**强化学习理论学习与代码实现**

# 马尔科夫决策过程和贝尔曼等式（上）

在本章中将介绍有限马尔科夫决策过程（finite MDPS）的形式化问题，该问题涉及评价反馈，又涉及到关联方面—在不同的情形下选择不同的动作。MDPs是对序贯决策问题的一种典型的形式化表达，其中动作不仅会影响瞬时奖励，还会影响之后的情形或者说是状态，再以此影响未来的奖励。因此MDPs会涉及延迟奖励，需要在瞬时和延迟奖励之间作出权衡。在MDPs中，我们评估各个状态s下各个动作a的值q\*(s,a)，或者估计给定最优动作选择下的各个状态的值V\*(s)。这些依赖状态的量，对于准确地将长期结果的信誉（credit）分配给各个动作选择来说是必不可少的。

## 学习目标

* 理解代理-环境接口；
* 理解什么是马尔科夫决策过程（MDPs）以及如何解释转移图（变迁图）；
* 理解值函数、行为值函数和策略函数；
* 理解值函数、行为值函数的贝尔曼等式和贝尔曼最优等式。

## 代理-环境接口

MDPs旨在为从交互中学习以实现目标的问题提供一个直观的框架。学习者和决策者被称为代理（agent），代理所能交互的其之外的所有事物，称为环境（environment）。两者持续进行交互：代理选择动作，然后环境对动作做出反馈并将新的情形呈现给代理。环境也会带来一种特殊的数值奖励，随着时间的推移，代理通过其行为选择寻求最大化其累积值。

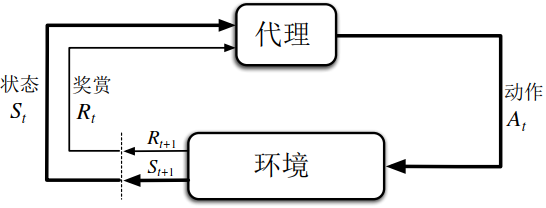


图1马尔可夫决策过程中的代理-环境交互

更具体地说，代理和环境在一系列离散的时间步t=0,1,2,3,…中交互。在每一个时间步t，代理接收环境状态St∈*S*的某一表达，并在此基础上选择一个动作At∈*A*(s)，一个时间步之后，在某种程度上作为其动作的结果，代理收到一个数值奖励Rt+1∈*R*，并进入下一状态St+1，因此马尔科夫决策过程与代理一起产生可如下的序列或轨迹：

S0,A0,R1,S1,A1,R2,S2,A2,R3,...

在有限MDP中，状态集、动作集与奖励集(*S*;*A*;*R*)，都只有有限个元素。在这一情况下，随机变量Rt与St都有明确的离散概率分布，且该分布仅依赖于前一状态与动作。也就是说，在给定前一状态与动作的情况下，这两个随机变量的特定值s’∈*S*及r∈*R*，出现的概率为：



函数p定义了MDP的动态性，等式中等号上的点提醒我们这是一个定义式，而非由先前的定义推导出的事实。动态函数p：*S*×*R*×*S*×*A*→[0;1]是有4个参数的、一般的确定性函数。其中“|”表示条件概率，但在这里其仅用于提醒我们：p指明了选择各个s与a的概率分布，且



在马尔科夫决策过程中，由p给出的概率可以完全描述环境的动态（完全可观的），也就是说，St和Rt取得各个可能值的概率只由前一个状态与动作，即St−1和At−1的具体值决定，而不受更早的状态或动作的任何影响。状态必须包括代理与环境的先前交互中所有能对未来产生影响的各方面的信息，如果符合此条件，那么我们称这样的状态拥有马尔科夫性质（Markov property）。

MDP框架抽象而灵活，可以以许多不同的方式应用于不同的问题。例如，时间步无需为真实时间的固定间隔，可以指任意连续的决策和动作阶段。动作可以是底层控制，例如加到机械臂马达上的电压，也可以是高层的决策，例如是否去吃午饭或去读研究生。**再例如，在我需要研究的调度问题中，动作可以是选择调度规则，或是直接选择工序。**一般来说，动作可以是任何我们希望学得如何决策的决定，而状态可以是任何我们认为可以帮助决策的信息。

关于代理与环境的界限，一般遵循的原则是：任何不能被代理任意更改的内容都被认为是在代理的外部，因此是其环境的一部分。我们并没有假设环境中的一切事物对于代理而言是未知的，例如，代理通常知道很多关于奖励是如何作为其行动和状态的函数来计算的，但是我们总是认为奖励计算对代理来说是外部的，因为它定义了代理所面临的任务，因此代理没有对其进行任意修改的能力。事实上，在一些情况下，代理可能知道环境运作的一切信息，但其面临的强化学习任务仍然非常困难，例如我们完全知道魔方这样的益智游戏的规则，但我们仍然无法解决它们。代理与环境间的边界代表了代理绝对控制能力的限制，而非其知识的限度。

## 目标和奖励

在强化学习中，代理的目的或目标可以通过一个被称为“奖励（reward）”的特殊信号进行形式化，且该信号从环境传递给代理。在每个时间步中，奖励是一个简单数字Rt∈*R*。代理的目标是使其获得的奖励总量最大化，这意味着最大化的不是瞬时奖励，而是长期累积奖励。我们可以把这个非正式的想法明确地表述为奖励假说（reward hypothesis）：

我们所说的目标和目的都可以看作是：最大化接收到的标量信号(称为奖励)的累积和的期望值。

使用奖励信号来形式化目标的概念，是强化学习最显著的特征之一。

尽管这种表示刚开始看起来有局限性，但是在实际应用中却被证明是及其灵活且具有广泛适用性。了解这一点的最好方法是考虑它是如何被使用的，或者可以如何被使用的例子。例如，为了让机器人学会走路，研究人员为机器人的每一步向前移动提供了与之成比例的奖励。在让机器人学习如何从迷宫中逃脱时，在逃跑前每走一步，奖励通常是1，以鼓励代理尽可能快地逃离。为了让机器人学会寻找和回收空汽水罐，人们可能会在大部分时间里给它零奖励，然后每回收一个汽水罐就给它+1奖励。当机器人撞到东西或有人对它大喊大叫时，也会给它负面的奖励。对于一个学习下跳棋或国际象棋的代理来说，自然的奖励是赢给予+1，输给予-1，平局和所有非终点位置给予0分。

## 回报和片段

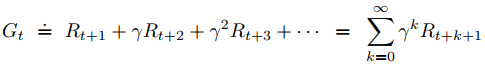
我们已经说过，代理的目标是最大化其长期获得的累积奖励，一般而言，我们希望最大化期望回报，其中回报（return）记为Gt，被定义为奖励序列的某一特定函数。在最简单的情况下回报就是奖励的和：



其中T是最后的时间步。这种方法当代理-环境的交互可以自然地分解为子序列时才有意义，这样的子序列称为片段（episode）。每一个片段都有一个终止状态，这个状态之后意味着整个过程重置成初始状态。

另一方面，在许多情形下，代理与环境的交互不能自然地划分为可辨别的分节，而是无限地继续下去。例如，可以很自然地以此种方式形式化一个持续进行的过程控制任务，或一个长生命周期的机器人应用。我们将这样的任务称为连续式任务（continuing task）。对连续式任务来说，上面的回报公式存在一定问题，因为最终时步T=∞,且我们希望最大化的回报可以很轻易地达到无穷（例如假设代理在每一时间步都接收+1的奖励），因此我们将采用一种概念上很复杂但数学表示上很简单的回报定义。

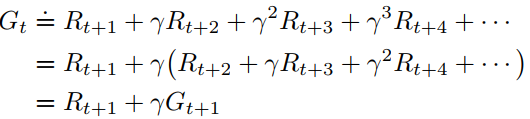
我们需要的另一个概念是折扣（discounting），根据此方法，代理试图选择可以最大化未来累积折扣奖励的动作：



其中γ为折扣率且0≤γ≤1。

折扣率决定了未来奖励的当前值：在k时间步之后接收的奖励，仅是其瞬时接收时的γk-1倍。如果γ<1,只要奖励序列{Rk}是有限的，上式的无限项之和也会是有限值；如果γ=0,代理只关心最大化瞬时奖励而变得短视：在这种情况下，代理的目标为选择At以仅仅最大化Rt+1。如果代理的每一个动作恰巧仅影响瞬时奖励而不会影响未来奖励，那么一个短视的代理可以通过分别最大化每一个瞬时奖励来最大化上式。但是一般来说，将瞬时奖励最大化会减少获得未来奖励的机会，从而降低回报。随着γ趋近于1，回报目标更强烈地考虑到未来的奖励回报，代理也变得更有远见。

相邻时间步中的回报彼此相关，其相关的方式对强化学习的理论与算法来说十分重要：



请注意如果我们定义GT=0，那么这对所有的t<T的时间步都成立，即使终止发生于t+1，这使得由奖励序列计算回报变得容易。

请注意折扣累积回报由无穷多项组成，但如果奖励为非0常数，且γ<1，那么回报值仍是有限的值。例如，如果奖励恒定为+1，那么回报为

