第九章　探索与利用

在强化学习问题中，探索和利用是一对矛盾：探索尝试不同的行为继而收集更多的信息，利用则是做出当前信息下的最佳决定。探索可能会牺牲一些短期利益，通过搜集更多信息而获得较为长期准确的利益估计；利用则侧重于对根据已掌握的信息而做到短期利益最大化。探索不能无止境地进行，否则就牺牲了太多的短期利益进而导致整体利益受损；同时也不能太看重短期利益而忽视一些未探索的可能会带来巨大利益的行为。因此如何平衡探索和利用是强化学习领域的一个课题。

根据探索过程中使用的数据结构，可以将探索分为：依据状态行为空间的探索 (state-action exploration) 和参数化搜索 (parameter exploration)。前者针对当前的每一个状态，以一定的算法尝试之前该状态下没有尝试过的行为；后者直接针对参数化的策略函数，表现为尝试不同的参数设置，进而得到具体的行为。

本章结合多臂赌博机实例一步步从理论角度推导得到一个有效的探索应该具备什么特征，随后介绍三类常用的探索方法：包括在前几章常用的衰减的 ϵ-贪婪探索、不确定优先探索以及利用信息价值进行探索等。

9.1　多臂赌博机

多臂赌博机 (图 9.1) 是一种博弈类游戏工具，它由多个拉杆，游戏这每当拉下一个拉杆后赌博机会随机给以一定数额的奖励，游戏者一次只能拉下一个拉杆，每个拉杆的奖励分布是相互独立的，且前后两次拉杆之间的奖励也没有关系。在这个场景中，赌博机相当于环境，个体拉下某一单臂赌博机的拉杆表示执行了一个特定的行为，赌博机会给出一个即时奖励 R，随即该状态序结束。因此多臂赌博机中的一个完整状态序列就由一个行为和一个即时奖励构成，与状态无关。

从上文的描述可以得出，多臂赌博机可以看成是由行为空间和奖励组成的元组 $\langle A, R\rangle$，假如一个多臂赌博机有 m 个拉杆，那么行为空间将由 m 个具体行为组成，每一个行为对应拉下某一

个拉杆。个体采取行为 a 得到的即时奖励 r 服从一个个体未知的概率分布：

$$R^a(r) = \mathbb{P}[r \mid a]$$

在 t 时刻，个体从行为空间 A 中选择一个行为 $a_t \in \mathbb{A}$, 随后环境产生一个即时奖励 $r_t \sim R^{a_t}$。

个体可以持续多次的与多臂赌博机进行交互，那么个体每次选择怎样的行为才能最大化来自多臂赌博机的累积奖励 $(\sum_{\tau=1}^{t} r_\tau)$ 呢?

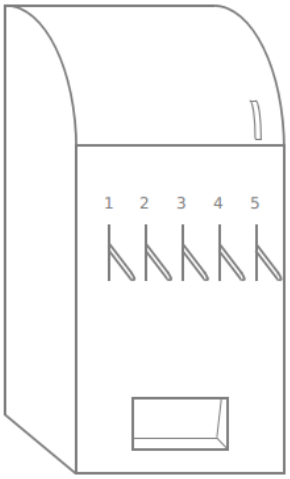

图 9.1: 多臂赌博机示意图

为了方便描述问题，定义行为价值 $Q(a)$ 为采取行为 a 获得的奖励期望：

$$Q(a) = \mathbb{E}[r \mid a])$$

假设能够事先知道哪一个拉杆能够给出最大即时奖励，那可以每次只选择对应的那个拉杆。如果用 V^* 表示这个最优价值，a^* 表示能够带来最优价值的行为，那么：

$$V^* = Q(a^*) = \underset{a \in A}{max} Q(a)$$

事实上不可能事先知道拉下哪个拉杆能带来最高奖励，因此每一次拉杆获得的即时奖励都与最优价值 V^* 存在一定的差距，定义这个差距为**后悔值** (regret):

$$l_t = \mathbb{E}[V^* - Q(a_t)]$$

每执行一次拉杆行为都会产生一个后悔值 l_t，随着拉杆行为的持续进行，将所有的后悔值加

起来，形成一个总后悔值:

$$L_t = \mathbb{E}\left[\sum_{\tau=1}^{t}(V^* - Q(a_\tau))\right]$$

这样最大化累计奖励的问题就可以转化为最小化总后悔值了。之所以要进行这样的转换，是由于使用后悔值来分析问题较为简单、直观。上式可以以另一种方式来重写。令 $N_t(a)$ 为到 t 时刻时已执行行为 a 的次数，Δ_a 为最优价值 $V*$ 与行为 a 对应的价值之间的差，那么总后悔值可以表示为:

$$\begin{aligned} L_t &= \mathbb{E}\left[\sum_{\tau=1}^{t} V^* - Q(a_\tau)\right] \\ &= \sum_{a\in A}\mathbb{E}\left[N_t(a)\right](V^* - Q(a)) \\ &= \sum_{a\in A}\mathbb{E}\left[N_t(a)\right]\Delta_a \end{aligned}$$

把总后悔值按行为分类统计可以看出，一个好的算法应该尽量减少执行那些价值差距较大的行为的次数。但个体无法知道这个差距具体是多少，可以使用蒙特卡罗评估来得到某行为的近似价值:

$$\hat{Q}_t(a) = \frac{1}{N_t(a)}\sum_{t=1}^{T} r_t\ 1(a_t = a) \approx Q(a)$$

理论上 V^* 和 $Q(a)$ 由环境动力学确定，因而都是静态的，随着交互次数 t 的增多，可以认为蒙特卡罗评估得到的行为近似价值 $(\hat{Q}_t(a))$ 越来越接近真实的行为价值 $(Q(a))$。图 9.2 是不同探索程度的贪婪策略总后悔值与交互次数的关系:

对于完全贪婪的探索方法，其总后悔值是线性的，这是因为该探索方法的行为选择可能会锁死在一个不是最佳的行为上；对于 $\epsilon-$ 贪婪的探索方法，总后悔值也是呈线性增长，这是因为每一个时间步，该探索方法有一定的几率选择最优行为，但同样也有一个固定小的几率采取完全随机的行为，导致总后悔值也呈现与时间之间的线性关系。类似的 softmax 探索方法与此类似。总体来说，如果一个算法永远存在探索或者从不探索，那么其总后悔值与时间的关系都是线性增长的。

能否找到一种探索方法，其对应的总后悔值与时间是次线性增长，也就是随着时间的退役总后悔值的增加越来越少呢？答案是肯定的，上图中衰减 ϵ-贪婪方法就是其中一种。下文将陆续介绍一些实际常用的探索方法。

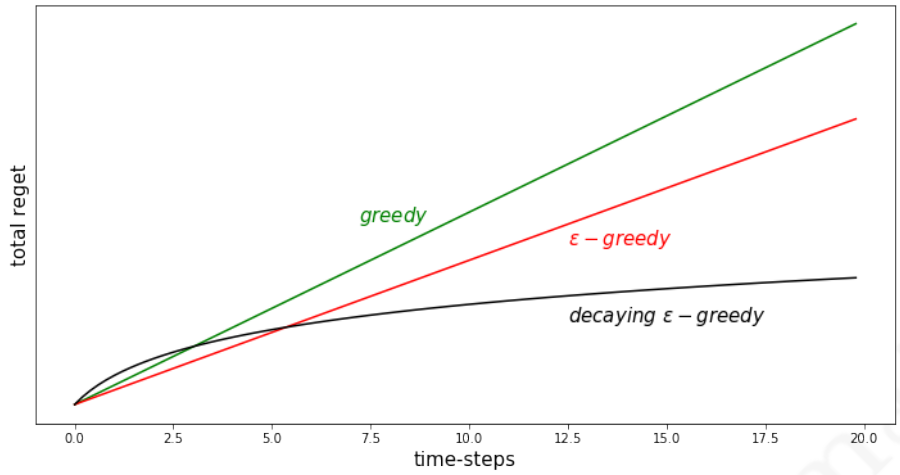

图 9.2: 不同探索程度贪婪策略的总后悔值

9.2 常用的探索方法

9.2.1 衰减的 ϵ-贪婪探索

衰减的 ϵ-贪婪探索是在 ϵ-贪婪探索上改进的，其核心思想是随着时间的推移，采用随机行为的概率 ϵ 越来越小。理论上随时间改变的 ϵ—t 由下式确定：

$$\epsilon_t = min\{1, \frac{c|A|}{d^2 t}\}, \quad d = \min_{a|\Delta_a>0} \Delta_i \in (0,1], \ c > 0 \tag{9.1}$$

其中 d 是次优行为与最优行为价值之间的相对差距。衰减的 ϵ-贪婪探索能够使得总的后悔值呈现出与时间步长的对数关系，但该方法需要事先知道每个行为的差距 Δ_a，实际使用是无法按照该公式来准确确定 ϵ_t 的，通常采用一些近似的衰减策略，这在之前几章已经有过介绍。

9.2.2 不确定行为优先探索

不确定行为优先探索的基本思想是，当个体不清楚一个行为的价值时，个体有较高的几率选择该行为。具体在实现时可以使用乐观初始估计、可信区间上限以及概率匹配三种形式。

乐观初始估计

乐观初始估计给行为空间中的每一个行为在初始时赋予一个足够高的价值，在选择行为时使用完全贪婪的探索方法，使用递增式的蒙特卡罗评估来更新这个价值：

$$\hat{Q}_t(a_t) = \hat{Q}_{t-1} + \frac{1}{N_t(a_t)}(r_t - \hat{Q}_{t-1}) \tag{9.2}$$

实际应用时，通常初始分配的行为价值为：

$$Q_*(a) = \frac{r_{max}}{1-\gamma}$$

不难理解，乐观初始估计由于给每一个行为都赋予了一个足够高的价值，在实际交互时根据奖励计算得到的价值多数低于初始估计，一旦某行为由于尝试次数较多其价值降低时，贪婪的探索将选择那些行为价值较高的行为。这种方法使得每一个可能的行为都有机会被尝试，由于其本质仍然是完全贪婪的探索方法，因而理论上这仍是一个后悔值线性增长的探索方法，但在实际应用中乐观初始估计一般效果都不错。

置信区间上限

试想一下如果多臂赌博机中的某一个拉杆一直给以较高的奖励而其它拉杆一直给出相对较低的奖励，那么行为的选择就容易得多了。如果多个拉杆奖励的方差较大，忽高忽低，但这些拉杆实际给出的奖励多数情况下比较接近时，那么选择一个价值高的拉杆就不那么容易了，也就是说这些大干虽然给出的奖励较接近，但实际上每一个拉杆奖励分布的均指却差距较大。可以通过比较两个拉杆价值的差距 (Δ) 以及描述其奖励分布相似程度的 KL 散度 ($KL(R^a||R^{a^*})$) 来判断总后悔值的下限。一般来说，差距越大后悔值越大；奖励分布的相似程度越高，后悔值越低。针对多臂赌博机，存在一个总后悔值下限，没有任何一个算法能做得比这个下限更好：

$$\lim_{t\to\infty} L_t \geq log\ t \sum_{a|\Delta_a>0} \frac{\Delta_a}{KL(R^a||R^{a^*})} \tag{9.3}$$

假设现在有一个三个拉杆的多臂赌博机，每一个拉杆给出的奖励服从一定的个体未知分布，先经过一定次数的对三个拉杆的尝试后，根据给出的奖励信息绘制得到如图 9.3 所示的奖励分布图。

根据图中提供的信息，在今后的行为选择中，应该有先尝试行为空间 $\{a_1, a_2, a_3\}$ 中的哪一个呢？从图中给出的三个拉杆奖励之间的相对关系，可以得出行为 $a3$ 的奖励分布较为集中，均值最高；行为 a_1 的奖励分布较为分散，均值最低；而行为 a_2 介于两者之间。虽然行为 a_1 对应的奖励均指最低，但其奖励分布较为分散，还有不少奖励超过了均值最高的行为 a_1 的平均均值，说明对行为 a_1 的价值估计较不准确，此时为了弄清楚行为 a_1 的奖励分布，应该优先尝试更多次行为 a_1，以尽可能缩小其奖励分布的方差。

从上面的分析可以看出，单纯用行为的奖励均值作为行为价值的估计进而指导后续行为的选择会因为采样数量的原因而不够准确，更加准确的办法是估计行为价值在一定可信度上的价值上限，比如可以设置一个行为价值 95% 的可信区间上限，将其作为指导后续行为的参考。如

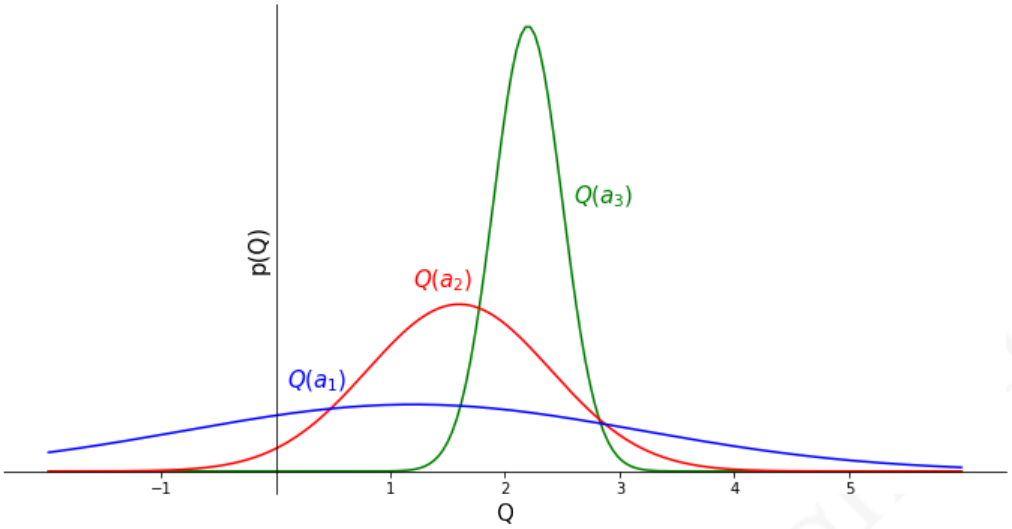

图 9.3: 根据实际交互奖励绘制的三个拉杆奖励分布

此一个行为的价值将有较高的可信度不高于某一个值:

$$Q(a) \leq \hat{Q}_t(a) + \hat{U}_t(a) \tag{9.4}$$

因此当一个行为的计数较少时，由均值估计的该行为的价值将不可靠，对应的一定比例的价值可信区间上限将偏离均值较多；随着针对某一行为的奖励数据越来越多，该行为价值在相同可信区间的上限将接近均值。因此，可以使用置信区间上限 (upper confidence bound, UCB) 作为行为价值的估计指导行为的选择。令:

$$a_t = \underset{a \in A}{argmax}(\hat{Q}_t(a) + \hat{U}_t(a)) \tag{9.5}$$

如果奖励的真实分布是明确已知的，那么置信区间上限可以较为容易地根据均值进行求解。例如对于高斯分布 95% 的置信区间上限是均值与约两倍标准差的和。对于分布未知的置信区间上限的计算可以使用公式 (9.6) 进行计算:

$$a_t = \underset{a \in A}{argmax}\left(Q(a) + \sqrt{\frac{2\ log\ t}{N_t(a)}}\right) \tag{9.6}$$

上式中 $Q(a)$ 是根据交互经历得到的行为价值估计，$N_t(a)$ 是行为 a 被执行的次数，t 是时间步长。这一公式的推导过程如下。

定理：令 $X_1, X_2, \ldots, X_t$ 是值在区间 $[0,1]$ 上独立同分布的采样数据，令 $\overline{X}_t = \frac{1}{\tau}\sum_{\tau=1}^{t} X_\tau$ 是采样数据的平均值，那么下面的不等式成立：

$$\mathbb{P}\left[\mathbb{E}[X] > \overline{X}_t + u\right] \leq e^{-2tu^2} \tag{9.7}$$

该不等式称为霍夫丁不等式 (Hoeffding's inequality)。它给出了总体均值与采样均值之间的关系。根据该不等式可以得到：

$$\mathbb{P}\left[Q(a) > \hat{Q}(a) + U_t(a)\right] \leq e^{-2N_t(a)U_t(a)^2}$$

该不等式同样描述了一个置信区间上限，假定某行为的真实价值有 p 的概率超过设置的可信区间上限，即令：

$$e^{-2N_t(a)U_t(a)^2} = p$$

那么可以得到：

$$U_t(a) = \sqrt{\frac{-\ log\ p}{N_t(a)}}$$

随着时间步长的增加，p 值逐渐减少，假如令 $p = t^{-4}$，上式则变为：

$$U_t(a) = \sqrt{\frac{2logt}{N_t(a)}}$$

由此可以得出公式 (9.6)。

依据置信区间上限 UCB 算法原理设计的探索方法可以使得总后悔值随时间步长满足对数渐进关系：

$$\lim_{t\to\infty} L_t \leq 8logt \sum_{a|\Delta_a>0} \Delta_a \tag{9.8}$$

经验表明，ϵ-贪婪探索的参数如果调整得当可以有很好的表现，而 UCB 在没有掌握任何信息的前提下也可以表现很好。

如果多臂赌博机中的每一个拉杆奖励服从相互独立的高斯分布，即:

$$R_a(r) = N(r; \mu_a, \sigma_a^2)$$

那么：

$$a_t = \underset{a\in A}{argmax}\left(\mu_a + c\sigma_a/\sqrt{N(a)}\right) \tag{9.9}$$

由于自然界许多现象都可以用高斯分布来近似描述，因此许多情况下可以使用上式来指导探索。

概率匹配

另一个基于不确定有限探索的犯法是概率匹配 (probability matching)，它通过个体与环境的实际交互的历史信息 h_t 估计行为空间中的每一个行为是最优行为的概率，然后根据这个概率来采样后续行为：

$$\pi(a \mid h_t) = \mathbb{P}\left[Q(a) > Q(a'), \forall a' \neq a \mid h_t\right]$$

实际应用中常使用汤姆森采样 (Thompson sampling)，它是一种基于采样的概率匹配算法，具体行为 a 被选择的概率由下式决定：

$$\begin{aligned} \pi(a \mid h_t) &= \mathbb{P}\left[Q(a) > Q(a'), \forall a' \neq a \mid h_t\right] \\ &= \mathbb{E}_{R|h_t}\left[1(a = \underset{a \in A}{argmax} Q(a))\right] \end{aligned} \tag{9.10}$$

以具有 n 个拉杆的多臂赌博机为例，假设选择第 i 个拉杆的行为 a_i 一共有 m_i 次获得了历史最高奖励，那么使用汤姆森采样算法的个体将按照：

$$\pi(a_i) = \frac{m_i}{\sum_{i=1}^{n} m_i}$$

给出的策略来选择后续行为。汤姆森采样算法能够获得随时间对数增长的总后悔值。

9.2.3 基于信息价值的探索

探索之所以有价值是因为它会带来更多的信息，那么能否量化被探索信息的价值和探索本身的开销，以此来决定是否有探索该信息的必要呢？这就涉及到信息本身的价值。

打个比方，对于一个有 2 个拉杆的多臂赌博机。假如个体当前对行为 a_1 的价值有一个较为准确的估计，比如是 100 元，这意味着执行行为 a_1 可以得到的即时奖励的期望；此外，个体虽然对于行为 a_2 的价值也有一个估计，譬如说是 70 元，但这个数字非常不准确，因为个体仅执行了非常少次的行为 a_2。那么获取 “较为准确的行为 a_2 的价值” 这条信息的价值有多少呢？这取决于很多因素，其中之一就是个体有没有足够多的行为次数来获取累计奖励，假如个体只有非常有限的行为次数，那么个体可能会倾向于保守的选择 a_1 而不去通过探索行为 a_2 而得到较为准确的行为 a_2 的价值。因为探索本身会带来一定几率的后悔。相反如果个体有数千次甚至更多的行为次数，那么得到一个更准确的行为 a_2 的价值就显得非常必要了，因为即使通过一定次数的探索 a_2，后悔值也是可控的。而一旦得到的行为 a_2 的价值超过 a_1，则将影响后续每一次行为的

选择。

为了能够确定信息本身的价值，可以设计一个 MDP，将信息作为 MDP 的状态构建对其价值的估计：

$$\tilde{M} = \langle \tilde{S}, A, \tilde{P}, R, \gamma \rangle$$

以有 2 个拉杆的多臂赌博机为例，一个信息状态对应于分别采取了行为 a_1 和 a_2 的次数，例如 $S_0 < 5, 3 >$ 可以表示一个信息状态，它意味着个体在这个状态时已经对行为 a_1 执行了 5 次，a_2 执行了 3 次。随后个体又执行了一个行为 a_1，那么状态转移至 $S_1 < 6, 3 >$。

由于基于信息状态空间的 MDP 其状态规模随着交互的增加而逐渐增加，因此使用表格式或者精确的求解这样的 MDP 是很困难的，通常使用近似架子和函数以及构建一个基于信息状态的模型并通过采样来近似求解。

虽然前文的这些探索方法都是基于与状态无关的多臂赌博机来讲述，但其均适用于存在不同状态转换条件下的 MDP 问题，只需将状态 s 代入相应的公式即可。