强化学习原理

第三章读书笔记

姓名：石若川 学号：2111381 专业：智能科学与技术

3 有限马尔可夫决策过程

马尔可夫决策过程（MDP）是序列决策的经典形式化表达，动作不仅影响当前的立即收益，还会影响后续状态和收益。在赌博机问题中，我们估计了每个动作 a 的价值 $q_*(a)$；而在 MDP 中，我们估计每个动作 a 在每个状态 s 中的价值 $q_*(s,a)$，或者估计给定最优动作下的每个状态的价值 $v_*(s)$。

3.1 “智能体-环境”交互接口

1. MDP 中的“智能体-环境”交互过程

MDP 是一种通过交互式学习来实现目标的理论框架。进行学习及实施决策的机器被称为智能体 (agent)。智能体之外所有与其相互作用的事物都被称为环境 (enuironment)。这些事物之间持续进行交互，智能体选择动作，环境对这些动作做出相应的响应，并向智能体呈现出新的状态。环境也会产生一个收益，通常是特定的数值，这就是智能体在动作选择过程中想要最大化的目标，如图3.1所示。

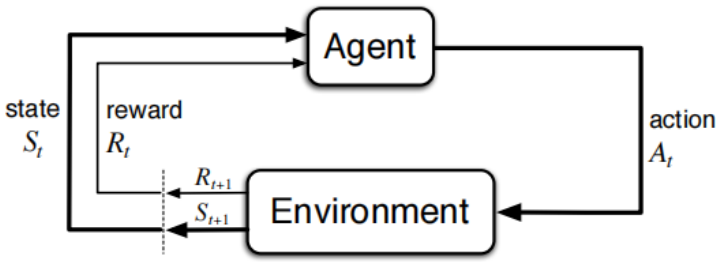

图 3.1: 智能体-环境的交互

2. 轨迹

在每个离散时刻 $t = 0, 1, 2, 3, \ldots$, 智能体和环境都发生了交互。在每个时刻 t，智能体观察到所在的环境状态的某种特征表达，$S_t \in \mathcal{S}$，并且在此基础上选择一个动作，$A_t \in \mathcal{A}(s)$。下一时刻，作为其动作的结果，智能体接收到一个数值化的收益，$R_{t+1} \in \mathcal{R} \subset \mathbb{R},$，并进入一个新的状态 S_{t+1}。从而，MDP 和智能体共同给出了一个序列或轨迹

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots$$

3. MDP 的动态特性

在有限 MDP 中，状态、动作和收益的集合 ($\mathcal{S}$、$\mathcal{A}$ 和 $\mathcal{R}$) 都只有有限个元素。在这种情况下，随机变量 R_t 和 S_t, 具有定义明确的离散概率分布，并且只依赖于前继状态和动作。也就是说，给定前继状态和动作的值时，这些随机变量的特定值，$s' \in \mathcal{S}$ 和 $r \in \mathcal{R}$ 在 t 时刻出现的概率是

$$p(s', r|s, a) \doteq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\} \tag{1}$$

4. 马尔可夫性

在马尔可夫决策过程中，由 p 给出的概率完全刻画了环境的动态特性。也就是说，S_t 和 R_t 的每个可能的值出现的概率只取决于前一个状态 S_{t-1} 和前一个动作 A_{t-1}，并且与更早之前的状态和动作完全无关。这样，状态就被认为是具有马尔可夫性的。

5. 环境信息

- 状态转移概率 $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$

$$p(s'|s, a) \doteq \Pr\{S_t = s'|S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in \mathcal{R}} p(s', r|s, a) \tag{2}$$

- “状态-动作”二元组的期望收益 $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

$$r(s, a) \doteq \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a] = \sum_{r \in \mathbb{R}} r \sum_{s' \in \mathcal{S}} p(s', r|s, a) \tag{3}$$

- “状态-动作-后继状态”三元组的期望收益 $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

$$r(s, a, s') \doteq \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in \mathcal{R}} r \frac{p(s', r|s, a)}{p(s'|s, a)} \tag{4}$$

6. 环境与智能体

通常来说，智能体不能改变的事物都被认为是在外部的，即是环境的一部分。但我们并不是假定智能体对环境一无所知。例如，智能体通常会知道如何通过一个动作和状态的函数来计算所得到的收益。但是，我们通常认为收益的计算在智能体的外部，因为它定义了智能体所面临的任务，因此智能体必然无法随意改变它。智能体和环境的界限划分仅仅决定了智能体进行绝对控制的边界，而并不是其知识的边界。

MDP 框架是目标导向的交互式学习问题的一个高度抽象。任何目标导向的行为的学习问题都可以概括为智能体及其环境之间来回传递的三个信号: 一个信号用来表示智能体做出的选择 (行动)，一个信号用来表示做出该选择的基础 (状态)，还有一个信号用来定义智能体的目标 (收益)。实验表明，对于不同的任务，特定的状态和动作的定义差异很大，并且其性能极易受其表征方式的影响。

3.2 目标和收益

在强化学习中，智能体的目标被形式化表征为一种特殊信号，称为收益，它通过环境传递给智能体。在每个时刻，收益都是一个单一标量数值，$R_t \in \mathbb{R}$。智能体的目标是最大化其收到的总收益。使用收益信号来形式化目标是强化学习最显著的特征之一。

虽然基于收益信号来形式化目标的做法在一开始可能存在限制，但在实际中却显示出灵活性和广泛的可行性。例如，为了使机器人学习走路，研究人员在每个时刻都提供了与机器人向前运动成比例的收益。在训练机器人学习 (如逃脱迷宫) 的过程中，成功逃脱前每个时刻的收益都是-1，这会鼓励智能体尽快逃脱。

如果我们想要它为我们做某件事，我们提供收益的方式必须要使得智能体在最大化收益的同时也实现我们的目标。因此，我们设立收益的方式要能真正表明我们的目标。特别地，收益信号并不是传授智能体如何实现目标的先验知识。例如，国际象棋智能体只有当最终获胜时才能获得收益，而并非达到某个子目标，比如吃掉对方的子或者控制中心区域。如果实现这些子目标也能得到收益，那么智

能体可能会找到某种即使绕开最终目的也能实现这些子目标的方式。例如，它可能会找到一种以输掉比赛为代价的方式来吃对方的子。收益信号只能用来传达什么是你想要实现的目标，而不是如何实现这个目标。

3.3 回报与分幕

1. 累计回报与幕

在最简单的情况下，回报是收益的总和

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T \tag{5}$$

其中 T 是最终时刻。

智能体和环境的交互能被自然地分成一系列子序列，我们称每个子序列为幕。每幕都以一种特殊状态结束，称之为终结状态。随后会重新从某个标准的起始状态或起始状态的分布中某个状态样本开始。即使结束的方式不同，下一幕的开始状态与上一幕的结束方式完全无关。因此这些幕可以被认为在同样的终结状态下结束，只是对不同的结果有不同的收益。具有这种分幕重复特性的任务称为分幕式任务。智能体-环境交互持续不断发生时，这类任务称为持续性任务。

2. 折扣累计回报

回报公式5描述持续性任务时，最大化回报很容易趋于无穷。因此，引入折扣的概念。智能体选择 A_t 来最大化期望折扣累计回报

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{6}$$

其中，γ 是一个参数，$0 \le \gamma \le 1$，称为折扣率。如果 $\gamma = 0$，那么智能体就是“目光短浅的”，只关心最大化当前利益。随着 γ 接近 1，智能体变得有远见了，将更多地考虑将来的利益。

相邻时刻的回报可以写为递归形式

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

公式6中只要收益是非零常数且 $\gamma < 1$，则回报是有限的。比如，若收益是 $+1$，那么回报就是

$$G_t = \sum_{k=0}^{\infty} = \frac{1}{1-\gamma}$$

3.4 分幕式和持续性任务的统一表示法

对于回报的定义包括两种情况：一是将回报定义为有限项的总和，如式5所示；而在另一种情况中，我们将回报定义为无限项的总和，如式6所示。这两者可以通过一个方法进行统一，即把幕的终止当作一个特殊的吸收状态的入口，它只会转移到自己并且只产生零收益。例如，考虑状态转移图

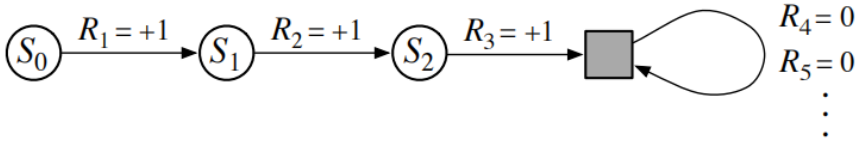

图 3.2: 状态转移图

这里的方块表示与幕结束对应的吸收状态。从 S_0 开始，我们就会得到收益序列 +1,+1,+1,0,0,0,$\cdots$。总之，无论我们是计算前 T 个收益 (这里 T=3) 的总和，还是计算无限序列的全部总和，我们都能得到相同的回报。即使我们引人折扣，这也仍然成立。因此，一般来说，我们可以根据式6来定义回报。我们也可以把回报表示为

$$G_t \doteq \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k \tag{7}$$

3.5 策略和价值函数

1. 策略和价值函数

严格地说，策略是从状态到每个动作的选择概率之间的映射。如果智能体在时刻 t 选择了策略 π，那么 $\pi(a|s)$ 就是当 $S_t = s$ 时 $A_t = a$ 的概率。

将策略 π 下状态 s 的价值函数记为 $v_\pi(s)$，即从状态 s 开始，智能体按照策略 π 进行决策所获得的回报的概率期望值。对于 MDP，我们可以正式定义 v_π 为

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s\right], \text{for all} s \in \mathcal{S} \tag{8}$$

其中，$\mathbb{E}_\pi[\cdot]$ 表示在给定策略 π 时一个随机变量的期望值，t 可以是任意时刻。我们把函数 v_π 称为策略 π 的状态价值函数。

类似地，我们把策略 π 下在状态 s 时采取动作 a 的价值记为 $q_\pi(s,a)$。这就是根据策略 π，从状态 s 开始，执行动作 a 之后，所有可能的决策序列的期望回报

$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s, A_t = a\right] \tag{9}$$

称 q_π 为策略 π 的动作价值函数。

2. 蒙特卡洛方法

价值函数 v_π 和 q_π 都能从经验中估算得到。比如，如果一个智能体遵循策略 π，并且对每个遇到的状态都记录该状态后的实际回报的平均值，那么，随着状态出现的次数接近无穷大，这个平均值会收敛到状态价值 $v_\pi(s)$。如果为每个状态的每个动作都保留单独的平均值，那么类似地，这些平均值也会收敛到动作价值 $q_\pi(s,a)$。我们将这种估算方法称作蒙特卡洛方法，因为该方法涉及从真实回报的多个随机样本中求平均值。

3. 贝尔曼公式

对于任何策略 π 和任何状态 s，s 的价值与其可能的后继状态的价值之间存在以下关系

$$
\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t \mid S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a) \Big[r + \gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s'] \Big] \\
&= \sum_a \pi(a|s) \sum_{s', r} p(s', r|s, a) \Big[r + \gamma v_\pi(s') \Big], \quad \text{for all} s \in \mathcal{S}
\end{aligned} \tag{10}
$$

式10被称为 v_π 的贝尔曼公式，它用等式表达了状态价值和后继状态价值之间的关系。如图3.3所示，从一个状态向后观察所有可能到达的后继状态。其中空心圆表示一个状态，而实心圆表示一个“状态-动作”二元组。从状态 s 开始，并将其作为根节点，智能体可以根据其策略 π，采取动作集合中的任意一个动作 (图中显示了三个动作)。对每一个动作，环境会根据其动态特性函数 p，以一个后继状态 s(图中显示了两个状态) 及其收益 r 作为响应。贝尔曼方程10对所有可能性采用其出现概率进行了加权平均。这也就说明了起始状态的价值一定等于后继状态的 (折扣) 期望值加上对应的收益的期望值。

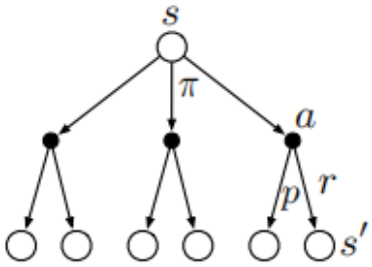

图 3.3: v_π 的回溯图

3.6 最优策略和最优价值函数

若对于所有的 $s \in \mathcal{S}, \pi \geqslant \pi'$，那么应当 $v_\pi(s) \geqslant v_{\pi'}(s)$。总会存在至少一个策略不劣于其他所有的策略，这就是最优策略。尽管最优策略可能不止一个，我们还是用 π_* 来表示所有这些最优策略。它们共享相同的状态价值函数，称之为最优状态价值函数，记为 v_*，其定义为：对于任意 $s \in \mathcal{S}$，

$$
v_*(s) \doteq \max_\pi v_\pi(s) \tag{11}
$$

最优的策略也共享相同的最优动作价值函数，记为 q_*，其定义为：对于任意 $s \in \mathcal{S}$，$a \in \mathcal{A}$

$$
q_*(s, a) \doteq \max_\pi q_\pi(s, a) \tag{12}
$$

对于“状态-动作”二元组 (s,a)，这个函数给出了在状态 s 下，先采取动作 a，之后按照最优策略去决策的期望回报。因此，我们可以用 v_* 来表示 q_*，

$$
q_*(s, a) = \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a] \tag{13}
$$

因为 v_* 是策略的价值函数，它必须满足贝尔曼方程 (式10) 中状态和价值的一致性条件。但因为它是最优的价值函数，所以 v_* 的一致性条件可以用一种特殊的形式表示而不拘泥于任何特定的策略。这就是贝尔曼最优方程。我们可以直观地理解为，贝尔曼最优方程阐述了一个事实：最优策略下各个

状态的价值一定等于这个状态下最优动作的期望回报。

式14和15为 v_* 的贝尔曼最优方程的两种形式。

$$\begin{aligned} v_*(s) &= \max_{a\in\mathcal{A}(s)} q_{\pi_*}(s,a) \\ &= \max_a \mathbb{E}_{\pi_*}[G_t \mid S_t=s, A_t=a] \\ &= \max_a \mathbb{E}_{\pi_*}[R_{t+1}+\gamma G_{t+1} \mid S_t=s, A_t=a] \\ &= \max_a \mathbb{E}[R_{t+1}+\gamma v_*(S_{t+1}) \mid S_t=s, A_t=a] \qquad (14) \\ &= \max_a \sum_{s',r} p(s',r|s,a)\big[r+\gamma v_*(s')\big] \qquad (15) \end{aligned}$$

式16为 q_* 的贝尔曼最优方程。

$$\begin{aligned} q_*(s,a) &= \mathbb{E}\left[R_{t+1}+\gamma \max_{a'} q_*(S_{t+1},a') \,\middle|\, S_t=s, A_t=a\right] \\ &= \sum_{s',r} p(s',r|s,a)\left[r+\gamma \max_{a'} q_*(s',a')\right] \qquad (16) \end{aligned}$$

图3.4展示了 v_* 和 q_* 的贝尔曼最优方程中进行的状态和动作的扩展过程，左图表示15，右图表示16

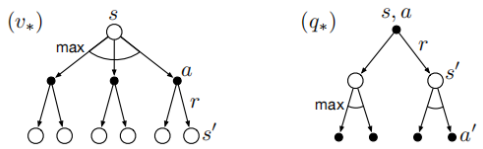

图 3.4: v_* 和 q_* 的回溯图

对于有限 MDP 来说，v_π 的贝尔曼最优方程 (式15) 有独立于策略的唯一解。贝尔曼最优方程实际上是一个方程组，每个状态对应一个方程等式。如果环境的动态变化特性 p 是已知的，那么原则上可以用解非线性方程组的方法来求解 v_* 方程组。类似地，我们也可以求得 q_* 的一组解。

显式求解贝尔曼最优方程给出了找到一个最优策略的方法。但是，这个方法很少是直接有效的。这类似于穷举搜索，预估所有的可能性，计算每种可能性出现的概率及其期望收益。这种解法至少依赖于三条在实际情况下很难满足的假设：(1) 我们准确地知道环境的动态变化特性;(2) 有足够的计算资源来求解 (3) 马尔可夫性质。

在强化学习中，我们通常只能用近似解法来解决这类问题。许多不同的决策方法都被视为近似求解贝尔曼最优方程的途径。例如，启发式搜索方法可以被视为对公式15的等号右侧项进行一定深度的展开，以形成一个概率“树”然后用启发式评估函数来估算“叶子”节点的”(类似于 A^* 这样的启发式搜索方法几乎都基于分幕式的情况)。动态规划算法与贝尔曼最优方程的关系更近，它们都是基于实际经历过的历史经验来预估未来的长期或全局期望值。

3.7 最优性和近似算法

按照上面的方法，智能体可以很好地学习到最优策略，但是最优策略通常需要极大量的计算资源。对最优性的完备严格定义对提到的学习方法进行了组织，并且提供了对这些不同学习算法的理论性质

的理解方式。但实际情况是，这些都是理想情况，真实情况下的智能体只能采用不同程度的近似方法。即使我们有个关于环境动态变化特性的完备精确模型，贝尔曼最优方程仍然不能简单地计算出一个最优策略。现在智能体面临的一个关键问题就是能用多少计算力，特别是每一步能用的计算力。

存储容量也是一个很重要的约束。价值函数、策略和模型的估计通常需要大量的存储空间。在状态集合小而有限的任务中，用数组或者表格来估计每个状态 (或 “状态-动作二元组) 是有可能的。我们称之为表格型任务，对应的方法我们称作表格型方法。但在许多实际情况下，经常有很多状态是不能用表格中的一行来表示的。在这些情况下，价值函数必须采用近似算法，这时通常使用紧凑的参数化函数表示方法。

3.8 总结

本章介绍了强化学习的各个组成部分。在强化学习中，智能体及其环境在一连串的离散时刻上进行交互。这两者之间的接口定义了一个特殊的任务：动作由智能体来选择，状态是做出选择的基础，而收益是评估选择的基础。智能体内部的所有事物对于智能体来说是完全可知、完全可控的。而智能体外部的事物则是部分可控的，可能完全知道其知识，也可能不完全知道。策略是一个智能体选择动作的随机规则，它是状态的一个函数。智能体的目标，就是随着时间的推移来最大化总的收益。

当强化学习用完备定义的转移概率描述后，就构成了马尔可夫决策过程 (MDP)。有限 MDP 是指具有有限状态、动作、收益集合的 MDP。目前大多数强化学习理论都只局限于讨论有限 MDP，但是相关方法和思路的应用范围却可以更加广泛。

回报是智能体要最大化的全部未来收益的函数 (最大化概率期望值)。根据不同任务和是否希望对延迟的收益打折扣，它也有不同的定义。非折扣形式适用于分幕式任务，在这类任务中，智能体-环境交互会被自然分解为幕；折扣形式适用于持续性任务，在这类任务中，智能体-环境交互不会被分解为幕，而是会无限制持续下去。我们通过定义合理的回报形式，可以用同样的公式来统一描述分幕式和持续性任务这两种情况。

一旦智能体确定了某个策略，那么该策略的价值函数就可以对每个状态或 “状态-动作” 二元组给出对应的期望回报值。最优价值函数对每个状态或 “状态-动作” 二元组给出了所有策略中最大的期望回报值。一个价值函数最优的策略就叫作最优策略。对于给定的 MDP，尽管状态或 “状态-动作” 二元组对应的最优价值函数是唯一的，但最优策略可能会有好多个。在最优价值函数的基础上，通过贪心算法得到的策略肯定是一个最优策略。贝尔曼最优方程是最优价值函数必须满足的一致性条件，原则上最优价值函数是可以通过这个条件相对容易地求解得到的。强化学习问题可以用不同的方式表示，这取决于智能体起初能获得的先验知识。对于完备知识问题，智能体会有一个完整精确的模型来表示环境的动态变化。如果环境是 MDP，那么这个模型就包含了一个四参数转移函数 p(式1)。对于非完备知识问题来说，我们则没有完整的环境模型。

即使智能体有一个完整精确的环境模型，智能体通常也没有足够的计算能力在每一时刻都全面地利用它。可用的存储资源也是另一个重要的约束。精确的价值函数、策略和模型都需要占用存储资源。在大多数实际问题中，环境状态远远不是一个表格能装下的，这时就需要近似方法来解决问题。