CS229 Lecture notes

原作者：[Andrew Ng](http://cs229.stanford.edu/)（[吴恩达](http://open.163.com/movie/2008/1/M/C/M6SGF6VB4_M6SGHFBMC.html)）

翻译：[CycleUser](https://zhuanlan.zhihu.com/python-kivy)

Part XIII

# 强化学习（Reinforcement Learning）和控制（Control）

这一章我们就要学习强化学习（reinforcement learning）和适应性控制（adaptive control）了。在监督学习（supervised learning）中，我们已经见过的一些算法，输出的标签类 y 都是在训练集中已经存在的。这种情况下，对于每个输入特征 x，都有一个对应的标签作为明确的“正确答案（right answer）”。与之相反，在很多的连续判断（sequential decisions making）和控制（control）的问题中，很难提供这样的明确的显示监督（explicit supervision）给学习算法。例如，假设咱们制作了一个四条腿的机器人，然后要编程让它能走路，而我们并不知道怎么去采取“正确”的动作来进行四条腿的行走，所以就不能给他提供一个明确的监督学习算法来进行模仿。在强化学习（reinforcement learning）的框架下，我们就并不提供监督学习中那种具体的动作方法，而是只给出一个奖励函数（reward function），这个函数会告知学习程序（learning agent） 什么时候的动作是好的，什么时候的是不好的。在四腿机器人这个样例中，奖励函数会在机器人有进步的时候给出正面回馈，即奖励，而有退步或者摔倒的时候给出负面回馈，可以理解成惩罚。接下来随着时间的退役，学习算法就会解决如何选择正确动作以得到最大奖励。

强化学习（Reinforcement learning，下文中缩写为 RL）已经成功用于多种场景了，例如无人直升机的自主飞行，机器人用腿来运动，手机的网络选择，市场营销策略筛选，工厂控制，高效率的网页索引等等。我们对强化学习的探索，要先从马尔可夫决策过程（Markov decision processes，缩写为 MDP）开始，这个概念给出了强化学习问题的常见形式。

# 1 马尔可夫决策过程（Markov decision processes）

一个马尔可夫决策过程（Markov decision process）由一个元组（tuple）： (S, A, {Psa}, γ, R)组成，其中元素分别为：

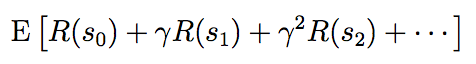
1. S 是一个状态集合（a set of states）。（例如，在无人直升机飞行的案例中，S 就可以是直升机所有的位置和方向的集合。）
2. A 是一个动作结合（a set of actions）。（例如，还以无人直升机为例，A 就可以是遥控器上面能够操作的所有动作方向。）
3. Psa 为状态转移概率（state transition probabilities）。对于每个状态 s ∈ S 和动作 a ∈ A， Psa 是在状态空间上的一个分布（a distribution over the state space）。后面会再详细讲解，不过简单来说， Psa 给出的是在状态 s 下进行一个动作 a 而要转移到的状态的分布。
4. γ ∈ [0, 1] 叫做折扣因子（discount factor）。
5. R : S × A → R 就是奖励函数（reward function）。（奖励函数也可以写成仅对状态 S 的函数，这样就可以写成 R : S → R。）

马尔可夫决策过程（MDP）的动力学（dynamics）过程如下所示：于某个起始状态 s0 启动，然后选择某个动作 a0 ∈ A 来执行 MDP 过程。根据所选的动作会有对应的结果，MDP 的状态则转移到某个后继状态（successor state），表示为 s1，根据 s1 ∼ Ps0a0 得到。然后再选择另外一个动作 a1，接下来又有对应这个动作的状态转移，状态则为 s2 ∼ Ps1a1。接下来在选择一个动作 a2，就这样进行下去。可以用下面的过程来作为表示：



通过序列中的所有状态 s0, s1, . . . 和对应的动作 a0, a1, . . .，就你能得到给出的总奖励值，即总收益函数（total payoff）为 R(s0,a0) + γR(s1,a1) + γ2R(s2,a2) + ···。如果把奖励函数只作为仅与状态相关的函数，那么这个值就简化成了 R(s0) + γR(s1) + γ2R(s2) + ···。多数情况下，我们都用后面这种仅为状态的函数这种形式，虽然扩展到对应状态-动作两个变量的函数 R(s,a) 也并不难。

强化学习的目标就是找到的一组动作，能使得总收益函数（total payoff）的期望值最大：



注意，在时间步长（timestep） t 上的奖励函数（reward）通过一个参数（factor）γt而进行了缩减（discounted）。因此，要使得期望最大化，就需要尽可能早积累符号为正的奖励（positive rewards），而尽量推迟负面奖励（negative rewards，即惩罚）的出现。在经济方面的应用中，其中的 R(·) 就是盈利金额（amount of money made），γ 也可以理解为利润率（interest rate）的表征，这样有自然的解释（natural interpretation），例如今天的一美元就比明天的一美元有更多价值。

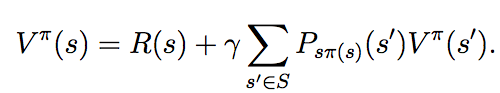
有一种策略（policy），是使用任意函数 π : S → A，从状态（states）到动作（actions）进行映射（mapping）。如果在状态 s，采取动作 a = π(s)，就可以说正在执行（executing）某种策略（policy） π。然后还可以针对策略函数（policy）π 来定义一个值函数（value function）：



Vπ(s) 就是从状态 s 开始，根据 π 给出的动作来积累的部分奖励函数（discounted rewards）的期望总和（expected sum）。1

1 实际上这里我们用 π 这个记号来表示，严格来说不太正确，因为 π 并不是一个随机变量，不过在文献里面这样表示很多，已经成了某种事实上的标准了。

给定一个固定的策略函数（policy） π，则对应的值函数 V π 满足贝尔曼等式（Bellman equations）：



This says that the expected sum of discounted rewards Vπ(s) for starting in s consists of two terms: First, the immediate reward R(s) that we get rightaway simply for starting in state s, and second, the expected sum of future discounted rewards. Examining the second term in more detail, we see that the summation term above can be rewritten Es′∼Psπ(s) [V π(s′)]. This is the expected sum of discounted rewards for starting in state s′, where s′ is distributed according Psπ(s), which is the distribution over where we will end up after taking the first action π(s) in the MDP from state s. Thus, the second term above gives the expected sum of discounted rewards obtained after the first step in the MDP.

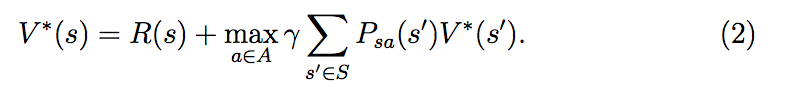
这也就意味着，从状态 s 开始的这个部分奖励（discounted rewards）的期望总和（expected sum） Vπ(s) 由两部分组成：首先是在状态 s 时候当时立即获得的奖励函数值 R(s)，也就是上面式子的第一项；另一个就是第二项，即后续的部分奖励函数值（discounted rewards）的期望总和（expected sum）。对第二项进行更深入的探索，就能发现这个求和项（summation term）可以写成 Es′∼Psπ(s) [V π(s′)] 的形式。这种形式也就是从状态 s′ 开始的这个部分奖励（discounted rewards）的期望总和（expected sum） Vπ(s′)，此处的 s′ 是根据 Psπ(s) 分布的，在 MDP 过程中从状态 s 采取第一个动作 π(s) 之后，确定了这个分布所在的空间。因此，上面的第二项实际上也就是给出了在 MDP 过程中第一步之后的部分奖励（discounted rewards）的期望总和（expected sum）。

贝尔曼等式（Bellman’s equations）可以有效地解出 Vπ。尤其是在一个有限状态的 MDP 过程中，即 (|S| < ∞)，我们可以把每个状态 s 对应的 V π (s) 的方程写出来。这样就得到了一系列的 |S | 个线性方程，有 |S | 个变量（也就是对应每个状态的未知的 Vπ(s) ），这些 Vπ(s) 都很容易解出来。

然后可以定义出**最优值函数（optimal value function）**



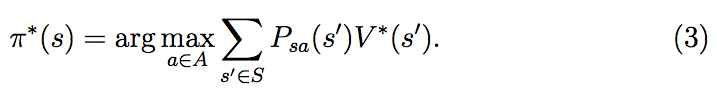
换一种说法，这个值也就是能用任意一种策略函数（policy）来获得的，最佳的可能部分奖励（discounted rewards）的期望总和（expected sum）。另外对于最优值函数（optimal value function），也有一个版本的贝尔曼等式（Bellman’s equations）：



上面这个等式中的第一项，还是跟之前一样的，还是即时奖励函数值。第二项是在采取了动作 a 之后的所有动作 a 的部分奖励（discounted rewards）的未来期望总和（expected future sum）的最大值。要确保理解这个等式，并且要明白为什么这个等式有意义。

（译者注：抱歉，这里的这个 discounted rewards 弄得我不知道怎么翻译才顺，意思表达得很狗，非常抱歉。）

另外还定义了一个策略函数（policy） π∗ : S → A，如下所示



注意，这里的 π∗(s) 给出的动作 a 给出的在上面等式（2）当中能够使 “max” 项取最大值。对于每个状态 s 和每个策略函数（policy）π，都有：



上面的第一个等式关系表明，对应策略函数（policy） π∗的值函数（value function）V π∗ 等于对于每个状态 s 的最优值函数 V ∗。右边的不等式则表明，π∗ 的值至少也等于任意其他策略函数的值。也就是说，上面在等式（3）当中定义的这个 π∗ 就是最佳策略函数（optimal policy）。

注意，这个 π∗ 有一个有趣的特性，它是所有状态 s 下的最佳策略。具体来讲，并不是说只是如果从某个状态 s 开始 MDP 过程，这个 π∗ 是对应这个状态的最佳策略，而如果从某个别的状态 s′ 开始就有其他的最佳策略。而是对于所有的状态 s，都是同样的一个策略函数 π∗ 能够使得等式（1）中的项目取得最大值。这也就意味着无论 MDP 过程的初始状态（initial state）如何，都可以使用同样的策略函数 π∗。

# 2 值迭代（Value iteration）和策略迭代（policy iteration）

现在我们要讲两种算法，都能很有效地解决有限状态的马尔可夫决策过程问题（finite-state MDPs）。目前为止，我们只考虑有限状态和动作空间的马尔可夫决策过程，也就是状态和动作的个数都是有限的，即|S| < ∞, |A| < ∞。

第一种算法，值迭代（value iteration），过程如下所述：

1. 对每个状态 s, 初始化 V (s) := 0.

2. 重复直到收敛 {

对每个状态，更新规则 

}

这个算法可以理解成，利用贝尔曼等式（Bellman Equations）（2）重复更新估计值函数（estimated value function）。

在上面的算法的内部循环体中，有两种进行更新的方法。首先，我们可以为每一个状态 s 计算新的值 V (s)，然后用新的值覆盖掉所有的旧值。这也叫做同步更新（synchronous update）。在这种情况下，此算法可以看做是实现（implementing）了一个“贝尔曼备份运算符（Bellman backup operator）”，这个运算符接收值函数（value function）的当前估计（current estimate），然后映射到一个新的估计值（estimate）。（更多细节参考作业题目中的内容。）

另外一种方法，就可以使用异步更新（asynchronous updates）。使用这种方法，就可以按照某种次序来遍历（loop over）所有的状态，然后每次更新其中一个的值。

无论是同步还是异步的更新，都能发现最终值迭代（value iteration）会使 V 收敛到 V ∗ 。找到了 V ∗ 之后，就可以利用等式（3）来找到最佳策略（optimal policy）。

除了值迭代（value iteration）之外，还有另外一种标准算法可以用来在马尔可夫决策过程（MDP）中寻找一个最佳策略（optimal policy）。这个策略循环（policy iteration）算法如下所述：

1. 随机初始化 π。

2. 重复直到收敛{

(a) 令 V := V π.

(b) 对每个状态 s，令 

}

因此，在循环体内部就重复计算对于当前策略（current policy）的值函数（value function），然后使用当前的值函数（value function）来更新策略函数（policy）。（在步骤 b 中找到的策略 π 也被称为对应 V 的贪心策略（greedy with respect to V）。）注意，步骤 a 可以通过解贝尔曼等式（Bellman’s equation）来实现，之前已经说过了，在固定策略（fixed policy）的情况下，这个等式只是一系列有 |S| 个变量（variables）的 |S| 个线性方程（linear equations）。

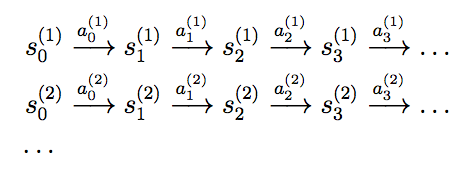
在上面的算法迭代了某个最大迭代次数之后，V 将会收敛到 V ∗，而 π 会收敛到 π∗。

值迭代（value iteration）和策略迭代（policy iteration）都是解决马尔可夫决策过程（MDPs）问题的标准算法， 而且目前对于这两个算法哪个更好，还没有一个统一的一致意见。对小规模的 MDPs 来说，策略迭代（policy iteration）通常非常快，迭代很少的次数就能瘦脸。然而，对有大规模状态空间的 MDPs，确切求解 V π就要涉及到求解一个非常大的线性方程组系统，可能非常困难。对于这种问题，就可以更倾向于选择值迭代（value iteration）。因此，在实际使用中，值迭代（value iteration）通常比策略迭代（policy iteration）更加常用。

# 3 学习一个马尔可夫决策过程模型（Learning a model for an MDP）

目前为止，我们已经讲了 MDPs，以及用于 MDPs 的一些算法，这都是基于一个假设，即状态转移概率（state transition probabilities）以及奖励函数（rewards）都是已知的。在很多现实问题中，却未必知道这两样，而是必须从数据中对其进行估计。（通常 S，A 和 γ 都是知道的。）

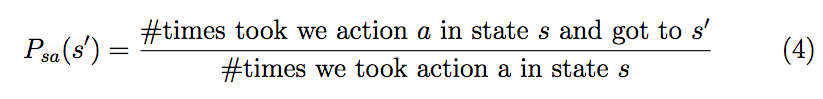
例如，加入对倒立摆问题（inverted pendulum problem，参考习题集 4），在 MDP 中进行了一系列的试验，过程如下所示：



Here, si(j) is the state we were at time i of trial j, and ai (j) is the corresponding action that was taken from that state. In practice, each of the trials above might be run until the MDP terminates (such as if the pole falls over in the inverted pendulum problem), or it might be run for some large but finite number of timesteps.

其中 si(j) 表示的是第 j 次试验中第 i 次的状态，而 ai (j) 是该状态下的对应动作。在实践中，每个试验都会运行到 MDP 过程停止（例如在倒立摆问题（inverted pendulum problem）中杆落下（pole falls）），或者会运行到某个大但有限个数的时间步长（timesteps）。

有了在 MDP 中一系列试验得到的“经验”，就可以对状态转移概率（state transition probabilities）推导出最大似然估计（maximum likelihood estimates）了：



Psa(s’)= (在状态 s 执行动作 a 而到达状态 s’ 花的时间)/(在状态 s 执行动作 a 花的时间) (4)

或者，如果上面这个比例出现了“0/0”的情况，对应的情况就是在状态 s 之前没进行过任何动作 a，这样就可以简单估计 Psa(s′) 为 1/|S|。（也就是说把 Psa 估计为在所有状态上的均匀分布（uniform distribution）。）

注意，如果在 MDP 过程中我们能获得更多经验信息（观察更多次数），就能利用新经验来更新估计的状态转移概率（estimated state transition probabilities），这样很有效率。具体来说，如果我们保存下来等式（4）中的分子（numerator）和分母（denominator）的计数（counts），那么观察到更多的试验的时候，就可以很简单地累积（accumulating）这些计数数值。计算这些数值的比例，就能够给出对 Psa 的估计。

利用类似的程序（procedure），如果奖励函数（reward） R 未知，我们也可以选择在状态 s 下的期望即时奖励函数（expected immediate reward） R(s) 来当做是在状态 s 观测到的平均奖励函数（average reward）。

学习了一个 MDP 模型之后，我们可以使用值迭代（value iteration）或者策略迭代（policy iteration），利用估计的状态转移概率（transition probabilities）和奖励函数，来去求解这个 MDP 问题。例如，结合模型学习（model learning）和值迭代（value iteration），就可以在未知状态转移概率（state transition probabilities）的情况下对 MDP 进行学习，下面就是一种可行的算法：

1. 随机初始化 π 。

2. 重复 {

(a) 在 MDP 中执行 π 作为若干次试验（trials）。

(b) 利用上面在 MDP 积累的经验（accumulated experience），更新对 Psa 的估计（如果可以的话也对奖励函数 R 进行更新）。

(c) 利用估计的状态转移概率（estimated state transition probabilities）和奖励函数（rewards），应用值迭代（value iteration），得到一个新的估计值函数（estimated value function） V。

(d) 更新 π 为与 V 对应的贪婪策略（greedy policy）。

}

我们注意到，对于这个特定的算法，有一种简单的优化方法（optimization），可以让该算法运行得更快。具体来说，在上面算法的内部循环中，使用了值迭代（value iteration），如果初始化迭代的时候不令 V = 0 启动，而是使用算法中上一次迭代找到的解来初始化，这样就有了一个更好的迭代起点，能让算法更快收敛。

# 4 连续状态的马尔可夫决策过程（Continuous state MDPs）

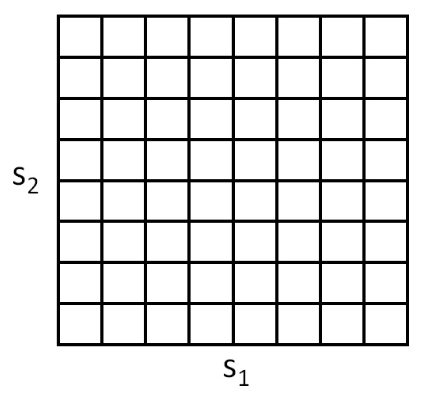
目前为止，我们关注的都是有限个状态（a finite number of states）的马尔可夫决策过程（MDPs）。接下来我们要讲的就是有无限个状态（an infinite number of states）的情况下的算法。例如，对于一辆车，我们可以将其状态表示为 (x, y, θ, x ̇,y ̇,θ ̇)，其中包括位置（position） (x, y)，方向（orientation）θ， 在 x 和 y 方向的速度分量 x ̇ 和 y ̇，以及角速度（angular velocity）θ ̇。这样，S = R6 就是一个有无限的状态集合，因为一辆车的位置和方向的个数是有无限可能的2。与此相似，在习题集4 中看到的倒立摆问题（inverted pendulum）中，状态也有，其中的 θ 是杆的角度。在直升机飞行的三维空间中，状态的形式则为，其中包含了滚动角（roll）φ，俯仰角（pitch）θ，以及偏航角（yaw）ψ，这几个角度确定了直升机在三维空间中的运动方向。在本节中，我们考虑状态空间为 S = Rn 的情况，并描述此种情况下解决 MDPs 的方法。

2从理论上讲，θ 是一个方向（orientation），所以更应当把 θ 的取值空间写为 θ ∈ [−π, π]，而不是写为实数集合 θ ∈ R；不过在我们讨论的问题中，这种区别不要紧。

## 4.1 离散化（Discretization）

解决连续状态 MDP 问题最简单的方法可能就是将状态空间（state space）离散化（discretize），然后再使用之前讲过的算法，比如值迭代（value iteration）或者策略迭代（policy iteration）来求解。

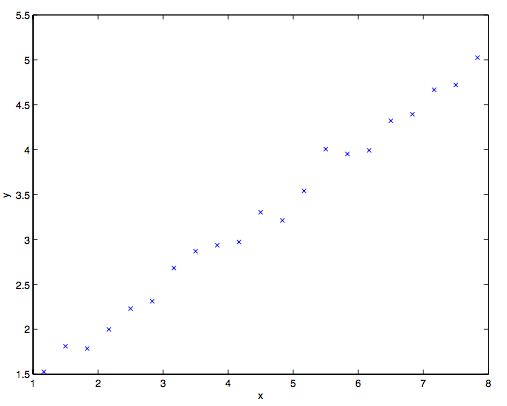
例如，假设我们有一个二维状态空间（s1，s2），就可以用下面的网格（grid）来将这个状态空间离散化：



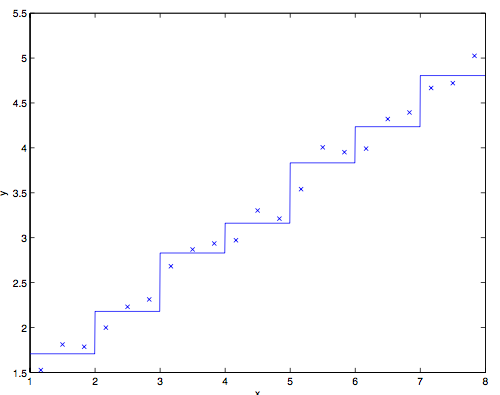
如上图所示，每个网格单元（grid cell）表示的都是一个独立的离散状态 s̄。这样就可以把一个连续状态 MDP 用一个离散状态的 (S ̄, A, {Ps ̄a}, γ, R) 来进行逼近，其中的S ̄ 是离散状态集合，而{Ps ̄a} 是此离散状态上的状态转移概率（state transition probabilities），其他项目同理。然后就可以使用值迭代（value iteration）或者策略迭代（policy iteration）来求解出离散状态的 MDP (S ̄, A, {Ps ̄a}, γ, R) 的 V ∗(s ̄) 和 π∗(s ̄)。当真实系统是某种连续值的状态 s ∈ S，而有需要选择某个动作来执行，就可以计算对应的离散化的状态 s ̄，然后执行对应的动作 π∗(s ̄)。

这种离散化方法（discretization approach）可以解决很多问题。然而，也有两个缺陷（downsides）。首先，这种方法使用了对 V ∗ 和 π∗ 相当粗糙的表征方法。具体来说，这种方法中假设了在每个离散间隔（discretization intervals）中的值函数（value function）都去一个常数值（也就是说，值函数是在每个网格单元中分段的常数。）。

要更好理解这样表征的的局限性，可以考虑对下面这一数据集进行函数拟合的监督学习问题：



很明显，上面这个数据适合使用线性回归。然而，如果我们对 x 轴进行离散化，那么在每个离散间隔中使用分段常数表示，对同样的数据进行拟合，得到的曲线则如下所示：



This piecewise constant representation just isn’t a good representation for many smooth functions. It results in little smoothing over the inputs, and no generalization over the different grid cells. Using this sort of representation, we would also need a very fine discretization (very small grid cells) to get a good approximation.

这种分段常数表示，对于很多的光滑函数，都不能算好。这会导致输入值缺乏平滑（little smoothing over the inputs），而且在不同的望各单元中间也没有进行扩展（generalization）。使用这种表示方法，我们还需要一种非常精细的离散化过程（也就是网格单元要非常小），才能得到一个比较好的近似估计。

第二个缺陷可以称之为维度的诅咒（curse of dimensionality）。设 S = Rn ，然后我们队每个 n 维度状态离散成 k 个值。这样总共的离散状态的个数就是 kn。在状态空间 n 的维度中，这个值会呈指数级增长，对于大规模问题就不好缩放了。例如，对于一个 10 维的状态，如果我们把每个状态变量离散化成为 100 个值，那么就会有 10010 = 1020 个离散状态，这个维度太大了，远远超过了当前桌面电脑能应付的能力之外。

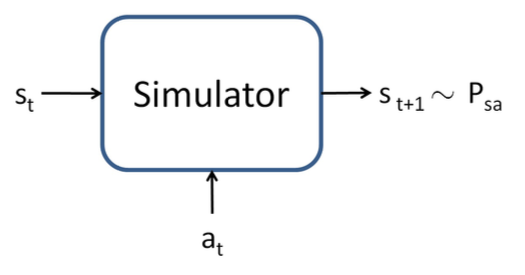
根据经验法则（rule of thumb），离散化通常非常适合用于 1 维和 2 维的问题（而且有着简单和易于快速实现的优势）。对于 4 维状态的问题，如果使用一点小聪明，仔细挑选离散化方法，有时候效果也不错。如果你超级聪明，并且还得有点幸运，甚至也有可能将离散化方法使用于 6维问题。不过在更高维度的问题中，就更是极其难以使用这种方法了。

## 4.2 值函数近似（Value function approximation）

现在我们来讲另外一种方法，能用于在连续状态的 MDPs 问题中找出策略，这种方法也就是直接对进行近似 V ∗，而不使用离散化。这个方法就叫做值函数近似（value function approximation），在很多强化学习的问题中都有成功的应用。

### 4.2.1 使用一个模型或模拟器（Using a model or simulator）

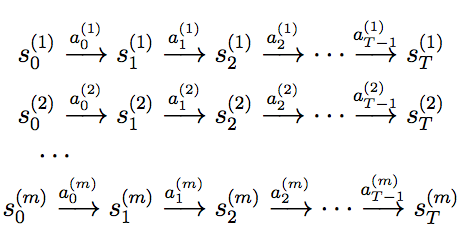
要开发一个值函数近似算法，我们要假设已经有一个对于 MDP 的模型，或者模拟器。简单来看，一个模拟器就是一个黑箱子（black-box），接收输入的任意（连续值的）状态 st 和动作 at，然后输出下一个状态 st+1，这个新状态是根据状态转移概率（state transition probabilities） Pstat 取样（sampled）得来：



There’re several ways that one can get such a model. One is to use physics simulation. For example, the simulator for the inverted pendulum in PS4 was obtained by using the laws of physics to calculate what position and orientation the cart/pole will be in at time t + 1, given the current state at time t and the action a taken, assuming that we know all the parameters of the system such as the length of the pole, the mass of the pole, and so on. Alternatively, one can also use an off-the-shelf physics simulation software package which takes as input a complete physical description of a mechanical system, the current state st and action at, and computes the state st+1 of the system a small fraction of a second into the future.3

3Open Dynamics Engine (http://www.ode.com) is one example of a free/open-source physics simulator that can be used to simulate systems like the inverted pendulum, and that has been a reasonably popular choice among RL researchers.

An alternative way to get a model is to learn one from data collected in the MDP. For example, suppose we execute m trials in which we repeatedly take actions in an MDP, each trial for T timesteps. This can be done picking actions at random, executing some specific policy, or via some other way of choosing actions. We would then observe m state sequences like the following:

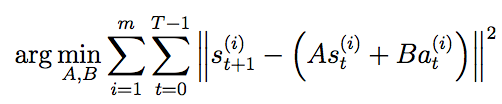


We can then apply a learning algorithm to predict st+1 as a function of st and at.

For example, one may choose to learn a linear model of the form



using an algorithm similar to linear regression. Here, the parameters of the model are the matrices A and B, and we can estimate them using the data collected from our m trials, by picking



(This corresponds to the maximum likelihood estimate of the parameters.) Having learned A and B, one option is to build a deterministic model, in which given an input st and at, the output st+1 is exactly determined. Specifically, we always compute st+1 according to Equation (5). Alternatively, we may also build a stochastic model, in which st+1 is a random function of the inputs, by modelling it as



where here εt is a noise term, usually modeled as εt ∼ N (0, Σ). (The covariance matrix Σ can also be estimated from data in a straightforward way.)

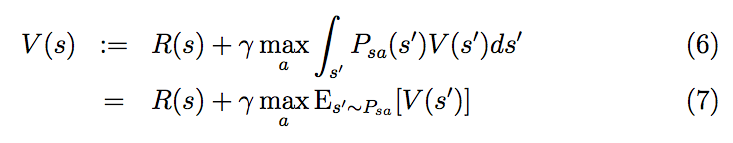
Here, we’ve written the next-state st+1 as a linear function of the current state and action; but of course, non-linear functions are also possible. Specifically, one can learn a model st+1 = Aφs(st) + Bφa(at), where φs and φa are some non-linear feature mappings of the states and actions. Alternatively, one can also use non-linear learning algorithms, such as locally weighted linear regression, to learn to estimate st+1 as a function of st and at. These approaches can also be used to build either deterministic or stochastic simulators of an MDP.

### 4.2.2 Fitted value iteration

We now describe the fitted value iteration algorithm for approximating the value function of a continuous state MDP. In the sequel, we will assume that the problem has a continuous state space S = Rn, but that the action space A is small and discrete.4

4In practice, most MDPs have much smaller action spaces than state spaces. E.g., a car has a 6d state space, and a 2d action space (steering and velocity controls); the inverted pendulum has a 4d state space, and a 1d action space; a helicopter has a 12d state space, and a 4d action space. So, discretizing ths set of actions is usually less of a problem than discretizing the state space would have been.

Recall that in value iteration, we would like to perform the update



(In Section 2, we had written the value iteration update with a summation rather than an integral over states; the new notation reflects that we are now working in continuous states rather than discrete states.) The main idea of fitted value iteration is that we are going to approximately carry out this step, over a finite sample of states s(1), . . . , s(m). Specifically, we will use a supervised learning algorithm—linear regression in our description below—to approximate the value function as a linear or non-linear function of the states:



Here, φ is some appropriate feature mapping of the states. For each state s in our finite sample of m states, fitted value iteration will first compute a quantity y(i), which will be our approximation to R(s)+γmaxaEs′∼Psa[V(s′)](the right hand side of Equation7). Then, it will apply a supervised learning algorithm to try to get V (s) close to R(s) + γ maxa Es′∼Psa [V (s′)] (or, in other words, to try to get V (s) close to

y(i)). In detail, the algorithm is as follows:

1. Randomly sample m states s(1), s(2), . . . s(m) ∈ S. 2. Initialize θ := 0.

3. Repeat {

For i = 1, . . . , m {

For each action a ∈ A {

Sample s′1, . . . , s′k ∼ Ps(i)a (using a model of the MDP).

Set 

// Hence, q(a) is an estimate of 

Set y(i) = maxa q(a).

// Hence, y(i) is an estimate of.

}

// In the original value iteration algorithm (over discrete states)

// we updated the value function according to V (s(i)) = y(i).

// In this algorithm, we want V (s(i)) ≈ y(i), which we’ll achieve

// using supervised learning (linear regression).

Set 

}

Above, we had written out fitted value iteration using linear regression as the algorithm to try to make V (s(i)) close to y(i). That step of the algorithm is completely analogous to a standard supervised learning (regression) problem in which we have a training set (x(1),y(1)),(x(2),y(2)),...,(x(m),y(m)), and want to learn a function mapping from x to y; the only difference is that here s plays the role of x. Even though our description above used linear regression, clearly other regression algorithms (such as locally weighted linear regression) can also be used.

Unlike value iteration over a discrete set of states, fitted value iteration cannot be proved to always to converge. However, in practice, it often does converge (or approximately converge), and works well for many problems. Note also that if we are using a deterministic simulator/model of the MDP, then fitted value iteration can be simplified by setting k = 1 in the algorithm. This is because the expectation in Equation (7) becomes an expectation over a deterministic distribution, and so a single example is sufficient to exactly compute that expectation. Otherwise, in the algorithm above, we had to draw k samples, and average to try to approximate that expectation (see the definition of q(a), in the algorithm pseudo-code).

Finally, fitted value iteration outputs V , which is an approximation to V∗. This implicitly defines our policy. Specifically, when our system is in some state s, and we need to choose an action, we would like to choose the action

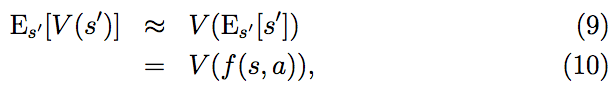


The process for computing/approximating this is similar to the inner-loop of fitted value iteration, where for each action, we sample s′1,...,s′k ∼ Psa to approximate the expectation. (And again, if the simulator is deterministic, we can set k = 1.)

In practice, there’re often other ways to approximate this step as well. For example, one very common case is if the simulator is of the form st+1 = f(st,at) + εt, where f is some determinstic function of the states (such as f(st,at) = Ast + Bat), and ε is zero-mean Gaussian noise. In this case, we can pick the action given by



In other words, here we are just setting εt = 0 (i.e., ignoring the noise in the simulator), and setting k = 1. Equivalently, this can be derived from Equation (8) using the approximation



where here the exception is over the random s′ ∼ Psa. So long as the noise terms εt are small, this will usually be a reasonable approximation.

However, for problems that don’t lend themselves to such approximations, having to sample k|A| states using the model, in order to approximate the expectation above, can be computationally expensive.