不再变化	2.15	3.5	5	经A绕行

每轮使用上一轮价值；G始终为0。最终路径S→A→B→G，回报2.15分。

10.3.4 Bellman与策略优化：价值传播

终点收益怎样改变起点的选择

收益沿转移关系回传



圆内颜色表示价值

同步更新 4

$\gamma = 0.9$

起点当前选择

绕行

$V(G) = 0$

$V(S) = 2.15$

图：同步价值迭代；圆内颜色与读数为当前轮价值，金色点表示回传方向

先更新B，再影响A，最后改变S的选择；每轮只读取上一轮价值。

10.3.4 Bellman与策略优化：两种求解方式

策略迭代交替评价与改进

固定当前策略 求其V或Q → 比较每个动作 选择更高价值 → 重新评价新策略 直到策略稳定

阶段	S的策略	S的价值	改进依据
评价初始策略	直接到G	2分	A、B价值为3.5、5
改进并再评价	经A绕行	2.15分	$-1+0.9\times3.5>2$

有限状态动作、奖励有界且 $0\leq\gamma<1$ 时，同步精确价值迭代收敛到唯一最优价值。

10.3.4 Bellman与策略优化：训练与执行

模型规划与经验学习的分界

环节	使用的信息	得到的结果
模型规划	已知转移概率和奖励函数	价值迭代或策略迭代的解
交互训练	实际采样的状态、动作、奖励	估计价值并更新策略参数
固定策略评估	未用于更新的多个回合	回报、成功率与失败分布
执行控制	当前观测和固定策略	动作送至接口并由控制器跟踪

练习与讨论

课堂练习：回报与状态是否充分

- **问题1：**保持奖励不变，折扣改为0.8。两条路径的回报各是多少？哪条更优？
- **问题2：**恢复折扣0.9。S点以1/4概率直接到达、3/4概率绕行，状态价值是多少？
- **问题3：**两次观测位置相同、速度相反。仅凭位置能否决定下一状态的分布？

计算依据：绕行奖励为−1、−1、5；直接到达得2分；到G真正终止。

参考要点

参考要点：折扣、平均与状态

$$R_{\text{via}} = -1 - 0.8 + 0.8^2 \times 5 = 1.4 < 2$$

$$V_\mu(S) = \frac{1}{4} \times 2 + \frac{3}{4} \times 2.15 = 2.1125$$

- **问题3参考：**位置通常不足以描述运动；可加入速度或用历史估计隐含状态。
- **判据：**给定当前表示和动作之后，历史信息是否仍改变下一状态分布？

本节小结

状态、回报与策略优化

- **系统位置：**智能策略生成行为或参考，外部控制接口和底层控制器完成运动执行。
- **建模方法：**状态要包含必要信息，奖励与折扣定义目标，终止决定后续价值。
- **价值关系：**回报是路径结果；Q固定第一步动作，V按策略平均动作价值。
- **易错点：**奖励不是回报；策略评价不同于取最大值；训练变好不保证全局最优。