

中文课程名称

课程讲义模板

将封面图命名为 `cover.jpg` 并放入 `figures/` 目录后，会自动显示在这里。

教师： 授课教师
学期： 2026 春季
单位： 学校 / 机构名称
日期： 2026 年 7 月 8 日

目录

第一讲 第一部分	1
第二讲 强化学习基础	2
2.1 强化学习建模方法	2
2.1.1 马尔可夫决策过程	3
2.1.2 关键元素介绍	4
2.1.3 优化目标	5
2.2 强化学习的基本优化算法	7
2.2.1 基于价值函数的方法	7
2.2.2 基于策略优化的方法	7
2.2.3 Actor-Critic 框架	7
2.3 强化学习的进阶优化算法	7
2.3.1 近端策略优化方法	7
2.3.2 离线强化学习方法	7
2.4 强化学习的应用设计方法	7
2.4.1 强化学习的建模步骤	7
2.4.2 状态空间与动作空间设计	7
2.4.3 奖励函数设计	7
2.4.4 强化学习训练过程设计	7
2.5 强化学习的前沿方向	7
2.5.1 多智能体强化学习	7
2.5.2 面向样本重放的强化学习方法	7
2.5.3 强化学习方法的稳定性改进	7
2.6 强化学习方法的挑战	7
2.6.1 面向大规模决策的挑战	7
2.6.2 强化学习训练的不稳定性和泛化性问题	7
2.6.3 强化学习方法的奖励构建问题	7
2.7 本章小结	7
2.8 图片示例	8
2.9 例题与练习	9

第三讲 第二讲标题	10
3.1 知识点一	10
3.1.1 更小的层级	10
3.2 长表格示例	10
附录 A 附录	11
A.1 常用 LaTeX 片段	11
参考资料	12

第一讲 第一部分

第二讲 强化学习基础

强化学习作为机器学习的重要范式之一，主要研究智能体如何在与环境的持续交互中，根据反馈信号不断调整行为策略，从而实现长期收益最大化。与监督学习依赖静态标注样本不同，强化学习更关注动态决策过程中的策略改进问题，因此广泛应用于机器人控制、博弈决策、自动驾驶、推荐系统和交互式智能体等场景。

本章将围绕强化学习的基础理论与方法体系展开，重点介绍其建模框架、基本算法和应用设计方法。首先介绍马尔可夫决策过程、智能体、环境、状态、动作、奖励、策略与优化目标等核心概念；随后梳理基于价值函数的方法、基于策略优化的方法、Actor-Critic 框架，以及近端策略优化和离线强化学习等进阶算法；在此基础上，进一步讨论强化学习在具体任务中的建模步骤，包括状态空间、动作空间、奖励函数和训练过程设计。

值得注意的是，本部分并不以复杂数学推导和理论证明为主要目标，而是侧重于从概念、算法和应用设计三个层面帮助读者理解强化学习的基本思想与方法结构。最后，本部分还将简要介绍多智能体强化学习、样本重放、稳定性改进等前沿方向，以及大规模决策、训练不稳定、泛化能力不足和奖励构建困难等现实挑战，为后续理解生成式人工智能中的强化学习应用奠定基础。

2.1 强化学习建模方法

强化学习的核心在于对“智能体如何在环境中连续决策并根据反馈改进策略”这一问题进行形式化建模。与监督学习中给定输入样本并学习其对应标签不同，如图2.1所示，强化学习所面对的是一个动态交互过程：智能体需要在某一状态下选择动作，环境根据该动作发生状态转移并返回奖励信号，智能体再依据反馈调整后续行为。由此可见，强化学习并不只是解决某一次预测是否准确的问题，而是关注一系列决策行为在长期意义上能否带来更优结果。

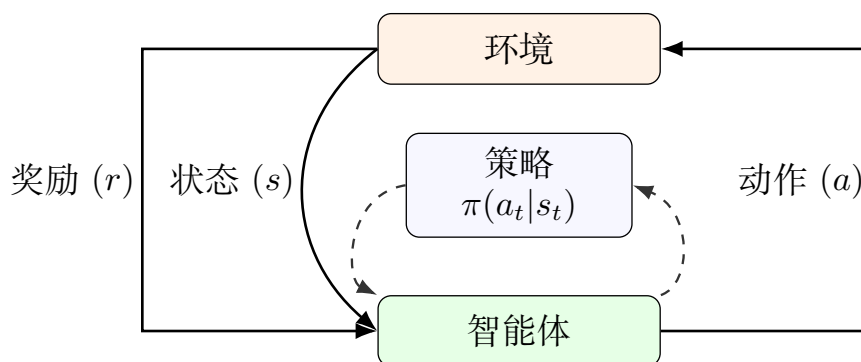


图 2.1: 强化学习的基本建模框架

为了刻画这一交互过程，强化学习通常以马尔可夫决策过程作为基本数学框架。该框架将复杂的智能决策问题抽象为状态、动作、状态转移、奖励函数和策略等基本要素，并通过“当前状态决定未来演化”的马尔可夫性假设，建立起对序列决策问题的统一描述。在这一框架下，智能体的学习目标不再是简单地最大化某一步动作的即时收益，而是通过不断优化策略，使其在长期交互过程中获得尽可能高的累积奖励。

2.1.1 马尔可夫决策过程

马尔可夫决策过程 (Markov Decision Process, MDP) 是强化学习中描述序列决策问题的基本数学框架。强化学习所关注的并不是单次输入与输出之间的静态映射关系，而是智能体在连续交互过程中如何根据环境状态选择动作，并依据环境反馈不断改进策略。为了对这一过程进行形式化刻画，MDP 将智能体与环境之间的交互抽象为状态、动作、状态转移和奖励反馈所构成的动态过程，从而为后续算法设计提供统一的建模基础。

MDP 的核心假设是马尔可夫性，即未来状态的变化只依赖于当前状态和当前动作，而与更早之前的历史状态和动作无关。其形式化表达为：

$$P(s_{t+1} \mid s_t, a_t, s_{t-1}, a_{t-1}, \dots, s_0, a_0) = P(s_{t+1} \mid s_t, a_t) \quad (2.1)$$

其中， s_t 表示智能体在时刻 t 所处的状态， a_t 表示智能体在该状态下执行的动作， s_{t+1} 表示执行动作后环境转移到的下一状态。该假设意味着，只要当前状态 s_t 已经充分包含影响未来决策的信息，那么过去的交互历史就可以被当前状态所概括，智能体只需依据当前状态进行决策。

通常，一个马尔可夫决策过程可以表示为五元组：

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma) \quad (2.2)$$

其中， $\mathcal{S}$ 表示状态空间， $\mathcal{A}$ 表示动作空间， P 表示状态转移概率， r 表示奖励函数， $\gamma \in [0, 1]$ 表示折扣因子。状态转移概率用于刻画智能体在状态 s_t 下执行动作 a_t 后转移到下一状态 s_{t+1} 的可能性：

$$P(s_{t+1} \mid s_t, a_t) \quad (2.3)$$

奖励函数用于表示智能体在特定状态下执行动作并产生状态转移后所获得的反馈，通常可写为：

$$r_t = r(s_t, a_t, s_{t+1}) \quad (2.4)$$

在确定性环境中，如果下一状态 s_{t+1} 可以由当前状态 s_t 和动作 a_t 唯一确定，则奖励函数也可简化表示为：

$$r_t = r(s_t, a_t) \quad (2.5)$$

在 MDP 框架中，智能体与环境的交互会形成一条轨迹：

$$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \dots, s_T) \quad (2.6)$$

其中， T 表示交互过程的终止时刻。强化学习算法通常通过采样或估计大量类似轨迹，分析不同策略所能获得的长期收益，并据此优化智能体的决策方式。

2.1.2 关键元素介绍

在马尔可夫决策过程的形式化框架下，强化学习问题通常由智能体、环境、状态、动作、奖励、策略和轨迹等关键元素共同构成。这些元素分别对应决策过程中的不同环节：智能体负责作出决策，环境提供交互对象与反馈，状态刻画当前决策条件，动作表示智能体的行为选择，奖励反映动作结果的优劣，策略决定智能体如何依据状态选择动作，而轨迹则记录了智能体与环境连续交互的完整过程。理解这些元素之间的关系，是掌握强化学习建模方法的基础。

- **智能体 (Agent)**。智能体是强化学习中的决策主体，负责感知环境状态、选择动作，并在不断交互中调整自身策略。智能体可以是游戏中的棋手、机器人控制系统、自动驾驶车辆，也可以是推荐系统中的排序模型。其核心特征在于能够根据环境反馈不断改进行为，而不是仅执行预先设定的固定规则。
- **环境 (Environment)**。环境是智能体进行交互的外部对象。智能体执行动作后，环境会发生相应变化，并向智能体返回新的状态和奖励信号。例如，在五子棋任务中，棋盘局面和对手可以共同构成环境；在机器人控制任务中，物理空间、障碍物和目标位置构成环境。环境决定了智能体动作产生的后果。
- **状态 (State)**。状态表示智能体在某一时刻能够观察或感知到的环境信息，通常记为 s_t 。状态是智能体进行动作选择的依据。状态表示是否充分，直接影响强化学习模型的效果。如果状态信息不足，智能体可能无法判断当前处境；如果状态表示过于复杂，则会增加学习难度。以五子棋为例，当前棋盘上黑白棋子的分布就可以作为状态表示。
- **动作 (Action)**。动作表示智能体在当前状态下可以执行的行为，通常记为 a_t 。动作空间可以是离散的，也可以是连续的。五子棋中的动作是“在某一空位落子”，属于离散动作；机器人控制中的动作可能是关节角度、速度或力矩变化，通常属于连续动作。动作空间的规模和结构会显著影响强化学习问题的难度。
- **奖励 (Reward)**。奖励是环境对智能体动作结果的反馈信号，通常记为 r_t 。奖励可以用于表示动作带来的即时收益，也可以反映任务完成情况。奖励函数的设计是强化学习建模中的关键环节，因为它直接决定智能体学习的目标方向。例如，在五子棋中，胜利可以获得正奖励，失败可以获得负奖励；在路径规划中，到达目标可以获得奖励，而碰撞障碍物则会受到惩罚。

- **策略 (Policy)**。策略表示智能体在给定状态下选择动作的规则，通常记为 π 。在随机策略中，策略可表示为条件概率分布：

$$\pi(a_t | s_t) \quad (2.7)$$

即智能体在状态 s_t 下选择动作 a_t 的概率。如果策略是确定性的，则每个状态对应一个固定动作；如果策略是随机性的，则策略给出不同动作被选择的概率。强化学习的核心目标就是不断优化策略，使智能体在长期交互过程中获得更高回报。

- **轨迹 (Trajectory)**。轨迹是智能体与环境连续交互形成的状态、动作和奖励序列，通常表示为：

$$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \dots, s_T) \quad (2.8)$$

轨迹记录了智能体从初始状态到终止状态的完整决策过程。强化学习算法通常通过采样大量轨迹，估计不同策略的表现，并利用这些经验改进策略。

上述关键元素共同构成了强化学习问题的基本结构。智能体根据状态选择动作，环境根据动作产生状态转移并返回奖励，智能体再依据奖励反馈调整策略。强化学习的学习过程正是在这一循环中不断展开的。

2.1.3 优化目标

强化学习的优化目标是学习一个最优策略，使智能体在与环境的持续交互过程中获得尽可能高的长期累积奖励。与只关注单一步骤预测结果的学习方法不同，强化学习并不以最大化某一时刻的即时奖励 r_t 为最终目标，而是关注一系列连续动作在整体上所带来的长期收益。因此，强化学习中的策略优化本质上是一种面向序列决策过程的长期收益最大化问题。

设智能体与环境交互形成的状态-动作序列为：

$$\tau = \{(s_1, a_1), (s_2, a_2), \dots, (s_T, a_T)\} \quad (2.9)$$

其中， T 表示交互序列的总步数， τ 表示在某一策略下生成的一条完整交互轨迹。在不考虑折扣因子的简化情况下，该轨迹上的长期累积奖励可表示为：

$$R(\tau) = \sum_{t=1}^T r_t \quad (2.10)$$

其中， $R(\tau)$ 表示轨迹 τ 所对应的总累积奖励， r_t 表示智能体在时刻 t 获得的即时奖励。

在一般情况下，未来奖励的重要性通常会随着时间推移而降低，因此也可以引入

折扣因子 $\gamma \in [0, 1]$ ，将长期累积奖励表示为折扣回报：

$$R(\tau) = \sum_{t=1}^T \gamma^{t-1} r_t \quad (2.11)$$

其中， γ 用于调节未来奖励在当前决策中的重要程度。当 γ 趋近于 1 时，智能体更加重视长期收益；当 γ 较小时，智能体更加关注近期奖励。

由于状态-动作序列 τ 的产生依赖于智能体所采用的策略，不同策略会诱导出不同的轨迹分布。设智能体的策略为 $\pi_\theta(\cdot)$ ，其中 θ 表示策略参数，则在策略 π_θ 下，轨迹 τ 出现的概率可记为 $\text{Pr}_\theta(\tau)$ 。因此，需要考虑所有可能轨迹上的累积奖励期望，由此定义性能函数：

$$J(\theta) = \mathbb{E}_{\tau \sim \text{Pr}_\theta(\tau)}[R(\tau)] \quad (2.12)$$

进一步展开可得：

$$J(\theta) = \sum_{\tau \in \mathcal{D}_\tau} \text{Pr}_\theta(\tau) R(\tau) = \sum_{\tau \in \mathcal{D}_\tau} \text{Pr}_\theta(\tau) \sum_{t=1}^T \gamma^{t-1} r_t \quad (2.13)$$

其中， $\mathcal{D}_\tau$ 表示所有可能轨迹构成的轨迹空间， $\text{Pr}_\theta(\tau)$ 表示在策略 π_θ 下生成轨迹 τ 的概率， $J(\theta)$ 则用于衡量当前策略 $\pi_\theta(\cdot)$ 在长期交互过程中的整体表现。

由此，强化学习的训练目标可以表示为求解使性能函数最大化的最优策略参数：

$$\theta^* = \arg \max_{\theta} J(\theta) \quad (2.14)$$

相应地，最优策略可表示为：

$$\pi_{\theta^*} = \arg \max_{\pi_\theta} J(\theta) \quad (2.15)$$

这一优化目标说明，强化学习并不是简单追求某一步动作的即时收益最大化。某些动作虽然能够带来较高的短期奖励，却可能导致后续状态恶化，从而降低整个决策序列的长期收益；相反，某些动作在当前时刻的奖励可能较低，但有助于智能体在未来获得更高回报。因此，强化学习的关键在于使智能体学会在即时收益与未来收益之间进行权衡，并通过不断优化策略参数 θ ，最大化性能函数 $J(\theta)$ ，最终实现长期累积奖励的最大化。

2.2 强化学习的基本优化算法

2.2.1 基于价值函数的方法

2.2.2 基于策略优化的方法

2.2.3 Actor-Critic 框架

2.3 强化学习的进阶优化算法

2.3.1 近端策略优化方法

2.3.2 离线强化学习方法

2.4 强化学习的应用设计方法

2.4.1 强化学习的建模步骤

2.4.2 状态空间与动作空间设计

2.4.3 奖励函数设计

2.4.4 强化学习训练过程设计

2.5 强化学习的前沿方向

2.5.1 多智能体强化学习

2.5.2 面向样本重放的强化学习方法

2.5.3 强化学习方法的稳定性改进

2.6 强化学习方法的挑战

2.6.1 面向大规模决策的挑战

2.6.2 强化学习训练的不稳定性和泛化性问题

2.6.3 强化学习方法的奖励构建问题

2.7 本章小结

学习目标

- 了解本课程的学习目标、评价方式与课堂要求。
- 熟悉讲义中的例题、练习、阅读材料与作业模块。
- 掌握如何在模板中插入章节、表格、图片和重点提示。

这里填写本讲的正文内容。中文可直接输入，不需要额外转码。若要突出关键词，可使用 **关键词强调**，也可以使用普通的 **加粗**、斜体等格式。

重点提示

这里可以写课堂重点、易错点、考试提示，或者需要学生特别记住的概念。

表格建议使用 booktabs 提供的 `\toprule`、`\midrule`、`\bottomrule`，视觉会比普通竖线表格更清爽。

表 2.1: 课程安排示例

周次	主题	任务
第 1 周	课程导论与学习方法	阅读讲义
第 2 周	核心概念一	完成练习
第 3 周	核心概念二	小组讨论

2.8 图片示例

将图片放入 `figures/` 目录，然后用 `\includegraphics` 插入。下面的示例会在 `figures/example.png` 存在时显示图片，否则显示一个占位框。

图片占位：将图片保存为 `figures/example.png`

图 2.2: 图片插入示例

2.9 例题与练习

例题 1

这里放置例题内容。可以包含公式：

$$a^2 + b^2 = c^2$$

也可以写解题步骤或分析。

课堂练习

1. 请用自己的话概括本讲的三个重点。
2. 根据表 2.1，说明第 2 周需要完成什么任务。
3. 观察图 2.2，写出你能得到的信息。

第三讲 第二讲标题

3.1 知识点一

从这里开始写第二讲内容。新增章节时，复制下面结构即可：

```
\chapter{新的一讲}
\section{小节标题}
正文内容……
```

3.1.1 更小的层级

如果一讲内部内容较多，可以继续使用 `\subsection` 或 `\subsubsection` 组织层次。

3.2 长表格示例

如果表格跨页，可以使用 `longtable`。

表 3.1: 长表格结构示例

模块	内容	备注
课前	预习阅读材料	可附链接或二维码图片
课中	讲授、讨论、练习	建议加入例题与即时反馈
课后	作业、复盘、扩展阅读	可列出提交要求

附录 A 附录

A.1 常用 LaTeX 片段

% 插入普通图片

```
\begin{figure}[htbp]
  \centering
  \includegraphics[width=0.8\linewidth]{figures/your-image.png}
  \caption{图片标题}
\end{figure}
```

% 插入三线表

```
\begin{table}[htbp]
  \centering
  \caption{表格标题}
  \begin{tabular}{lll}
    \toprule
    A & B & C \\
    \midrule
    内容 & 内容 & 内容 \\
    \bottomrule
  \end{tabular}
\end{table}
```

参考资料

这里可以列出教材、论文、网站或其他阅读材料。

参考文献

- [1] 作者. 书名. 出版社, 年份.
- [2] 网站名称. 文章标题. <https://example.com>