MDP框架是抽象且灵活的，可以以许多不同的方式应用于许多不同的问题。 例如，时间步长不必指实时的固定间隔； 他们可以指代决策和行动的任意连续阶段。 这些动作可以是低级控制（例如，施加到机械臂的电机上的电压），也可以是高级决定（例如，是否吃午餐或去读研究生）。同样，各州可以采取多种形式。 它们可以完全由低水平的感觉（例如直接读取传感器的读数）确定，也可以更高级和抽象，例如房间中物体的符号描述。 构成状态的某些原因可能是基于对过去感受的记忆或甚至完全是精神上的或主观的。例如，代理人可能处于不确定对象在哪里的状态，或者只是在某种明确定义的意义上感到惊讶。 同样，某些动作可能完全是思维或计算上的。 例如，某些操作可能会控制座席选择思考的内容或将注意力集中在什么地方。 一般而言，行动可以是我们想要学习如何做出决定的任何决定，而状态可以是我们所知道的可能对做出决定有用的任何事情。特别是，代理人和环境之间的边界通常与机器人或动物身体的物理边界不同。 通常，边界比该边界更靠近主体。 例如，通常应将机器人及其感应硬件的电机和机械链接视为环境的一部分，而不是代理的一部分。 同样，如果将MDP框架应用于人或动物，则应将肌肉，骨骼和感觉器官视为环境的一部分。 奖励也大概是在自然和人工学习系统的物理内部计算的，但被认为是在代理外部的。

我们遵循的一般规则是，任何不能由代理人随意更改的事物都被视为在其外部，因此是其环境的一部分。 我们不假定代理程序不了解环境中的所有内容。 例如，代理经常对如何根据其行为以及所采取的状态来计算其酬劳非常了解。 但是我们始终认为奖励计算是主体的外部行为，因为它定义了主体所面临的任务，因此必须超出其任意更改的能力。 实际上，在某些情况下，特工可能知道有关其环境如何运作的所有信息，并且仍然面临艰巨的强化学习任务，就像我们可能确切地知道像Rubik的魔方那样的难题是如何工作的，但仍然无法解决。 代理人与环境的边界代表了代理人绝对控制权的限制，而不是其知识的极限。

可以将代理人与环境的边界放置在不同的地方以达到不同的目的。 在复杂的机器人中，可能同时运行许多不同的代理，每个代理都有自己的边界。 例如，一个代理可以做出高层决策，这些决策构成了实现高层决策的较低层代理所面对的状态的一部分。 在实践中，一旦选择了特定的状态，行动和奖励，并因此确定了特定的决策任务，就确定了主体与环境的边界。

MDP框架是对基于交互的目标导向学习问题的相当抽象。 它提出，无论感觉，记忆和控制装置的细节如何，以及目标试图达到的目的是什么，任何学习目标导向行为的问题都可以简化为在代理与其环境之间来回传递的三个信号： 一种信号代表代理做出的选择（动作），一种信号代表做出选择的依据（状态），另一种信号定义代理的目标（回报）。 该框架可能不足以有效地表示所有决策学习问题，但是事实证明它是广泛有用和适用的。

当然，特定的状态和动作因任务而异，并且它们的表示方式会极大地影响性能。 在强化学习中，就像在其他类型的学习中一样，这种代表性的选择目前比艺术更多地是艺术。

在本书中，我们提供了一些有关表示状态和动作的良好方法的建议和示例，但是我们的主要重点是选择表示形式后学习如何表现的一般原则。

例3.1：生物反应器假设正在应用强化学习来确定生物反应器（用于生产有用化学品的大量营养物质和细菌）的瞬时温度和搅拌速率。 在这样的应用中，动作可能是目标温度和目标搅拌速度，这些目标温度和目标搅拌速度会传递到较低级别的控制系统，进而直接激活加热元件和电机以达到目标。 状态可能是热电偶和其他感官读数，可能经过过滤和延迟，再加上代表大桶和目标化学品成分的符号输入。 奖励可能是生物反应器生产有用化学物质的速率的瞬时测量。 请注意，这里的每个状态都是传感器读数和符号输入的列表或向量，每个动作是由目标温度和搅拌速度组成的向量。 强化学习任务通常具有带有这种结构化表示的状态和动作。 另一方面，奖励始终是单个数字。

示例3.2：取放机器人考虑在重复的取放任务中使用强化学习来控制机器人手臂的运动。 如果我们想学习快速，平稳的运动，学习代理将必须直接控制电动机，并且需要有关机械连杆的当前位置和速度的低延迟信息。 在这种情况下，动作可能是在每个关节处施加到每个电动机的电压，状态可能是关节角度和速度的最新读数。 对于成功拾取并放置的每个对象，奖励可能为+1。 为了鼓励平稳的运动，在每个时间步长上，都可以根据运动瞬间的晃动来给予较小的负面奖励。

练习3.1设计适合您MDP框架的三个示例任务，为每个模型标识其状态，操作和奖励。 使这三个示例尽可能彼此不同。 该框架是抽象且灵活的，可以以多种不同方式应用。 在您的至少一个示例中以某种方式扩展其限制。 ⇤练习3.2 MDP框架是否足以代表所有目标导向的学习任务？ 您能想到任何明显的例外吗？

练习3.3考虑驾驶问题。 您可以根据油门，方向盘和制动器来定义动作，即身体与机器相遇的位置。

或者，您可以将它们定义得更远-例如，橡胶与道路的交汇处，将您的行为视为轮胎扭矩。 或者，您可以在更远的地方对它们进行定义，例如，大脑与身体相遇的地方，就是抽动肌肉来控制四肢。 或者，您可以提高到一个很高的水平，并说您的动作是您选择开车前往的地方。

在代理与环境之间划清界限的正确水平和正确位置是什么？

在什么基础上优先选择线路的哪个位置？ 是否有某个根本原因偏爱某个地点而不是另一个地点？

习题3.4给出一个类似于示例3.3的表，但对于p（s0，r | s，a）。 它应该有s，a，s0，r和p（s0，r | s，a）的列，并且每4元组有一行，其中p（s0，r | s，a）> 0。

3.2目标和奖励在强化学习中，根据从环境到主体的特殊信号（称为奖励）来形式化主体的目的或目标。 在每个时间步长处，奖励都是一个简单的数字Rt 2R。非正式地，代理商的目标是使所获得的奖励总额最大化。 这意味着从长远来看，不是最大化立即回报，而是最大化累积回报。 我们可以清楚地将这种非正式的想法表述为奖励假设：我们可以将目标和目的所指的所有含义都很好地理解为接收到的标量信号（称为奖励）的累加总和的期望值的最大化。

使用奖励信号来形式化目标的概念是强化学习的最鲜明特征之一。

尽管根据奖励信号制定目标乍看起来似乎很局限，但在实践中已证明它是灵活的并且广泛适用。 最好的查看方法是考虑如何使用或可能使用它的示例。 例如，为了使机器人学会走路，研究人员在每个时间步上都提供了与机器人的向前运动成比例的奖励。 在使机器人学会如何从迷宫中逃脱时，对于逃脱之前经过的每个时间步长，奖励通常是-1。 这鼓励特工尽快逃脱。 为了使机器人学会发现并收集空的汽水罐以进行回收，人们通常会给它零的报酬，然后每收集一罐的报酬为+1。 当机器人碰到东西或有人大喊大叫时，可能还会想给它负面的奖励。 对于要学习下棋或下棋的特工，自然的奖励是：获胜+1，输失败-1，抽奖和所有非终局位置0。

您可以看到所有这些示例中发生的事情。 代理总是学会最大化其报酬。 如果我们希望它为我们做些事情，我们必须以某种方式提供回报，以使代理商最大化地实现我们的目标。 它因此至关重要的是，我们建立的奖励能够真正表明我们想要实现的目标。

特别是，奖励信号不是向代理人传授有关如何实现我们想要做的事的先验知识的地方。5例如，下棋的代理人应该仅因实际获胜而获得奖励，而不是因为获得子目标而获得奖励。 夺取对手的棋子或控制棋盘中心。 如果对实现这些子目标有所奖励，那么代理可能会找到一种方法来实现它们而没有实现真正的目标。 例如，它可能找到一种方法来夺取对手的棋子，甚至以输掉比赛为代价。 奖励信号是您与业务代表沟通想要获得的结果的方式，而不是您想要如何获得的信号。6

3.3返回和情节到目前为止，我们已经非正式地讨论了学习的目标。 我们已经说过，代理商的目标是从长远来看获得最大的累积报酬。 如何正式定义呢？ 如果在时间步t之后收到的奖励序列表示为Rt + 1，Rt + 2，Rt + 3，...。 。 。，那么我们希望最大化该序列的哪个精确方面？

通常，我们试图使期望收益最大化，其中将收益（表示为Gt）定义为奖励序列的某些特定函数。 在最简单的情况下，回报是报酬的总和：Gt。= Rt + 1 + Rt + 2 + Rt + 3 +·+ RT（3.7）其中T是最后的时间步长。 这种方法在具有最终时间步长的自然概念的应用中很有意义，也就是说，当代理人与环境之间的相互作用自然分解为子序列时，我们将其称为情节7，例如玩游戏，穿越迷宫， 或任何形式的重复互动。 每个情节以称为终端状态的特殊状态结束，然后重置为标准起始状态或从起始状态的标准分布中抽样。 即使您认为情节以不同的方式结束（例如赢得或输掉一场比赛），下一个情节也将独立于上一个情节的结束而开始。 因此，可以将这些情节全部视为以相同的终末状态结束，并为不同的结果提供不同的奖励。 具有此类情节的任务称为情节任务。 在情节任务中，有时我们需要将所有非终结状态的集合（表示为S）与所有状态加上终结状态的集合（表示为S +）区分开。 终止时间T是一个随机变量，通常随情节而变化。

另一方面，在许多情况下，代理人与环境之间的相互作用不会自然地分解为可识别的事件，而是持续不断地不受限制。 例如，这将是制定正在进行的过程控制任务的自然方法，或者是使用寿命长的机器人的应用程序。 我们称这些持续的任务。 对于最终任务，收益公式（3.7）存在问题，因为最终时间步长为T = 1，而我们试图最大化的收益可能很容易无限。（例如，假设代理在每个时间步长都获得+1的回报。）因此，在本书中，我们通常使用收益的定义，该收益的概念在概念上稍微复杂一些，但在数学上却简单得多。

我们需要的其他概念是折扣。 根据这种方法，代理尝试选择操作，以便将来获得的折扣奖励的总和最大化。 特别是，它选择At来最大化期望的折现收益：Gt。= Rt + 1 +＃Rt + 2 +＃2Rt + 3 +··= 1X k = 0＃kRt + k + 1，（3.8）其中＃ 是参数0＃1，称为折现率。

折现率决定了未来奖励的现值：未来k个时间步长收到的奖励的价值仅为立即收到奖励的价值的＃k-1倍。 如果＃<1，则只要奖励序列{Rk}有界，（3.8）中的无穷和就具有有限值。 如果＃= 0，则代理仅在最大化即时回报方面是“近视”的：在这种情况下，其目标是学习如何选择At以便仅最大化Rt + 1。 如果每个特工的行为恰好只影响即时奖励，而不是将来的回报，那么近视特工可以通过分别最大化每个即时奖励来最大化（3.8）。 但是总的来说，最大程度地发挥即时奖励的作用会减少获得未来奖励的机会，从而减少回报。 当＃接近1时，回报目标会更强烈地考虑未来的回报； 代理商变得更有远见。

在连续时间步长上的收益以一种对强化学习的理论和算法很重要的方式相互关联：Gt。= Rt + 1 +＃Rt + 2 +＃2Rt + 3 +＃3Rt + 4 +··· = Rt + 1 +＃＃Rt + 2 +＃Rt + 3 +＃2Rt + 4 +···$ = Rt + 1 +＃Gt + 1（3.9）注意，这适用于所有时间步t <T，即使 如果终止发生在t + 1，假设我们定义GT =0。这通常使计算奖励序列的回报变得容易。

请注意，尽管收益（3.8）是无限多个项的总和，但是如果奖励为非零且恒定，则仍然是有限的，如果＃<1。例如，如果奖励为常数+1，则收益 是Gt = 1X k = 0 #k = 1 1-＃。 （3.10）练习3.5第3.1节中的方程式适用于持续的情况，需要稍作修改以适用于情节任务。 通过给出（3.3）的修改版本，表明您知道所需的修改。例3.4：杆平衡此任务的目的是向沿轨道移动的手推车施加力，以防止铰接在手推车上的杆掉落：如果杆掉落到给定角度，则发生故障。 从垂直方向或手推车在轨道上行驶。 每次故障后，磁极将重置为垂直。 这项任务可以看作是情节性的，自然事件是反复尝试平衡极点的过程。 在这种情况下，对于没有发生故障的每个时间步长，奖励可以为+1，因此每次返回的收益将是直到故障为止的步数。 在这种情况下，永远成功的平衡将意味着无限的回归。 或者，我们可以使用折价将极平衡视为一项持续的任务。 在这种情况下，每次失败的回报为-1，而在其他所有时间的回报为零。 然后，每次返回将与-＃K-1相关，其中K是故障之前的时间步数（以及以后故障的时间）。 在任何一种情况下，通过使磁极保持尽可能长的时间平衡，可以最大化返回。

习题3.6假设您将平衡点视为一个偶发性任务，但也使用了折现，除−1失败外，所有奖励均为零。 那么每次的回报是多少？ 这个收益与折扣，持续制定该任务的收益有何不同？ ⇤练习3.7假设您正在设计一个运行迷宫的机器人。 您决定为从迷宫中逃脱获得+1的奖励，而在其他所有时间获得0的奖励。 任务似乎自然分解为几集-连续穿过迷宫-因此您决定将其视为情节任务，目的是最大化预期的总奖励（3.7）。 运行学习代理一段时间后，您发现它从迷宫中逃脱并没有任何改善。 怎么了？ 您是否已有效地向代理传达了您想要实现的目标？ ⇤练习3.8假设＃= 0.5，并且收到以下奖励序列R1 = −1，R2 = 2，R3 = 6，R4 = 3，并且R5 = 2，T =5。G0，G1，。 。 。，G5？ 提示：向后工作。 ⇤练习3.9假设＃= 0.9，奖励顺序为R1 = 2，然后是7s的无限顺序。 什么是G1和G0？ ⇤练习3.10证明（3.10）中的第二等式。 ⇤