强化学习笔记

概念

学习本质上源于与环境的互动，互动是知识的主要来源。接下来我们将研究强化学习——专注于**目标导向的、基于互动的学习**，区别于其他机器学习范式，强调智能体通过试错与环境交互来优化行为策略。

强化学习的目标是学习在不同情境下应采取何种动作，以**最大化累积数值奖励**。通过**试错学习**(不被告知正确动作，需通过尝试发现最优策略)以及**延迟奖励**(动作影响当前及未来情境与奖励，需考虑长期后果)来实现。

范式	学习方式	是否有标签	目标
监督学习	从带标签样本中学习	是（正确动作已知）	泛化到新情境
无监督学习	发现数据中的结构	否	聚类、降维等
强化学习	从试错与奖励中学习	否（仅有奖励信号）	**最大化长期奖励**

强化学习的核心挑战是**探索与利用之间**的权衡。利用指的是选择已知高奖励动作。**探索**指的是尝试新动作以发现潜在更优策略。**困境**在于，不能只追求探索或只追求利用，否则都会导致任务失败。过度利用则会错失更好策略；过度探索则会奖励低下。

强化学习从**完整、交互、目标导向的智能体**出发，强调：感知环境状态、采取影响环境的动作、在不确定性中追求目标。传统的监督学习不考虑动作与反馈循环，规划方法则忽略实时决策与模型获取；强化学习整合学习、规划、模型构建，解决**如何嵌入更大决策系统**的问题。

例子

例子	环境	动作的影响	目标	不确定性体现	如何通过经验改进
国际象棋大师走棋	棋局（对手、棋子位置）	影响下一个棋局状态和后续可走的棋	赢得比赛	对手回应不可完全预测	完善对棋局和走法可取性的直觉判断，提升棋艺
炼油厂自适应控制器	炼油厂（储罐液位、温度、压力等）	影响产量、成本、质量的实时状态	优化产量/成本/质量权衡（按边际成本）	实际运行条件偏离初始设定，需动态调整	通过运行经验优化控制策略
新生小羚羊学习奔跑	地形、自身身体状态	影响移动能力与生存机会	站立并奔跑（避免被捕食）	初次站立和行走的结果不确定	提高奔跑效率，逐步掌握运动技能

例子	环境	动作的影响	目标	不确定性体现	如何通过经验改进
移动机器人决策	房间布局、电池状态、垃圾分布	决定是否继续探索或返回充电，影响未来电量与任务能力	寻找垃圾并保持电力充足（避免耗尽）	充电站位置、路径难度、探索收益不确定	根据过去寻找充电器的速度和难易程度调整决策
准备早餐	厨房环境、身体状态（饥饿、偏好）	动作序列影响任务效率与结果（如溢出牛奶）	成功准备并享用早餐（获得营养）	物品位置、牛奶倒出速度、身体状态变化等不确定	学会简化流程，优化物品获取顺序和行为策略

共性

共同特征	说明
互动性	所有例子中，智能体通过动作影响环境，并根据环境反馈调整行为。
目标导向	每个任务都有明确目标，智能体可通过直接感知判断进展（如是否赢棋、是否吃饱、电池是否耗尽）。
延迟后果	当前动作影响未来状态和机会，正确决策需考虑间接和长期影响。
不确定性	动作结果无法完全预测，必须持续感知并响应（如菲尔观察牛奶是否溢出）。
经验驱动改进	智能体通过时间积累的经验提升性能，行为随学习而优化。
先验知识的作用	初始知识（来自经验、设计或进化）影响学习起点，但**与环境的互动**对适应具体任务至关重要。

要素

强化学习系统包含以下四个主要子元素：**策略**、**奖励信号**、**价值函数**，以及可选的**环境模型**。除智能体和环境外，这些是构成强化学习系统的核心组成部分。

策略

策略是学习智能体在给定时间的行为方式，用于将感知到的环境状态映射到应采取的动作。策略是智能体的核心，**仅凭策略就足以确定行为**；其可以是随机的，为每个动作指定选择概率。

奖励信号

在每个时间步，环境向智能体发送一个单一数字，称为**奖励**。奖励定义强化学习问题的**目标**，智能体的唯一目标是**最大化其长期接收到的总奖励**。奖励信号决定对智能体而言什么是“好”或“坏”的事件，是问题的**即时且决定性特征**，类比于生物系统中的愉悦或痛苦体验。奖励是**改变策略的主要依据**；若某动作后获得低奖励，策略可能被调整，以在未来类似状态下选择其他动作。

价值函数

价值函数用于评估**长期收益。**一个状态的**价值**，是智能体从该状态出发，在未来可以期望累积的**总奖励量。奖励**表示状态的**即时可取性**；与其相比**价值**表示状态的**长期可取性**，考虑后续可能的状态及其中的奖励。但**奖励是首要的**，价值是作为对奖励的预测而存在的，是**次要的。**没有奖励，就不可能有价值；估计价值的**唯一目的**是为了获得更多奖励。但虽然奖励是根本，但**决策依赖于价值判断。**智能体选择能导向**高价值状态**的动作，而非仅追求高即时奖励的状态，以实现长期奖励最大化。但确定价值比确定奖励困难得多。奖励由环境直接给出；价值则必须根据智能体在其整个生命周期中观察到的经验序列进行**估计和重新估计。几乎所有强化学习算法的核心组成部分，都是高效估计价值的方法。**

环境模型

环境模型是用于模仿环境行为的组件，或允许对环境未来行为进行推断的机制。基于给定当前状态和动作，模型可预测下一个状态和下一个奖励。其住哟啊用于支持**规划**；在实际经历之前，通过考虑可能的未来情境来决定行动方案。

局限性

状态作为**策略**和**价值函数**的输入；作为**环境模型**的输入和输出。强化学习的局限性在于严重依赖于**状态**（state）的概念。

第一部分：表格型解法

在强化学习中，**表格型解法**（Tabular Solution Methods）是指在状态空间和动作空间较小且有限的情况下，将价值函数或策略以**表格形式**显式存储和更新的求解方法。这类方法将每个状态（或状态-动作对）作为表中的一个条目，直接记录其对应的价值估计或策略概率。由于所有状态和动作都可以被完整枚举并存储，这些方法能够通过对环境的交互或建模，逐步计算出精确的价值函数和最优策略。