**POMDP建模**

现在让我们将焦点切换到部分可观的问题。在马尔科夫决策过程中只解决行动中的不确定性，但他们假设世界的状态可以有把握地确定。对于完全可观马尔可夫决策过程，以及存在很多很成熟的算法来解决问题。现在，我们考虑更一般的情况，其中的状态不是完全可观的，也即缺乏预测性。缺乏观测性是指传感器只能在可能的状态估计一个P后验分布，我们称之为belief。此集合也称为部分可观马尔可夫决策过程，或POMDPs。  
在本章的其余部分要解决的核心问题是：对于POMDPs，制定规划和控制算法。最优策略算法仅存在于有限的世界，其中状态空间，动作空间，观测空间，以及计划时域T均为有限的。不幸的是，这些精确的方法的计算是十分复杂的。除此之外，对于连续的情况下，最有名的算法都是近似的。  
在MDPs中的中心更新方程如下

1

在POMDPs中，我们应用非常接近的想法，然而，状态并不是可观的。所以，我们只可以得到状态的后验分布belief state b。因此对于POMDPs的计算方程如下

2

并且我们使用这个方程来生成控制

3

不幸的是，在belief空间计算价值要比在状态空间复杂的多。belief是一个概率分布，因此，在POMDPs中的价值是有关概率分布的函数。如果状态空间是有限的，由于belief space是所有状态空间的分布，belief state就是连续的。对于连续状态空间的问题，这个情况更加复杂，在这种情况下，belief space是一个无线维的连续体。除此之外，公式2和公式3对于所有的beliefs 进行积分。由于belief space的复杂度，积分并不容易精确计算或者有效地近似，  
幸运的是，对于有限的情况，存在确切的解法。

**有限环境**

下面考虑一种十分重要的特殊情况，有限环境。在这里，我们假设我们有一个有限状态空间，在每个状态具有有限的动作，有限数量的观测，和有限计划时域.在这些情况下，最优值函数可以被精确计算，也即具有最优策略。然而这并不是对所有情况都很明显，即使状态空间是有限的，事实上可以有无限多的可能的belief。然而，在这一点，下面我们建立一个众所周知的结果，最优值函数是凸的并且由有限多个线性部分组成。因此，即使该值函数被定义在一个连续体，它可以在计算机中以浮点数的精度来表示。  
说明性的例子  
一个有限的世界的一个例子示于图16.1。这个具体的例子极端简单，其唯一的作用是在讨论一般的解决方案前帮助熟悉POMDPs的关键问题。我们注意到，图16.1具有四种状态，标记至中，两个动作， 和，和两个观察，标记和。初始状态是随机抽取来自两个顶端状态，和。现在机器人给出一个选择：执动作作，这将具有高概率(但不总是)瞬间移动到相应的其他状态。另一种可选择方案，它可以执动作作，这导致在这两个底状态，或S4的概率中的一个的转变所示。在中，机器人将收到一个大的正收益，而在S4的收益为负。这两个state是终端的状态，也就是说，一旦进入任务就结束了。因此，机器人的目标是在状态时要执动作作。如果机器人知道它是在什么状态，执行任务是非常简单：简单地套用动作，直到状态是，然后应用动作。然而，机器人不知道其状态。相反，它可以感知两个观测和。不幸的是，这两个观测可能在这两种状态，虽然观察一比其他的概率不同，这取决于什么状态的机器人是由于机器人并不知道它的初始状态是或，它必须仔细跟踪过去的观测来计算它的belief。因此，关键的策略问题是：什么时候应该机器人尝试动作？具有多大的置信度去执动作作，需要多长时间来尝试这个动作？这些和其他问题将在下面进一步描述。

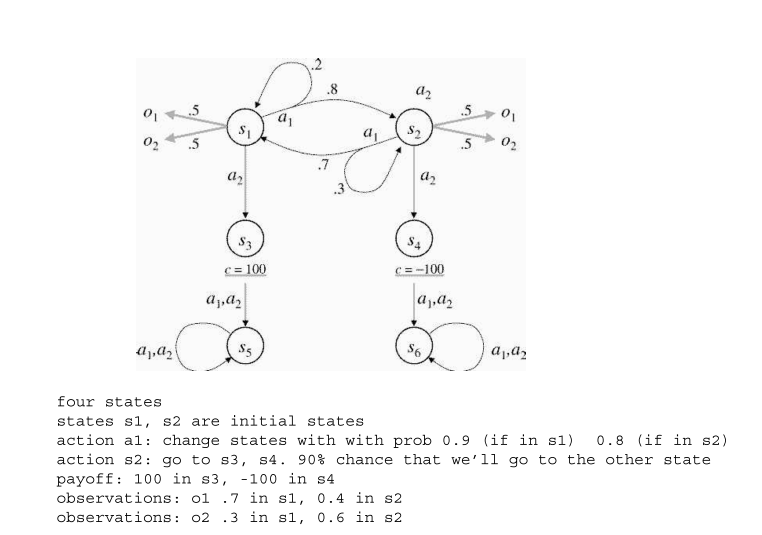
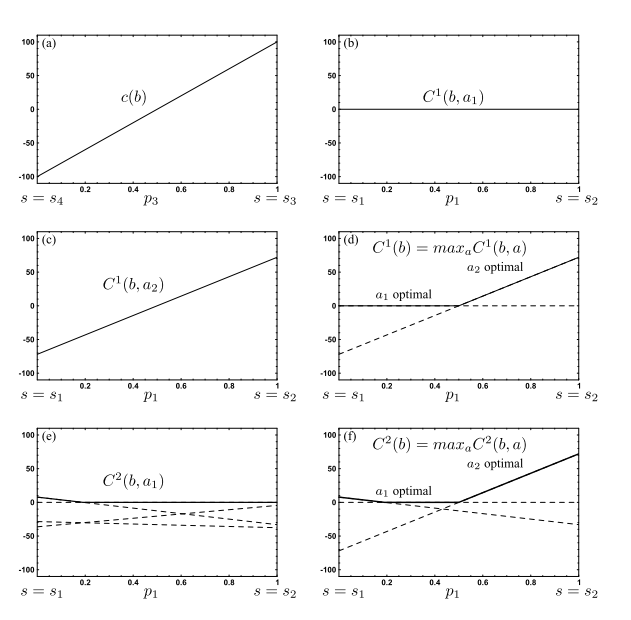
  
本节的目标是开发一种在有限状态和有限时域T的情况下来计算最优值函数的算法。让我们表示用T来表示计划时域和用表示状态。利用状态空间的有限，

图16.1有限状态环境，用来说明价值迭代的belief空间。

我们注意到，belief B(S)由n个概率给出  
(16.4)  
其中。  
belief必须满足的条件  
(16.5)  
因为最后一个条件的，b(s)可通过n-1个参数而不是n个参数指定。剩余的概率的可以根据如下公式计算：  
(16.6)。  
因此，如果状态空间是有限的，并且大小为n的话，则belief是一个(n-1)维时域。  
在我们的图16.1示出的例子中，可能出现的是由三个数值的概率值指定的belief，因为有四个状态。然而，操作将状态和状态分开。因此，该机器人可能遇到的唯一的不确定性是状态和或者状态和之间的的混淆。两者均可以由一个单一的数值的概率值来表示，由此，belief状态的唯一连续成分是一维的。这使得这个例子方便绘制值的函数。  
现在，让我们专注于一个问题，我们是否可以计算出最优值函数，准确地在有限的领域最优策略。如果没有，我们可能会被迫以接近它。起初，人们可能会得出这样的结论计算最优值函数是不可能的，由于这一事实，即belief空间是连续的。但是，我们观察到的有限的世界，值函数有一个特殊的形状：它是由有限多个线性件。这使得能够计算在有限时间内的最优值函数。我们还注意到，价值函数是凸和连续这后两种特性也适用于最优值函数在连续状态空间和无限的规划视野。  
我们开始的belief状态b的直接回报我们的考虑。回想一下，b的收益是由下概率分布b的收益c中的给定期望：  
(16.7)  
使用b唯一地由概率指定的事实，我们可以写  
(16.8)，它对于确实是线性。  
图16.2期望收益c(b)为belief参数 的功能，假设机器人无论是在状态或。  
有趣的是绘制函数c(b)在置信度分布在两种状态和在我们的例子中，只有国家与非零回报。在状态回报是100，而在状态的回报是-100。图16.2a示出了函数c(b)为相信通过？0,0，第3页，限定的子空间？这是一种belief的空间，放置在美国和所有的可能性。这里的期望c(b)的绘制为  
函数的概率的。显然，如果时，环境状态是，并回报将是  
。另一方面，如果，环境的状态是和回报将是。在此期间，期望是线性的，从而导致在图16.2a所示的曲线图。  
这种考虑使我们能够计算价值函数规划时域。从现在开始，我们将只考虑belief空间的地方的子集  
在两种状态和所有的可能性。这种belief空间是由一个单一的参数参数化，的，因为和的值函数中是恒定的零动作：  
(16.9)  
因为无论机器人的真实状态，动作不会导致它的状态，使得它收到非零的回报。该值函数如果绘制如图16.2b。图片变得更有趣动作。如果环境的状态是，这个动作将导致的几率状态，其中机器人将收到的100的收益有的概率会结束在状态，在那里它会回报是-100。因此，在状态的预期收益为。类似的说法，状态的预期收益为-80。在这两者之间，期望是线性的，产生的价值函数  
 (16.10)  
这里我们使用的折现因子。这个函数表示在图16.2c，该类型的belief  
  
那么，什么是正确的动作选择策略？继最大化预计这将对收益的理由，最好的动作取决于我们目前的belief，假设它准确  
反映了我们对现实世界的知识。如果概率，最佳动作将，因为我们期望正回报。对于值小于0.5，最佳动作将，因为它避免了与动作相关的负面预期收益。对应的价值功能的动作特定值功能最大：  
(16.11)  
该值函数和动作选择了相应的策略由图16.2d，从而最大程度地说明了两个线性成分的固体图所示

  
由虚线。我们注意到，这个数值函数不是线性的任何更长的时间。相反，它是分段线性和凸出的。非线性产生的事实，不同的操作是最佳的belief空间的不同部分。  
从图16.2d的价值功能，我们可以得出结论，belief 是没有办法获得的收益大于0？当然，答案是否定的。该  
值函数是仅最适合于时域。对于较大的视野，也能够先执动作作，接着动作。执动作作具有两个有益效果：首先，它有助于我们估计当前状态更好，由于这样的事实，在我们可以感知，并且第二，以高概率从至的，或反之亦然改变状态。因此，一个良好的策略可能会执行的动​​作，直到我们有理由相信，机器人的状态是。然后，机器人应执行的动作。  
为了使这个比较正规的，让我们得出的最优值函数的时域，假设我们执行的动作。然后，两件事情可能发生：我们要么遵守或。让我们先假设我们观察到。那么新的belief状态可以使用贝叶斯滤波器来计算：  
(16.12)  
其中，是贝叶斯法则的正规化。同样，我们得到的是后概率：

(16.13) since we know that —after all, when executing action the state will  
either be or—we obtain for :   
(16.14)  
我们也注意到，变量是一个有用的概率：这是执行的动作 (无论后状态的)后观察的概率。  
我们现在有表情刻画因为我们执行ac-后belief  
化观察。那么，什么是这个国家的belief价值？答案是转播tained通过插入新的belief到价值函数的公式定义(16.11)，和贴现的结果：

(16.15)  
我们现在移动的因素出来的最大化和移动0.9里面，并取得：  
(16.16)，其是一个分段线性的，凸的函数的belief b的参数。  
推导的执行动作后观察是完全类似。特别是，我们得到的后验  
*with the normalizer   
the corresponding value is*

顺便说一句，我们也注意到，。这直接遵循这样的事实  
在贝叶斯正规化过滤是观察概率  
和事实恰好有两种可能的观测我们的例子中，和。  
现在让我们计算值中，这是在执动作作的预期值。显然，该值是值和，通过实际观察和分别的概率加权的混合物。放入数学符号，我们有  
(16.21)  
术语已经如上所定义。概率执动作作之后观察是。因此，我们有所有成分计算所需值中：  
(16.22)  
该表达式可以被重新表示为四个线性函数的最大值：  
图16.2e表示这四个线性函数的belief间隔。该值函数是这四个线性函数的最大值。由于是容易被看到的那样，二者的这些功能是足以定义最大;其他2顷在整个频谱较小。这使我们能够改写的为最大值的仅有的两个方面，而不是四个：  
(16.24)  
最后，我们要确定的值中，这是下面两个条件中的最大值：  
(16.25)  
正如很容易地验证，则第二项的最大化，中，是完全与上述相同的用于规划时域t = 1：  
因此，我们获得(16.24)和(16.26)：  
(16.26)  
(16.27)这个函数是最优值函数规划时域T,和其组成部分的belief子空间。正如很容易看到的，函数是分段线性和凸出的。特别是，它包括三个  
线性件。对于在最左边的两个片belief(​​即)，是最佳的动作。对于，这两个动作都是一样的好。对应于最右边的线性片的belief，即用的belief，有最佳的动作。如果为最优函数，下一个动作取决于在belief空间的初始点和在观察。  
价值函数是仅最适合于时域2。但是，我们的分析示出了几个重要的观点。  
首先，最优值函数的任何有限的视距是连续的，分段线性，凸。每个线性一块对应于一个不同的操作选择在将来某个时候，或者，可以进行不同的观察。价值函数的凸性表示相当直观的观察，即明知是al-方面都优于不知道。给定两个belief状态b和b’,的belief状态的混合值大于或等于该混合belief状态的值，对于某些混合参数：  
(16.28)  
二，直线条的数量可以极大地增长，特别是如果一个人不注重线性函数变得过时。在我们的玩具例子中，两个出的四个直线约束限定的)的是没有必要。 有效实施POMDPs的“诀窍”在于早识别过时线性函数如可能，从而使计算的价值函数时，没有计算被浪费。不幸的是，即使我们仔细排除一切不必要的线性约束，线性函数的数量仍然可以增长得非常迅速。这对对有限POMDPs的精确值迭代解决方案固有的扩展限制。

**belief空间中的价值迭代**

在上一节展示来如何计算有限的价值函数。现在让我们回到在belief空间上价值迭代的一般问题。特别是，在介绍这一章中，我们指出了价值函数，我们简要地重申这里的基本更新方程：  
(16.29)  
在这个和下面的章节中，我们将开发一个可在数字计算机上实现的通用的在belief空间计算价值迭代的算法。首先，我们注意到，公式(16.2)提出对所有的belief进行积分，其中每一个belief是一个概率分布。如果状态空间是有限的，并且状态数为n，所有的概率分布的空间是具有维数n-1的连续体。我们注意到，为了指定在n个离散事件的概率分布n-1个数值是必需的(第n参数可以省略，因为概率加起来为1)。如果状态空间是连续的，belief空间具有无限多的维度。因此，对所有的belief的积分似乎是计算艰巨的任务。然而，我们可以通过重新制定的问题，并对所有观测进行积分来代替。  
让我们看看条件概率P(b’|a，b)，其中规定了在分配后的belief b’给定一个belief b和动作a。如果只有b和a是已知的，该后验belief b’不是唯一的，而且P(b’|a，b)是belief的一个真正的概率分布。但是，如果如果我们也知道测量o’，执行动作时，后验b’是独特的，P(b’|a，b)是一个point-mass分布的退化。从动作执行之前的belief b，动作a，和随后的观测o’，贝叶斯滤波器计算一个单一的准确的后验概率belief b’。因此，我们得出结论，只要我们知道o’，对所有的belief的积分会退化。  
这种洞察力可以通过重新表达p利用  
(16.30)

其中，p(b’|a，b，o’)是由贝叶斯滤波器计算的point-mass 分布将积分代入价值迭代的定义(16.2)，我们得到  
(16.31)  
内积分  
(16.32)  
仅包含一个非零项。这其中b’是由b，a和o的分布通过贝叶斯滤波器计算得到。让我们用b(b，a，o’)来表示这个分布，那么有  
(16.33)  
我们注意到，正规化分母，，是值更新方程(16.31)的一个因素。因此，将代入(16.31)消除了这一量，从而导致了值迭代算法的递归描述：  
(16.34)  
这种形式是比在(16.2)更加方便，因为它只需要对所有可能的测量o’，而不是所有可能的belief分布b’进行积分，这种转变是派生此一章所有值迭代算法的一个重要观点。事实上，它是含蓄地用在上面的例子中，在一个新的价值函数是由有限多个分段线性函数混合获得的。  
下面，很容易根据积分获得最大化价值的动作。因此，我们会注意到(16.34)可以改写为下列两个等式：  
(16.35)  
(16.36)  
在这里，是时域t对于假设下一个动作a的belief b的函数。

**计算价值函数**

在我们的例子中，最优值函数是分段线性和凸的。这种情况，对于一般的有限情况均满足。对于所有具有有限时域的有限POMDPs的最优值函数都是分段线性和凸的。分段线性度意味着价值函数由线性函数的集合表示。如果是线性的，它可以通过一组系数来表示  
(16.37)  
其中，像往常一样，是一种belief分布的参数。在我们的示例中，分段线性和凸值函数可以通过K个线性函数集合的最大值表示  
(16.38)  
其中，分别表示第k个线性函数的参数，以及k表示线性片段的数量。一组有限的线性函数的最大值事实上是凸的分段线性的函数。

在值迭代，初始值函数由下式给出  
(16.39)  
我们注意到，这个数值函数是线性的，因此，它也是分段线性和凸。这种分配建立了递归值迭代算法的基础。  
我们现在衍生除一个计算值函数的递归方程。该公式假定上一步的值函数，由一个分段线性函数如上规定表示。作为派生的一部分，我们将假设是分段线性和凸的，是也分段线性和凸出的。下面将证明具有有限时域的值函数是分段线性和凸的。  
方程(16.35)和(16.36)定义如下：  
(16.40)  
(16.41)  
在有限的空间中，所有的积分可以通过有限和取代，我们得到：  
(16.43)  
belief 使用下面的表达式获得，该式为用和来替换(16.33)中的积分所得到：

(16.44)

如果belief b由参数表示，那么belief b’中的第i个参数由下式计算得到  
(16.45)

为了计算价值函数，我们将现在的条件和中衍生出更多实用的表达式中)，从公式(16.42)开始，具有参数，我们从得到了  
(16.46)  
代入(16.45)，我们得到  
(16.47)  
由于不依赖于i，后者变换是合法的。这种形式仍然包含一个难以计算的表达式：。然而，在有限的情况下这个表达式可以被消掉。因此，它不必被任何进一步考虑。  
在(16.42)中对量的推导类似于上面。特别是，我们不得不用唯一的分段线性和凸的值函数C^(T-1)取代直接回报函数c。然后可以得到  
(16.48)  
。如上所述，我们代入后验概率belief b(见(16.45))，并获得：  
(16.49)  
其中 不依赖于i或k，故可移出求和和最大化部分，得到：  
(16.50)  
该表达式也是分段线性和凸的可能并不明显。然而，对该式进行整理得到下式：  
(16.51)

在这种形式下，很容易验证，对于任何的固定值k，使其最大化的参数是确实线性的参数。因而，这些线性段的最大值为分段线性和凸的。此外，线性段的数量是有限的。  
现在让我们回到本节中的主要问题，即计算和时域T的值函数。我们简要地重申方程(16.42)：  
16.52)

让我们首先用公式(16.47)和(16.50)计算总和

(16.53)   
经过一些整理得到，  
(16.54)

将这个表达式代入公式16.52  
(16.55)  
我们注意到，项同时出现在在此表达式的分子和分母中从而抵消了，得到：  
(16.56)

这个消除是仅对有限POMDP满足的。所需的价值函数通过对于所有动作a最大化  
得到  
(16.57)  
很容易验证，在belief参数为的情况是确实分段线性和凸的。具体地，对于每个固定k(16.56)的大括​​号内的量是线性的。这些k个线性函数的最大值是一个分段线性函数，包括最多有k段。分段线性的凸函数的总和为再次分段线性和凸的。因此，因此对于所有观测ø？的乘积的总和再次产生一个分段线性凸函数。最后，对于在方程(16.57)中的所有动作的最大值也是分段线性的凸函数。因此，我们的结论是在任何有限时域T，是分段线性和凸出的。