**强化学习理论学习与代码实现**

# 蒙特卡洛方法

不同于前面的章节，这里不再假设具备环境的完整知识，蒙特卡洛方法仅仅需要经验——从与环境的实际或仿真交互中得到的状态、动作和奖励采样序列。从实际经验中学习是惊人的，因为它不需要事先了解环境的动态，但仍然可以获得最优行为，从仿真经验中学习同样有用。虽然模型是必需的，但是模型只需要生成样本转移，而不需要生成动态规划(DP)所需的所有可能转移的完整概率分布。令人惊讶的是，在许多情况下，根据期望的概率分布很容易产生抽样的经验，但无法得到显式形式的分布。

蒙特卡罗方法是一种基于平均采样回报的强化学习方法，为了确保可用定义良好的回报，这里我们仅为片段任务定义蒙特卡罗方法。也就是说，我们假设经验被划分为多个片段，并且无论选择什么动作，所有片段最终都会结束。只有在一个片段完成时，价值评估和策略才会发生变化。因此，蒙特卡罗方法是在片段与片段上递增的，而不是时间步与时间步(在线)。

为了处理非平稳性问题，我们采用了第三章为DP开发的通用策略迭代(GPI)思想，在DP中我们从MDP的知识中计算值函数，而在这里，我们从MDP的样本回报中学习值函数。值函数和相应的策略仍然以本质上相同的方式（GPI）交互以获得最优性。在DP章中，我们首先考虑预测问题(固定任意策略的v和q的计算)，然后考虑策略改进，最后考虑控制问题及其用GPI求解。从DP中提取的每一个概念都被扩展到蒙特卡罗的情况，此时只有样本经验是可用的。

## 学习目标

* 理解预测与控制之间的不同；
* 知道如何使用MC方法预测状态值和状态-动作值；
* 理解在策略初访MC控制算法；
* 理解离策略MC控制算法；
* 理解加权重要性采样；
* 理解MC算法相较于动态规划法的优点。

## 蒙特卡洛预测

首先考虑蒙特卡洛方法用于学习给定策略下状态值函数的学习，一个状态的值为期望的回报，即从该状态开始的期望累积未来折扣奖励。因此，根据经验估计状态值的一个明显方法是，简单地对访问该状态后观察到的回报进行平均，当观察到更多的回报时，平均值应该收敛到期望值。这个思想是所有蒙特卡罗方法的基础。

假设希望估计vπ(s)，即为给定通过遵循π和经过s获得的一些列片段的情况下，策略π下状态s的值。一个片段中每次状态s的出现称为对s的一次访问（visit），当然s可能在一个片段中可以被访问多次，将一个片段中的第一次访问称为对s的初访（first-visit）。初访蒙特卡洛方法使用初访状态s时回报的平均值估计vπ(s)，而每次访问（every-visit）蒙特卡洛方法对所有访问s时的回报取平均值。这两种方法非常类似，但有稍微不同的理论性质。初访MC的研究最为广泛，可以追溯到20世纪40年代，也是我们在本章中重点研究的对象，每次访问MC可以更自然地扩展到函数逼近和资格迹。以下为初访MC的伪代码，每次访问MC的不同在于不需要检查在片段中St是否之前经历过。

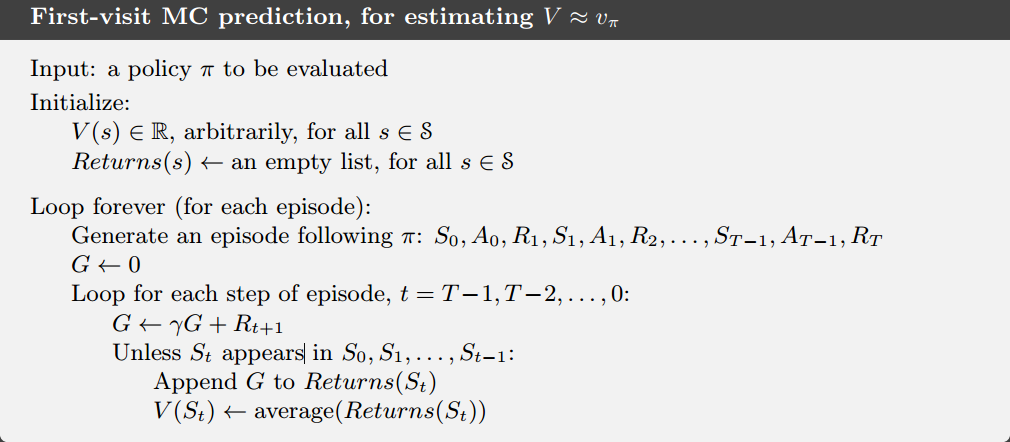


图1 初访MC算法

随着对s的访问（或者初访）趋于无穷，初访MC和每次访问MC都会收敛到vπ(s)，对于初访MC这是显而易见的，在这种情况下，每个回报都是有限方差下vπ(s)的独立同分布估计，根据大数定律，这些估计值的平均值序列收敛于其期望值，每一个平均值本身为无偏估计，误差的标准差依1/，其中n需要平均的回报数量。每次访问MC不是那么直截了当，但是其估计值也会二次收敛到vπ(s)。

## 蒙特卡洛动作值估计

如果无法获取模型，那么估计动作值（状态-动作对的值）比估计状态值有用的多。利用模型，仅使用状态值就足以确定策略，即通过一步前向观察，然后选择能带来最优的奖励和下一状态组合的动作。而在无模型的情况下，单凭状态值是不够的，还必须显示地估计每个动作的值，以便用于确定策略。因此蒙特卡洛方法的一个重要目标就是估计q\*，为此首先考虑动作值的策略评估问题。

动作值的策略评估问题就是估计qπ(s,a)，即从状态s开始，采取动作a，之后遵循策略π所能得到的期望回报。动作值估计的蒙特卡洛方法本质上与状态值估计相同，只是现在讨论的是对状态-动作对的访问，而不再是状态。如果状态s被访问并在该状态下执行动作s，则称在片段中对状态-动作对s,a进行了一次访问。每次访问MC方法将状态-动作对的值估计为对其所有访问之后回报的平均值，而初访MC中用于求平均的是每个片段中第一次访问的状态和采取的动作所获得的回报。和 之前一样，当访问每个状态-动作对的次数接近无穷时，这些方法以二次方式收敛到真实期望值。

唯一复杂的是，许多状态-动作对可能永远不会被访问。如果π是一个确定性策略，那么在遵循π时，将只观察来自每个状态的其中一个动作的返回。如果没有回报进行平均，其他动作的蒙特卡罗估计不会随着经验而改善。这是一个严重的问题，因为学习动作值的目的是帮助在每个状态中的可选动作中进行选择，为了比较这些可选方案，我们需要评估来自每个状态的所有动作的值，而不仅仅是当前偏好的动作。

这就是普遍存在的“maintaining exploration”问题，对于动作的策略评估问题，必须保证不断地进行探索，一种方式就是让每个状态-动作对都有一定的概率（大于0）被选中作为每个片段的开始节点，这样就能保证随着片段趋于无穷，所有的状态-动作对都能被访问无限次，我们将这个假设称为“exploring starts”。

这个“exploring starts”的假设有时是有用的，但在一般情况下却是不可靠的，特别是当直接与环境的实际交互中学习时，在那种情况下开始条件不太可能有太大帮助，因为真实情况下我们是不能去指定开始节点的。为了保证访问到所有的状态-动作对，最常用的方法是只考虑在每个状态下对所有动作选择概率都不为0的随机策略。在后面几节会讨论这种方法的两个重要变体，但是现在先保留“exploring starts”的假设并展示完整的蒙特卡洛控制方法。

## 蒙特卡洛控制

现在将考虑蒙特卡洛估计如何用于控制，也就是近似最优策略，总体思路是按照与DP章节相同的模式进行，即按照通用策略迭代（GPI）的思想。在GPI中，同时维护了近似策略和近似值函数，值函数反复更新以更逼近当前策略下的值函数，策略也利用当前值函数不断改进，如下图所示。这两种更改在某种程度上是相互矛盾的，因为它们都为另一种创建了一个移动的目标，但是它们一起使得策略和值函数都接近于最优。

首先，考虑经典策略迭代的蒙特卡洛版本，在这个方法中，交替执行完整的策略评估和策略改进，这个过程以任意策略π0开始，而以最优策略和最优值函数结束：



其中E表示一次完成的策略评估，I表示一次完整的策略改进。策略评估完全按照前面的章节中描述的那样进行，经历许多片段后，近似值函数渐进地逼近真实值函数。现在，假设确实观察了无数个片段，而且这些片段是使用起始点探索（exploring starts）产生的。在这些假设下，蒙特卡洛法将对任意的πk精确地计算qπk。

策略改进是通过对当前值函数贪婪地制定策略实现的，这种情况下就获得了动作值函数，因此不需要模型来构建贪婪策略。对任意的动作值函数q，对应的贪婪策略是对任意s∈S，可以确定性地选择使得动作值最大的动作的策略：

 （1）

然后通过构建每一个πk+1作为qπk的贪婪策略，将策略改进理论应用于πk和πk+1，因为对于所有的s∈S，



正如前面的章节中讨论的那样，策略改进理论保证了πk+1一律不劣于πk，或当两者都已经是最优策略时就一样好，这又保证了整个过程收敛到最优策略和最优值函数，通过这种方式，蒙特卡洛法可以在仅给定采样片段且没有环境动态知识的情况下，用于寻找最优策略。

为了简单地获得蒙特卡洛法的收敛性保证，这里我们做了两个不太可能实现的假设，一个是片段具有起始点探索，另一个是可以进行无限次策略评估。为了得到一个实用的算法，我们必须去掉这两个假设。我们把对第一个假设的考虑推迟到本章后面。

现在我们关注的假设是策略评估运行在无限次片段上，这个假设相对容易移除。实际上，同样的问题也会出现在经典的DP方法中，比如迭代策略评估，它也只渐进地收敛于真实值函数，对于DP和MC可以有两种方法解决这个问题。一种是坚持每一次策略评估中逼近qπk的观点，通过测量和假设来获得估计误差的大小和概率的界限，然后在每次策略评估期间采取足够的步数来确保这些界限足够小。这种方法可能完全令人满意，因为它在某种程度上保证了正确的收敛。然而，除了最小的问题之外，它还可能需要太多的片段才能在实践中发挥作用。

还有一种方法可以避免策略评估在名义上所需的无限个片段，我们放弃了在返回到策略改进之前试图完成策略评估，

首先，我们考虑如何计算任意策略π的状态值函数vπ，这在DP文献中被称为策略评估，我们也把它称为预测问题。回想上一章中，对于虽有的s∈S，



其中π(a|s)是策略π在状态s下执行动作a的概率，带下标π的期望表示其以遵循的策略π为条件。只要满足γ<1或者在策略π下，所有状态都会有终止状态，那么vπ就是存在且唯一的。

如果环境的动态转移特性是完全可知的，那么上式其实就是|S|个线性方程，带有|S|个未知数。原则上，可以进行直接求解，尽管有些繁琐。对于我们来说，迭代的计算方式是最适合的。假设我们有一系列的近似值函数v0,v1,v2...，每一个都把状态集合S+映射到实数集R。初始的近似值函数v0是任意选择的，然后后面的每一个近似值函数都使用vπ的贝尔曼方程（式（4））作为更新规则迭代求解。

 （5）

显然vk=vπ是这个更新规则的不动点，因为vπ的Bellman方程保证了我们在这种情况下是相等的。因此随着k→∞，序列{vk}在保证vπ存在的条件下一般收敛到vπ。这种算法称为迭代策略评估。

为了从vk得到后续的每一个近似vk+1，迭代策略评估针对每个状态s进行相同的操作：把当前状态s的旧值更新成一个新值，这个新值是由状态s后一状态的旧值和期望即时奖励，在待评估策略下沿着所有可能的状态转移概率求和得到。我们把这个迭代操作叫做期望更新（expected update）。迭代策略更新中的每次迭代都会更新一次所有的状态（full back up，完全备份），以得到vk+1的近似值函数。期望更新有多种不同的方式，这取决于当前更新的是状态（此处）还是状态-动作对，以及后一状态的估计值合并的方式。在DP算法中完成的所有更新都叫做期望更新，是因为这些更新都基于所有可能的下一状态的期望，而不是采用后一状态。

如果你想写一个描述迭代策略评估的程序，那么你应该使用两个数组，其中一个保存旧值vk(s)，另一个保存新值vk+1(s)。这样的话，新值就可以在不改变旧值的情况下被一个个计算出来。当然只使用一个数组更容易，可以用“in place”的方式更新值。这种方式下，某个状态更新后的值就会立即覆盖掉旧值。然后，根据状态更新的顺序，有时在式(5)的右边使用新值而不是旧值（可以这么理解：有时候恰好某个状态A的值更新较频繁，因为它与其他状态联系比较多，这样有可能下一次其他状态用到的A的值就是其更新两次或多次之后的）。“in place”算法同样也会收敛到vπ，实际上通常其收敛速度快于使用两个数组的版本，因为一旦可用就使用某状态的新值。我们认为更新是在状态空间的扫描中完成的。对于“in place”算法，状态值在扫描期间更新的顺序对收敛速度有显著影响。当我们考虑DP算法的时候，通常都考虑“in place”形式。

虽然策略迭代评估最终肯定会收敛，但是实际程序中我们需要设置一个终止条件。一个典型的终止条件就是当maxs∈S |vk+1(s)-vk(s)|的值非常小的时候，就让程序停止。

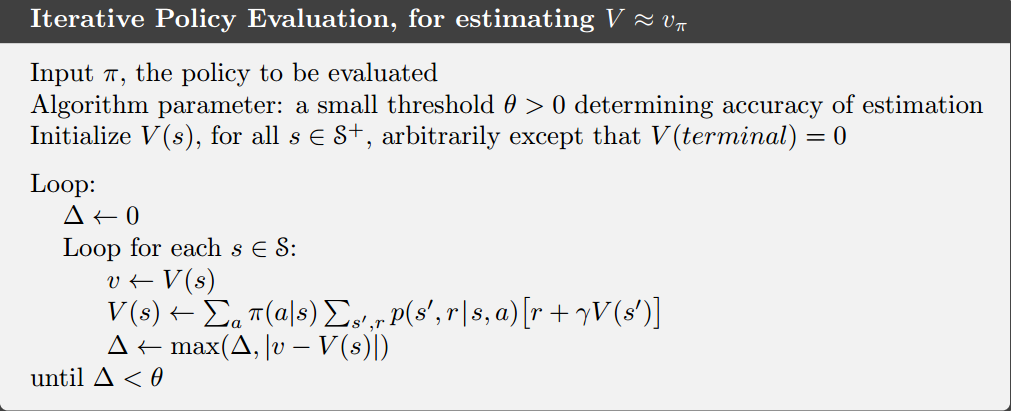


图2 迭代策略评估

## 策略改进

计算策略值函数的目的就是寻找更优的策略。假设已经确定了任一确定性策略π的值函数vπ，那么对于某一状态s，我们想要知道是不是可以改变一下策略来确定性地选择一个动作a≠v(s)。我们已经知道在状态s遵循当前策略的好坏，即vπ(s)，但是变成新策略后是变好了呢还是变坏了呢？一种方法就是在s考虑选择a，之后遵循现有的策略π，此时value迭代如下：

 （6）

关键标准是其是否大于vπ(s)，如果大于，说明在s选择a，之后遵循策略π会优于一直遵循策略π，那么在每次遇到状态s时仍然选择a会更好，新的策略在整体也更优。

这是一个称为策略改进定理（policy improvement theorem）的一般结果的特殊情况。设π和π’任意确定策略对，对于所有s∈S，有

 （7）

那么策略π’等同于或优于策略π，也就是说对于所有s∈S，其必会获得更大或相等的期望回报：

 （8）

而且，在任意状态下，如果式（7）存在严格不等式，那么在对应状态式（8）也必然是严格不等式。

策略改进理论适应于两个策略，原始的确定性策略π和改变的策略π’，π’与π的区别仅仅在于π’(s)=a≠π(s)。对于除了s的状态，式（7）成立因为两边相等，因此如果qπ(s,a)>vπ(s)，那么改变的策略优于π。

从式（7）出发，不断使用式（6）分解qπ，然后重新应用式（7）直到得到vπ’(s)：

 （9）

到目前为止，我们已经了解了如何在给定策略及其值函数的情况下，轻松地评估策略在单个状态下对特定动作的更改，很自然地可以考虑在所有状态下改变所有可能的动作，根据qπ(s,a)在每个状态选择看起来最好的动作。换句话说，考虑如下的贪婪策略：

 （10）

贪婪策略根据vπ采取短期看起来（一步向前看）最优的动作，其满足式（7）策略改进理论的条件，所以可知其等同于或优于原始策略。通过使新策略对原策略的值函数贪婪，从而对原策略进行改进，这样制定新策略的过程称为策略改进（policy improvement）。

假设新的贪婪策略π’和原策略π一样好，那么vπ’=vπ，根据式（10）对于所有s∈S，其满足：

 （11）

这和贝尔曼最优方程是一样的，因此vπ’一定是v\*，π和π’都是最优策略。换句话说，策略改进一定会给我们一个更好地策略，除非我们当前的策略已经是最优的了。

## 策略迭代

一旦一个策略π使用vπ得到了改进并得到了更优的策略π’，然后就可以计算vπ’，并再次对其进行改进从而得到更优的π’’，由此我们可以得到一个单调改进策略和值函数的序列：

 （12）

其中带有E的箭头表示策略评估，带有I的箭头表示策略改进，每个策略都保证比前一个策略有严格的改进(除非它已经是最优的)。由于有限的MDP只有有限数量的策略，因此该过程必然在有限次迭代中收敛到最优策略和最优值函数。

这种寻找最优策略的方法称为策略迭代，下面给出了其完整的算法。注意，每个策略评估本身就是一个迭代计算，从前一个策略的值函数开始。这通常会大大提高策略评估的收敛速度(可能是因为值函数在策略之间变化不大)。

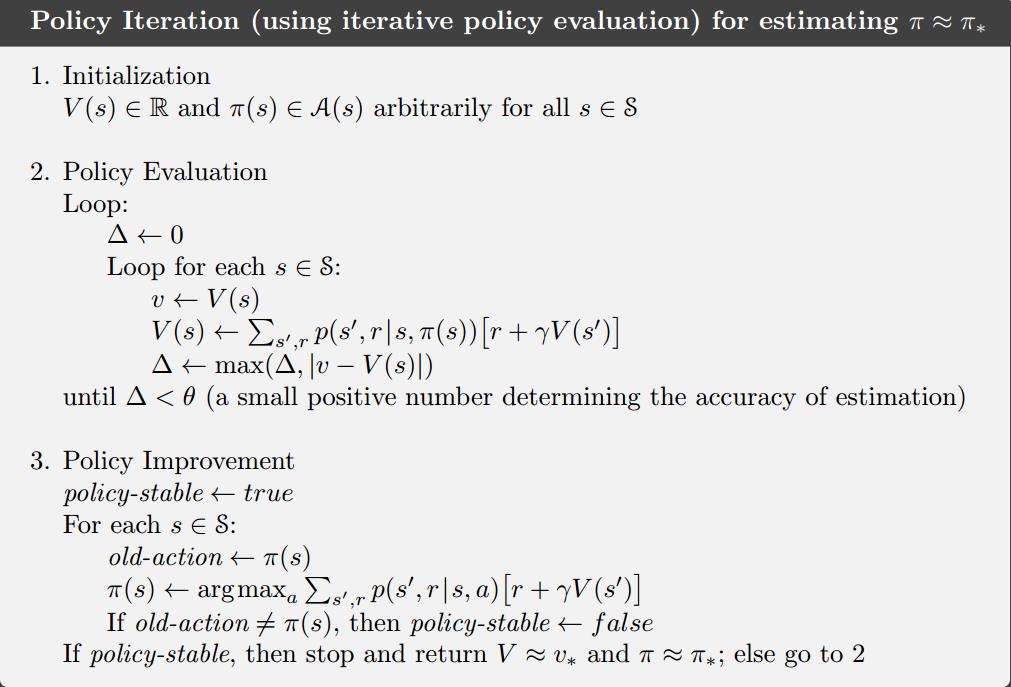


图3 策略迭代

## 值迭代

策略迭代的一个缺点是，它的每个迭代都包含策略评估，而策略评估本身可能是一个需要多次遍历状态集的冗长迭代计算。当然等到策略评估完成是好的，因为值函数会精确地收敛到vπ。我们必须等待道精确的收敛吗，或者我们能不能停下来？

实际上，可以采用多种方法截断策略迭代的策略评估步骤，而不丧失策略迭代的收敛性保证。其中一个特别的方式就是当所有状态都遍历一遍后，停止策略评估过程。这个算法就叫做值迭代（value iteration）。可以将其写为一个结合了策略改进和截断策略评估的简单更新操作：

 （13）

对于任意v0，序列{vk}在保证v\*存在的条件下可以收敛到v\*。

另外一个理解值迭代的方式是参考贝尔曼最优方程。我们注意到值迭代本质上是简单地把贝尔曼最优方程转化成一个更新规则，同意注意到值迭代的更新过程和之前的策略评估（式5）的过程是一模一样的，只是在值迭代过程中每次我们都需要所有动作中使得上式取最大值的那个。

最后，我们需要考虑一下值迭代如何终止。和策略评估相同，我们理论上需要无限次迭代使之精确收敛到v\*。在实际应用中，我们在相邻两次遍历状态的的差值非常小时，就停止运行。

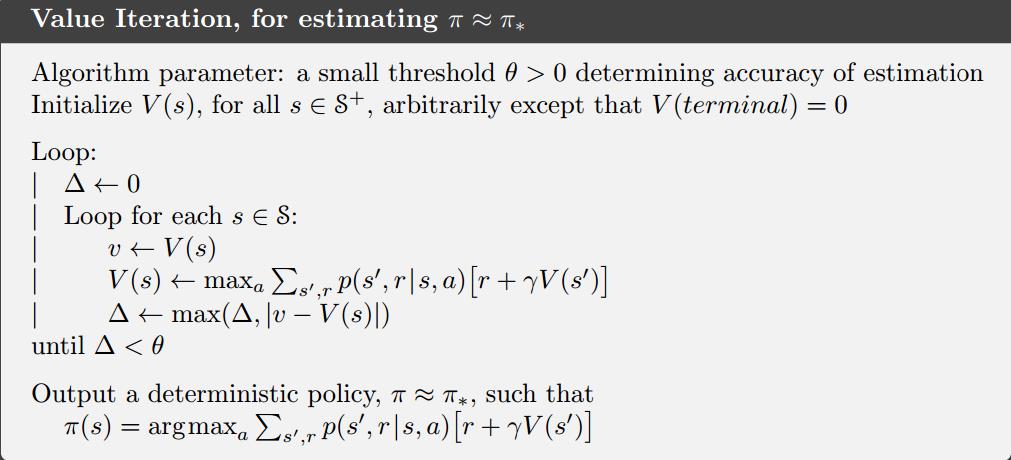


图4 值迭代

我们可以看出，值迭代高效地结合了策略评估和策略改进，通过在每个策略改进遍历之间插入多个策略评估遍历，通常可以实现更快的收敛。总体来讲，一般来说，整类截断策略迭代算法都可以看作是遍历序列，其中一些使用策略评估更新，另一些使用值迭代更新。因为仅有的不同之处只是上式（13）中的max操作，这只意味着是把max操作加入了策略评估过程。所有这些算法对于折扣优先MDP来说都会收敛到最优策略。

## 异步动态规划

到目前为止，我们讨论的DP算法的主要缺陷是我们必须在MDP的整个状态集上进行操作，即遍历状态集合。如果状态集数量非常大，那么就算是单次遍历也非常耗时。例如，西洋双陆棋有超过1020个状态，即使我们可以每秒对一百万个状态执行值迭代更新，完成一次遍历也需要超过一千年的时间。

异步DP算法是一种实时（in place）迭代的DP算法，它不是系统地对状态集进行扫描的。这些算法以任何顺序，使用任何其他状态可用的值更新状态值。某些状态的值可能会在其他状态的值更新一次之前多次更新。然而，为了正确收敛，异步算法必须继续更新所有状态的值：它不能忽略计算中某个点之后的任何状态。异步DP算法允许选择要更新的状态，具有很大的灵活性。

例如，其中一种异步值迭代使用值迭代更新（式（13））在每个步骤k上只更新（in place）一个状态sk的值。如果0<γ<1，只有在所有状态都在序列{sk}中无限次出现(序列甚至可以是随机的)的前提下，才能保证渐近收敛到v\*。(在非折扣的片段案例中，可能有一些更新顺序不会导致收敛，但相对容易避免这些情况。)类似地，可以将策略评估和值迭代更新混合使用，从而产生一种异步截断的策略迭代。

## 广义策略迭代

策略迭代由两个同步的交互过程组成，一个使值函数与当前策略保持一致(策略评估)，另一个根据当前值函数给出贪婪策略(策略改进)。在策略迭代中，这两个流程交替进行，每个流程在另一个流程开始之前完成，但这并不是真正必要的。例如，在值迭代中，在每个策略改进之间只执行一次策略评估迭代。在异步DP方法中，评估和改进过程以更细的粒度交织在一起。在某些情况下，在返回到另一个过程之前，会在一个过程中仅更新一个状态。只要这两个过程持续更新所有状态，最终的结果通常是相同的—收敛到最优值函数和最优策略。

我们使用术语通用策略迭代(GPI)来表示允许策略评估和策略改进过程交互的一般思想，而不关注两个过程的粒度和其他细节。几乎所有的强化学习方法都可以用GPI来很好地描述，也就是说，它们都具有可识别的策略和值函数，策略总是根据值函数进行改进，并且值函数总是被驱动到策略的值函数，如图4所示。如果评价过程和改进过程都稳定下来，即不再产生变化，那么值函数和策略一定是最优的。值函数只有在与当前策略相一致时才稳定，而策略只有对当前价值函数贪婪时才稳定。因此，只有当发现策略对自己的评估函数贪婪时，这两个过程才会稳定下来，这说明贝尔曼最优方程成立，因此策略和值函数都是最优的。

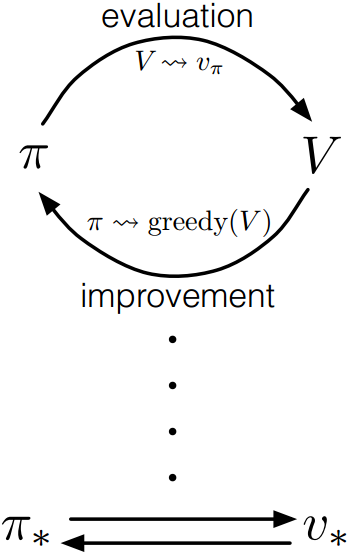


图5 GPI

GPI的评估和改进过程可以看作是竞争和合作的过程。它们相互竞争的意义在于它们向相反的方向拉扯。使策略对值函数贪婪通常会否定当前值函数相对于改变后的策略的正确性，值函数进行更新使之和当前策略相一致，也就意味着当前的策略对于更新后的值函数来说，不再贪婪。然而，在长期交互作用中，这两个子过程却可以找到同一个最终答案：最优的策略和最优的值函数。

我们也可以将GPI中评估和改进过程之间的相互作用考虑为两个约束或目标，例如，就像图5所示的二维空间中的两条线。虽然实际的几何比这复杂得多，但该图显示了实际情况。每个过程都将值函数或策略驱动到直线（表示目标的解）的其中一条。目标相互作用是因为这两条线不是正交的，直接朝一个目标前进会导致远离另一个目标。然而，联合过程不可避免地更接近优化的总体目标。图中的箭头对应于策略迭代的行为，因为每个策略迭代都将系统带到完全实现两个目标中的一个。在GPI中，也可以朝着每个目标采取较小的、不完整的步骤。在这两种情况下，这两个过程一起实现了优化的总体目标，尽管它们都没有尝试直接实现优化。

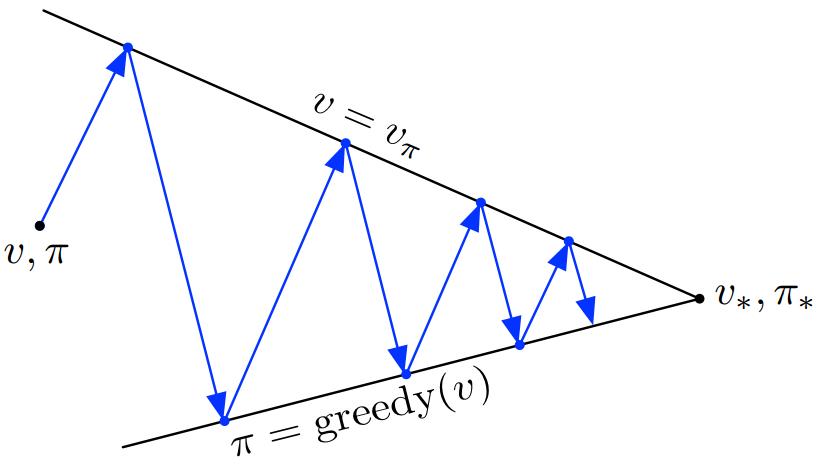


图6 迭代过程

## 总结

动态规划(DP)方法假设我们有一个完美的环境马尔可夫决策过程(MDP)模型。在实践中通常不是这样，但无论如何研究DP是很重要的。

策略评估：计算给定策略的状态值函数v(s)。在DP中，这是使用“完全备份”来完成的。在每一种状态下，我们一步前向观察每一个可能的行动和下一个状态。我们之所以能做到这一点，是因为我们有一个完美的环境模型。

完全备份基本上是将Bellman方程转换为更新。

策略改进：给定策略的正确状态值函数，我们就可以对它采取贪婪的动作(即在每个状态中选择最佳动作)。然后我们就可以保证改进这个政策，或者已经是最优时保持不变。

策略迭代：迭代地执行策略评估和策略改进，直到达到最优策略。

值迭代：我们不需要执行多个策略评估步骤来找到“正确的”v(s)，而是只执行一步并立即改进策略。实际上，这个收敛得更快。

通用策略迭代：迭代地进行策略评估和改进的过程。我们可以为每个步骤选择不同的算法，但基本思想是相同的。

DP方法自举：它们基于其他估计更新估计(向前一步)。

## 练习

### 策略评估

### 策略迭代

### 值迭代

### 赌徒问题