Chapter 1

动态规划

动态规划(Dynamic Programming简称(DP))是一种算法的总称。给定一个完美的环境模型作为马尔可夫决策过程(Markov decision process (MDP))，这种算法可以计算出最有策略。经典的动态规划算法在强化学习中被应用得很少，一是因为要已知完美模型，二是因为计算量很大，但在理论上这种算法仍然很重要。动态规划为理解本书呈现的方法提供了必不可少的基础。事实上，所有这些方法都可以被视为为达到动态规划同等效果的尝试，只不过它们的计算量较少，也不需要完美的环境模型。

从这章开始，我们通常会假设环境是一个有限的MDP。 [....Sachen kommen noch!]

Chapter 2

蒙特卡罗方法

这一章是我们本书探讨的第一个学习方法来评估价值函数和发现最优策略。不像以往的章节，这里我们不会假设一个完全已知的环境。蒙特卡罗方法仅需经验—从实际或模拟的与环境的互动中获得的状态样本序列，动作和奖励。从实际经验中学习是很值得关注的，因为它不需要环境动态的先验知识，却仍然能够做出最优决策。从模拟的经验中学习也是很强大的。尽管我们需要模型，这个模型也只需要生成样本转移(sample transitions),而不是所有可能转移的完整概率分布，像动态分布(DP)要求的那样。令人惊奇的是，在许多情况下，按照所需概率分布生成经验样本是简单的，而得到精确的分布却不可行。

[Sachen kommen noch!]

2.1 蒙特卡罗预测

[Sachen kommen hier noch!]

Example 5.1: Blackjack 二十四点纸牌赌博游戏的目的是得到尽可能大的点数，但这个点数不能超过21。所有人头牌(J, Q, K)算作10，A可以算作1或11。我们这里使用的版本是每一个玩家独立与庄家(dealer)对抗。游戏开始时，玩家与庄家各得两张牌。发牌者的一张牌面朝上为“明牌”，其他牌点数朝下为“暗牌”。如果玩家这时已经达到21点(一个A和一个10点),这个牌面叫自然*(natural)*。玩家赢得胜利，除非庄家也是自然，那么游戏

平局(draw)。如果玩家没有自然，他要继续一个一个地叫牌(*hits*)，直到他停牌(*sticks*)，或者超过21点就是爆牌(*goes bust*)。如果玩家爆牌，玩家输；如果他停牌，就轮到庄家叫牌。庄家叫牌或停牌的策略是固定的：大于等于17点时庄家停牌，没到则叫牌。如果庄家爆牌，则玩家胜；如果庄家没有爆牌，则其余玩家要揭开手中的牌，比较点数，点数大的取胜。

[Sachen kommen hier noch!]

Example 5.2: Soap Bubble 试想一个线框围成的闭环浸在肥皂水里形成的一个肥皂膜或肥皂泡。如果线框的几何形状是已知的不规则的形状，怎样才能计算出肥皂泡表面的形状？这个表面的特性是每一点受到周围点的合力为零(否则形状会改变)。这表示表面上任一点的高度是围绕在以这一点为中心的小圆中所有点的高度的平均值。另外，表面的边界必须与线框相接。通常解决这类问题的方法是用网格覆盖表面，并通过迭代计算解出网格点上表面的高度。表面边界的网格点被规定落在线框上，所有其他的点被调整为与它相邻四点高度的平均值。这个过程会进行迭代，就像动态规划的迭代策略评估，并最终收敛于一个近似的表面。

蒙特卡罗方法最初的设计目的就是用于解决类似的问题。

Chapter 3

时序差分学习(Temporal-Difference Learning)

如果我们要选出一个核心而新颖的强化学习的思想，它无疑是时序差分学习(TD)。时序差分学习是蒙特卡罗方法和动态规划思想的结合。类似蒙特卡罗方法，时序差分方法能够在没有环境动态模型的情况下直接通过原始经验学习。像动态规划，时序差分法更新评估是部分基于其他学习到的评估，而不用等待最后的结果(他们用自助法(bootstrap))。时序差分，动态规划与蒙特卡罗方法之间的关系是强化学习理论中一个不断重复出现的主题; 这一章是我们探索的开始。在我们结束之前，我们会看到这些思想和方法会相互渗透并在很多方面相互结合。特别地，在第七章我们介绍n步自助法(n-step bootstrapping),这个方法给时序差分和蒙特卡罗方法提供了一个桥梁，在第十二章我们会介绍TD(λ) 算法，它可以把这两种算法天衣无缝地统一起来。

3.1 时序差分预测

时序差分和蒙特卡罗这两种方法都是用经验来解决预测问题。已知一些经验遵从策略π，两种算法为在经验中经过的非终点状态S_t更新他们的

估计值V (V of v_π)。简单地说，蒙特卡罗方法要等到访问某一状态的回馈已知，才会把回馈作为$V(S_t)$的目标。一个简单的适合非固定环境的every-visit MC方法是：

$$V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)] \tag{3.1}$$