马尔可夫决策过程

本章开始包括后面的章节会涉及理论公式推导，建议读者在阅读之前先回顾一下概率论相关知识，尤其是条件概率、全概率期望公式等等。

马尔可夫决策过程

马尔可夫决策过程（Markov decision process，MDP）是强化学习的基本问题模型之一，它能够以数学的形式来描述智能体在与环境交互的过程中学到一个目标的过程。这里智能体充当的是作出决策或动作，并且在交互过程中学习的角色，环境指的是智能体与之交互的一切外在事物，不包括智能体本身。举个例子，比如我们要学习弹钢琴，在这个过程中充当决策者和学习者的我们人本身就是智能体，而我们的交互主体即钢琴就是环境。当我们执行动作也就是弹的时候会观测到一些信息，例如琴键的位置等，这就是状态。此外当我们弹下去的时候会收到钢琴发出的声音，也就是反馈，我们通过钢琴发出的声音来判断自己弹得好不好然后反思并纠正下一次弹的动作。当然如果这时候有一个钢琴教师在旁边指导我们，那样其实钢琴和教师就同时组成了环境，我们也可以交互过程中接收教师的反馈来提高自己的钢琴水平。

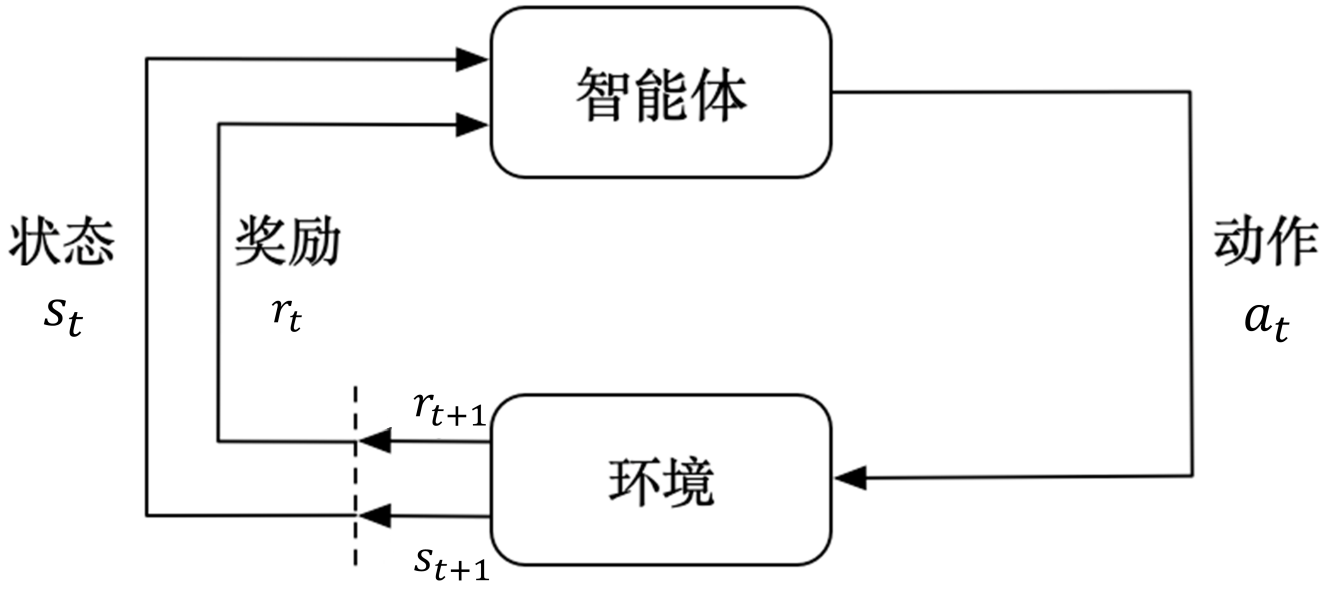

图 2.1 马尔可夫决策过程中智能体与环境的交互过程

如图 2.1 所示，它描述了马尔可夫决策过程中智能体与环境的交互过程。确切地说，智能体与环境之间是在一系列**离散的时步**[2]（time step）交互的，一般用 t 来表示，$t=0,1,2,...$[3]。在每个时步 t， 智能体会观测或者接收到当前环境的状态 s_t，根据这个状态 s_t 执行动作 a_t。执行完动作之后会收到一个奖励 $r{t+1}$[1]，*同时环境也会收到动作 a_t 的影响会变成新的状态 $s*{t+1}$*，并且在 $t+1$时步被智能体观测到。如此循环下去，我们就可以在这个交互过程中得到一串轨迹，可表示为：

$$s_0, a_0, r_1, s_1, a_1, r_2, \ldots, s_t, a_t, r_{t+1}, \ldots \tag{1}$$

其中奖励 r_{t+1}就相当于我们学习弹钢琴时收到的反馈，我们弹对了会收到老师的表扬，相当于一个正奖励，弹错了可能会接受老师的批评，相当于一个负奖励。前面讲到马尔可夫决策过程描述的是智能体在交互过程中学到一个目标的过程，而**这个目标通常是以最大化累积的奖励来呈现的**。换句话说，我们的目标是使得在交互过程中得到的奖励之和$r_1+r_2+...+r_T$尽可能最大，其中T表示当前交互过程中的最后一个时步，也就是最大步数，从$t=0$和$t+T$这一段时步我们称为一个回合（episode），比如游戏中的一局。

马尔可夫性质

现在我们介绍马尔可夫决策过程的一个前提，即马尔可夫性质，用公式表示如下：

$$P(s_{t+1}|s_t) = P(s_{t+1}|s_0, s_1, \ldots, s_t) \tag{2}$$

这个公式的意思就是在给定历史状态$s_0, s_1,...,s_t$的情况下，某个状态的未来只与当前状态s_t有关，与历史的状态无关。这个性质其实对于很多问题有着非常重要的指导意义的，因为这允许我们在没有考虑系统完整历史的情况下预测和控制其行为，随着我们对强化学习的深入会越来越明白这个性质的重要性。实际问题中，有很多例子其实是不符合马尔可夫性质的，比如我们所熟知的棋类游戏，因为在我们决策的过程中不仅需要考虑到当前棋子的位置和对手的情况，还需要考虑历史走子的位置例如吃子等。换句话说，它们不仅取决于当前状态，还依赖于历史状态。当然这也并不意味着完全不能用强化学习来解决，实际上此时我们可以用深度学习神经网络来表示当前的棋局，并用蒙特卡洛搜索树等技术来模拟玩家的策略和未来可能的状态，来构建一个新的决策模型，这就是著名的 AlphaGO 算法[@alphago]。具体的技术细节后面会展开，总之记住在具体的情境下，当我们要解决问题不能严格满足马尔可夫性质的条件时，是可以结合其他的方法来辅助强化学习进行决策的。

回报

前面讲到在马尔可夫决策过程中智能体的目标时最大化累积的奖励，通常我们把这个累积的奖励称为回报（Return），用G_t表示，最简单的回报公式可以写成：

$$G_t = r_1 + r_2 + \ldots + r_T \tag{3}$$

其中 T 前面提到过了，表示最后一个时步，也就是每回合的最大步数。这个公式其实只适用于有限步数的情境，例如玩一局游戏，无论输赢每回合总是会在有限的步数内会以一个特殊的状态结束，这样的状态称之为终止状态。但也有一些情况是没有终止状态的，换句话说智能体会持续与环境交互，比如人造卫星在发射出去后会一直在外太空作业直到报废或者被回收，这样的任务称之为持续性任务。在持续性任务中上面的回报公式是有问题的，因为此时 $T=\infty$。

为了解决这个问题，我们引入一个折扣因子（discount factor）γ，并可以将回报表示为：

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{4}$$

其中γ取值范围在0到1之间，它表示了我们在考虑未来奖励时的重要程度，控制着当前奖励和未来奖励之间的权衡。换句话说，它体现了我们对长远目标的关注度。当$\gamma=0$时，我们只会关心当前的奖励，而不会关心将来的任何奖励。而当γ接近1时，我们会对所有未来奖励都给予较高的关注度。

这样做的好处是会让当前时步的回报G_t跟下一个时步G_{t+1}的回报是有所关联的，即：

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \\ &= R_{t+1} + \gamma \left(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots \right) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned} \tag{5}$$

这个公式对于所有$t<T$都是存在的，在后面我们将贝尔曼公式的时候会明白它的重要性。

状态转移矩阵

截至目前，我们讲的都是有限状态马尔可夫决策过程（finite MDP），这指的是状态的数量必须是有限的（无论是离散的还是连续的。如果状态数是无限的，通常会使用另一种方式来对问题建模，称为泊松（Poisson）过程。这个过程又被称为连续时间马尔可夫过程，它允许发生无限次事件，每个事件发生的机会相对较小，但当时间趋近于无穷大时，这些事件以极快的速度发生。尽管泊松过程在某些应用领域中非常有用，但是对于大多数强化学习场景，还是用的有限状态马尔可夫决策过程。

既然状态数是有限的，那其实我们可以用一种状态流向图来表示智能体与环境交互过程中的走向。举个例子，假设学生正在上课，一般来讲从老师的角度来说学生会有三种状态，认真听讲、玩手机和睡觉，分别用s_1，s_2和s_3表示。注意，这里从老师的角度的意思是把老师当作智能体，学生跟教师组成环境，而如果把学生当作智能体，那么认真听讲、玩手机这些就只能理解成智能体作出的决策或动作，而不是状态了。这在实际问题中是很常见的，毕竟交互是相互的，强化学习中的环境也不是严格意义上的景止环境，它也可以是其他智能体。**有时智能体和环境的角色是能相互对调的，只要能各自建模成马尔可夫决策过程即可**，比如在竞技游戏中，我方角色可以当对方角色看作环境的一部分，对方角色也可以把我方角色看作环境的一部分，然后各自作出相应的决策。 回到我们举的例子，在马尔可夫决策过程中一般所有状态之间都是可以相互切换的，当学生在认真听讲时能切换到玩手机或者睡觉的状态，在睡觉时也可能继续睡觉，也可能醒过来认真听讲或者玩手机。

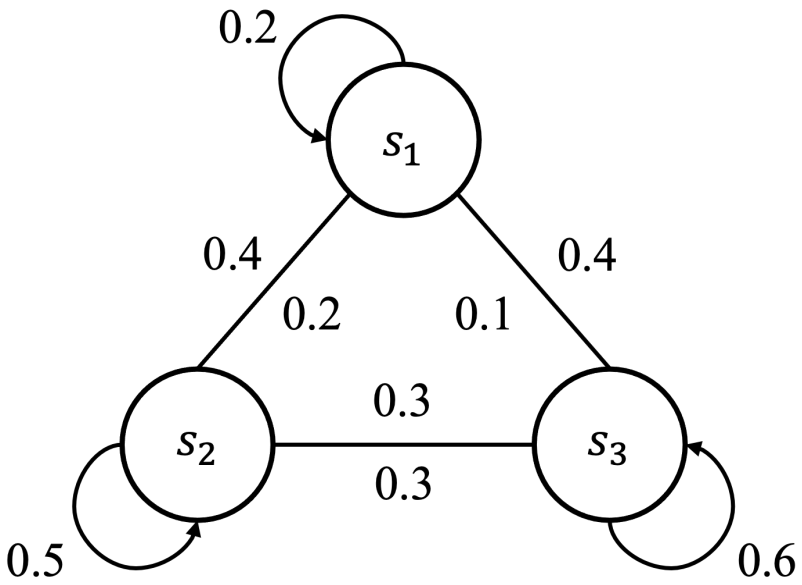

图 2.2 马尔可夫链

如图 2.2 所示，图中每个曲线箭头表示指向自己，比如当学生在认真听讲即处于状态 s_1 时，会有 0.2 的概率继续认真听讲。当然也会分别有 0.4 和 0.4 的概率玩手机（s_2）或者睡觉（s_3）。此外，当学生处于状态 s_2 时，也会有 0.2 的概率会到认真听讲的状态（s_1），像这种两个状态之间能互相切换的情况我们用一条没有箭头的线连接起来，参考无向图的表示。

整张图表示了马尔可夫决策过程中的状态流向，这其实跟数字电路中**有限状态机**的概念是一样的。严格意义上来讲，这张图中并没有完整地描述出马尔可夫决策过程，因为没有包涵动作、奖励等元素，所以一般我们称之为**马尔可夫链**（Markov Chain），又叫做离散时间的马尔可夫过程（Markov Process），跟马尔可夫决策过程一样，都需要满足马尔可夫性质。因此我们可以用一个概率来表示状态之间的切换，比如 $ P*{12}=P(S*{t+1}=s_2 | S_{t+1}=s_1) = 0.4$ 表示当前时步的状态是 s_1 即认真听讲时在下一个时步切换到 s_2 即玩手机的概率，我们把这个概率称为**状态转移概率（State Transition Probability）**。拓展到所有状态我们可以表示为：

$$P_{ss'} = P(S_{t+1} = s' | S_t = s) \tag{6}$$

即当前状态是 s 时，下一个状态是 s'的概率，其中大写的 S 表示所有状态的集合，即 $S={s_1,s_2,s_3}$ 。

由于状态数是有限的，我们可以把这些概率绘制成表格的形式，如下：

	$S_{t+1} = s_1$	**$S_{t+1} = s_2$**	**$S_{t+1} = s_3$**
$S_t = s_1$	0.2	0.4	0.4
$S_t = s_2$	0.2	0.5	0.3
$S_t = s_3$	0.1	0.3	0.6

在数学上也可以用矩阵来表示，如下：

$$P_{ss'} = \begin{bmatrix} 0.2 & 0.4 & 0.4 \\ 0.2 & 0.5 & 0.3 \\ 0.1 & 0.3 & 0.6 \end{bmatrix}$$

这个矩阵就叫做**状态转移矩阵（State Transition Matrix）**，拓展到所有状态可表示为：

$$P_{ss'} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} \tag{7}$$

其中 n 表示状态数，注意对于同一个状态所有状态转移概率加起来是等于 1 的，比如对于状态 s_1 来说，$p*{11}+ p*{12}+ \cdots+p_{1n}=1$。还有一个**非常重要的点就是，状态转移矩阵是环境的一部分**，跟智能体是没什么关系的，而智能体会根据状态转移矩阵来做出决策。在这个例子中老师是智能体，学生的状态不管是认真听讲、玩手机还是睡觉这些老师是无法决定的，老师只能根据学生的状态做出决策，比如看见学生玩手机就提醒一下上课认真听讲等等。

此外，在马尔可夫链（马尔可夫过程）的基础上增加奖励元素就会形成**马尔可夫奖励过程（Markov reward process, MRP）**，在马尔可夫奖励过程基础上增加动作的元素就会形成马尔可夫决策过程，也就是强化学习的基本问题模型之一。其中马尔可夫链和马尔可夫奖励过程在其他领域例如金融分析会用的比较多，强化学习则重在决策，这里讲马尔可夫链的例子也是为了帮助读者理解状态转移矩阵的概念。

本章小结

本章主要介绍了马尔可夫决策过程的概念，它是强化学习的基本问题模型之一，因此读者需要牢牢掌握。此外拓展了一些重要的概念，包括马尔可夫性质、回报、状态转移矩阵、轨迹、回合等，这些概念在我们后面讲解强化学习算法的时候会频繁用到，务必牢记。

---

1. 这里奖励表示成$r*{t+1}$*而不是*r_{t}，是因为此时的奖励是由于动作a_t和状态s_t来决定的，也就是执行完动作之后才能收到奖励，因此更强调是下一个时步的奖励。 ↩

2. 有些方法可以拓展到连续时间的情况，但为了方便，我们尽量只考虑离散时步的情况 ↩

3. 注意，这里的$t=0$和$t=1$之间是跟现实时间无关的，取决于智能体每次交互并获得反馈所需要的时间，比如在弹钢琴的例子中我们是能够实时接收到反馈的，但是比如我们的目标是考试拿高分的时候，每次考完试我们一般是不能立刻接收到反馈即获得考试分数的，这种情况下$t=0$和$t=1$之间会显得特别漫长。 ↩