Sarsa算法和Q-Learning

我们在上一篇文章中介绍了最基本的TD-Learning，其核心思想就是利用RM算法求解基于state value的贝尔曼方程。本文依然沿着TD-Learning的核心思路，继续介绍两种非常常见的算法：Sarsa算法以及Q-Learning。我们希望在本文结束时能够回答下面的几个问题：

1. Sarsa算法和Q-Learning相比于最基础的TD-Learning有何不同?
2. Sarsa算法和Q-Learning的原理和流程是什么?

我们知道，基础的TD-Learning是在求解基于state value的贝尔曼方程，然而我们知道，对于action value 也有其对应的贝尔曼方程。因此我们先来推导一下基于action value的贝尔曼方程：

根据期望的定义，我们可以得到下面的式子：

我们发现，上面这个式子的左边为 ，因此有：

根据贝尔曼公式，我们有：

对于 的处理，我们应用state value和action value的关系，能够得到：

代入，得：

我们现在来分析一下式子： ，前一部分表示在当前状态 下，我执行了一个动作 之后，环境转移到下一个状态 的概率，由model决定。后一部分表示在这个新的下一个状态 下我采取动作 的概率。由全概率公式的意义可知，这两部分相乘则可以表示在当前状态 下采取了一个动作 之后，转移到下一个状态，并采取行动的概率。因此我们可以用一个 来表示。因此上式可以简化为：

由期望的定义，我们就可以将上式写为基于 的贝尔曼期望方程：

有了上一篇文章的基础，我们这里可以很容易写出用RM算法求解上述方程的迭代流程，首先写出带噪观测：

因此，update rule就可以写为：

这个就是经典的Sarsa算法的核心公式。下面我们来推导一下Q-Learning的核心公式，我们先直接给出Q-Learning的update rule：

我们先给出结论：Q-Learning是利用RM算法求解贝尔曼最优公式。因此，我们来证明一下上面的式子就是贝尔曼最优公式，首先我们回顾一下Bellman Optimal Equation：

我们知道state value和action value又有下面的关系：

所以贝尔曼最优方程又等价于：

将其代入式子 ()，可得：

接下来，对于Q-Learning的Update rule，根据RM算法我们可以倒推得出：Q-Learning是在用RM算法求解下面这个方程：

对上式，我们将期望展开，得到：

对等式两边同时取最大值，可得：

因此，我们就证明出了Q-Learning就是在求解贝尔曼最优方程。我们在本文中所讨论的Sarsa算法和Q-Learning都是基于表格的，当state-action pair特别多时就会导致维度爆炸效应。因此，后续我们将引入函数近似，同时将深度学习方法引入。同样地，基于深度学习或函数近似的方法也分为value-based和policy-based。Value-based范畴中最经典的就是Deep Q-Network (DQN)，policy-based范畴中就是策略梯度法。