

强化学习基础篇（四）动态规划之迭代策略评估

1、迭代策略评估（Iterative Policy Evaluation）

在环境模型已知的前提下，对于任意的策略 π ，需要合理估算该策略带来的累积奖励期望以及准确衡量该策略的优劣程度，而策略评估（Policy Evaluation）可以实现这两个目标。

回顾一下策略 π 的具体定义：策略 π 是根据环境反馈的当前状态，决定智能体采取何种行动的指导方法。策略评估通过计算与策略对应的状态值函数 $v(s)$ ，以评估该策略的优劣。即给定一个策略，计算基于该策略下的每个状态的状态值 $v(s)$ 的期望，并用该策略下的最终状态值的期望来评价该策略。

策略评估通过迭代计算贝尔曼期望方程，已获得对应的状态值函数 $v(s)$ ，进而利用该状态值函数评估该策略是否最优，

2、迭代策略评估的过程

问题定义：

评估一个给定策略 π ，求对应的值函数 $v_\pi(s)$ 或者 $q_\pi(s, a)$ ，即解决预测（Prediction）问题。

解决方案：

- 直接求解贝尔曼方程，可以参考《强化学习基础篇（二）马尔科夫决策过程（MDP）》中MDP下贝尔曼方程的矩阵形式。可以在时间复杂度为 $O(n^3)$ 的情况下求得精确解。

$$v_\pi = (1 - \gamma P^\pi)^{-1} R^\pi$$

- 迭代解：迭代地应用Bellman期望方程进行求解， $v_1 \rightarrow v_2 \rightarrow \dots \rightarrow v_\pi$ 。

$$v(s) = \mathbb{E}[R_{t+1} + \gamma v(S_{t+1}) | S_t = s]$$

具体方法-同步反向迭代（synchronous backups）：

即在每次迭代过程中，对于第 $k+1$ 次迭代，所有状态 s 的价值用 $v_k(s')$ 计算并更新该状态第 $k+1$ 次迭代中使用的价值 $v_k(s)$ ，其中 s' 是 s 的后继状态。

同步(synchronous)的含义是每次更新都要更新完所有的状态；

备份(backup)，即 $v_{k+1}(s)$ 需要用到 $v_k(s')$ ，用 $v_k(s')$ 更新 $v_{k+1}(s)$ 的过程称为备份，更新状态 s 的值函数称为备份状态 s 。

使用数学描述这个过程为：

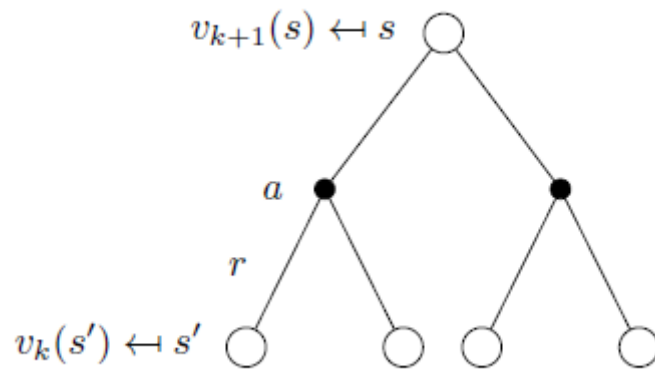
$$v_\pi(s) = \sum_{a \in A} \pi(a|s) (R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s'))$$

$$v_{k+1} = \sum_{a \in A} \pi(a|s) (R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_k(s')) \quad (v_k \text{ 为第 } k \text{ 次迭代得的函数})$$

$$v^{k+1} = R^\pi + \gamma P^\pi v^k$$

3、同步备份下的迭代式策略评价算法

一次迭代内，状态 s 的价值等于前一次迭代该状态的即时奖励与下 s 一个所有可能状态 s' 的价值与其概率乘积的和，如图示：



同步备份下的迭代式策略评价算法的伪代码如下：

算法 1 同步备份下的迭代式策略评价算法

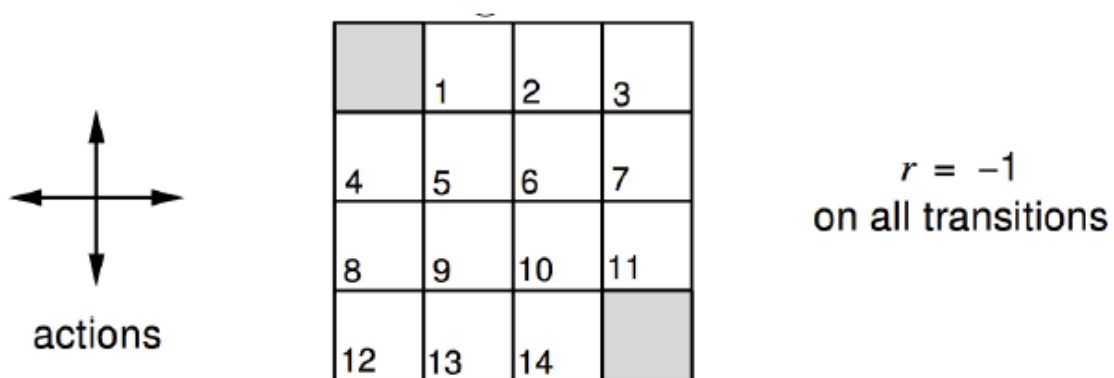
```

1: for  $k = 1, 2, \dots$  do
2:   for 所有的状态  $s \in \mathcal{S}$  do
3:     使用迭代式更新值函数  $v_{k+1}(s)$ 
4:   end for
5: end for

```

4、迭代策略评估在方格问题（Gridworld）中的示例

4.1、Gridworld描述：



已知条件为：

状态空间 S ：如图， $S_1 - S_{14}$ 为非终止状态， S_T 为终止状态（灰色方格所示的两个位置）

动作空间 A ：对于任何非终止状态可以有东南西北移动的四個动作。

转移概率 P ：任何试图离开方格世界的动作其位置将不会发生改变，其余条件下将100%地转移到动作指向的状态。

即时奖励 R ：任何在非终止状态间的转移得到的即时奖励均为-1.0，进入终止状态即时奖励为0

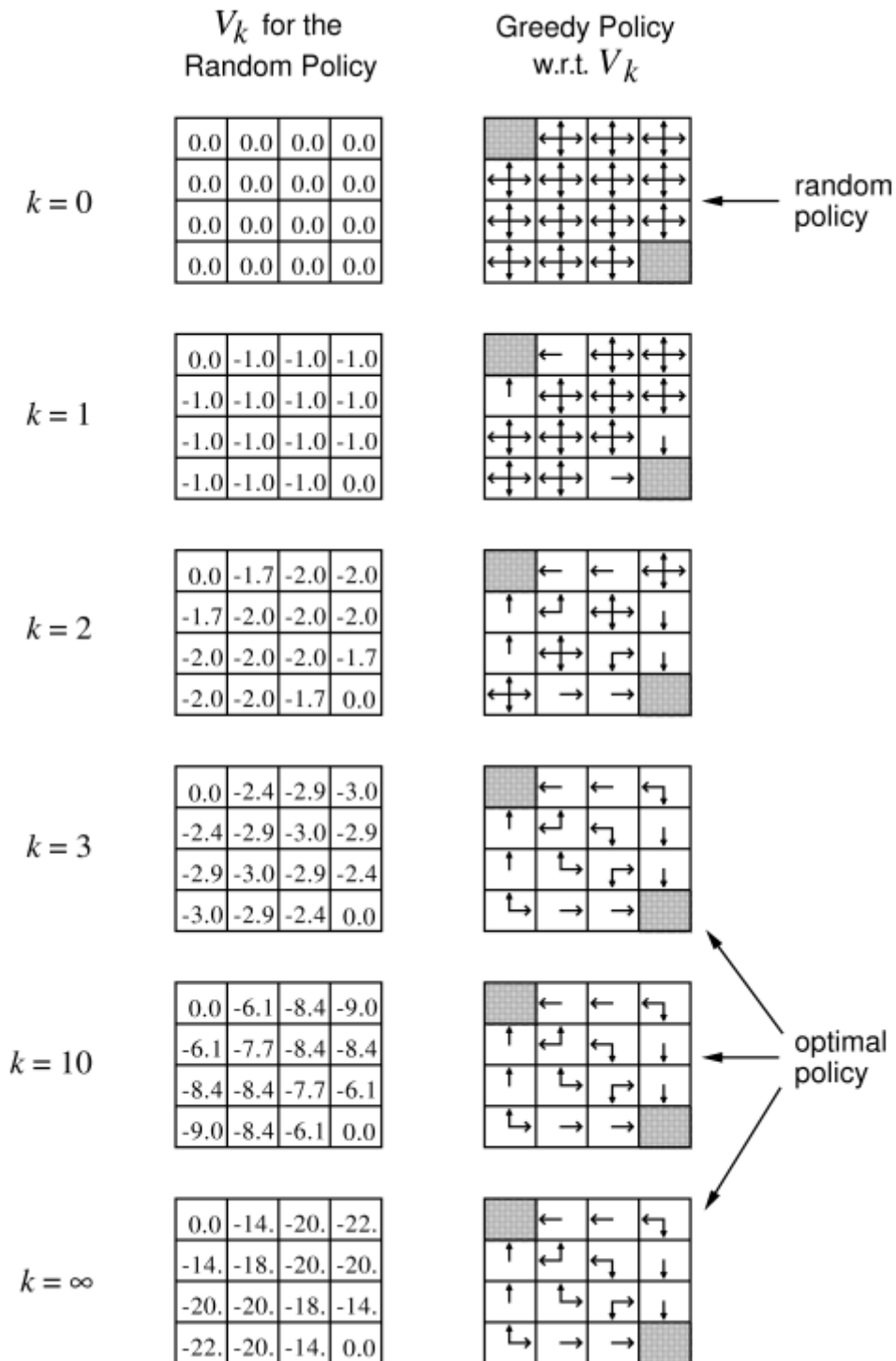
衰减系数 γ ：设定为常数1

当前策略 π ：智能体采用随机行动策略，在任何一个非终止状态下有均等的几率采取任一移动方向这个行为，即 $\pi(n|\cdot) = \pi(e|\cdot) = \pi(w|\cdot) = \pi(s|\cdot) = 0.25$ 。

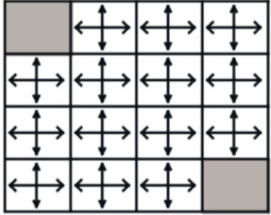
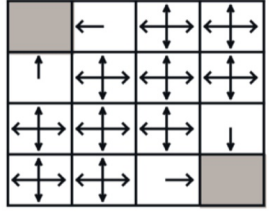
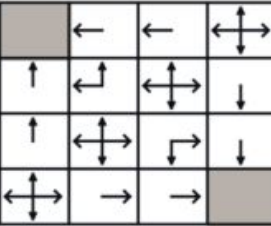
4.2、问题定义：

评估在这个方格世界里给定的策略。即求解该方格世界在给定策略下的（状态）价值函数，也就是求解在给定策略下，该方格世界里每一个状态的价值。

4.3、结果：



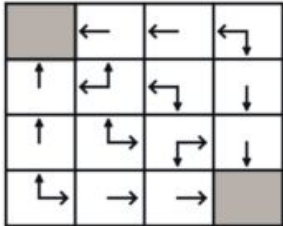
4.4、计算过程

迭代次数	状态价值函数	注释	Greedy Policy																
k = 0	<table> <tr><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td></tr> <tr><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td></tr> <tr><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td></tr> <tr><td>0.0</td><td>0.0</td><td>0.0</td><td>0.0</td></tr> </table>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	根据当前的状态价值，无法得出比随机策略更好的策略。	
0.0	0.0	0.0	0.0																
0.0	0.0	0.0	0.0																
0.0	0.0	0.0	0.0																
0.0	0.0	0.0	0.0																
k = 1	<table> <tr><td>0.0</td><td>-1.0</td><td>-1.0</td><td>-1.0</td></tr> <tr><td>-1.0</td><td>-1.0</td><td>-1.0</td><td>-1.0</td></tr> <tr><td>-1.0</td><td>-1.0</td><td>-1.0</td><td>-1.0</td></tr> <tr><td>-1.0</td><td>-1.0</td><td>-1.0</td><td>0.0</td></tr> </table>	0.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	0.0	<p>绿色的-1.0，注意↑试图离开方格世界，所以其位置将不会发生改变，而且这里的$v(s')$用的还是k=0 的值。</p> $ \begin{aligned} -1.0 &= \sum_{a \in A} \pi(a s) \left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v(s') \right) \\ &= \pi(n \uparrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times 0.0)) + \\ &\quad \pi(e \rightarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times 0.0)) + \\ &\quad \pi(s \downarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times 0.0)) + \\ &\quad \pi(w \leftarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times 0.0)) \\ &= 0.25 \times (-1.0 + 0.0) + \\ &\quad 0.25 \times (-1.0 + 0.0) + \\ &\quad 0.25 \times (-1.0 + 0.0) + \\ &\quad 0.25 \times (-1.0 + 0.0) \\ &= -1.0 \end{aligned} $ <p>蓝色的-1.0 计算方式同上，不再赘述。</p>	
0.0	-1.0	-1.0	-1.0																
-1.0	-1.0	-1.0	-1.0																
-1.0	-1.0	-1.0	-1.0																
-1.0	-1.0	-1.0	0.0																
k = 2	<table> <tr><td>0.0</td><td>-1.7</td><td>-2.0</td><td>-2.0</td></tr> <tr><td>-1.7</td><td>-2.0</td><td>-2.0</td><td>-2.0</td></tr> <tr><td>-2.0</td><td>-2.0</td><td>-2.0</td><td>-1.7</td></tr> <tr><td>-2.0</td><td>-2.0</td><td>-1.7</td><td>0.0</td></tr> </table>	0.0	-1.7	-2.0	-2.0	-1.7	-2.0	-2.0	-2.0	-2.0	-2.0	-2.0	-1.7	-2.0	-2.0	-1.7	0.0	<p>绿色的-1.7:</p> $ \begin{aligned} -1.7 &= \sum_{a \in A} \pi(a s) \left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v(s') \right) \\ &= \pi(n \uparrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-1.0))) + \\ &\quad \pi(e \rightarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-1.0))) + \\ &\quad \pi(s \downarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-1.0))) + \\ &\quad \pi(w \leftarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times 0.0)) \\ &= 0.25 \times (-1.0 - 1.0) + \\ &\quad 0.25 \times (-1.0 - 1.0) + \\ &\quad 0.25 \times (-1.0 - 1.0) + \\ &\quad 0.25 \times (-1.0 + 0.0) \\ &= 0.25 \times (-2.0 - 2.0 - 2.0 - 1.0) \\ &= 0.25 \times (-7.0) = -1.75 \end{aligned} $	
0.0	-1.7	-2.0	-2.0																
-1.7	-2.0	-2.0	-2.0																
-2.0	-2.0	-2.0	-1.7																
-2.0	-2.0	-1.7	0.0																

知乎 @搬砖的旺财

知乎 @搬砖的旺财

		<p>蓝色的-2.0:</p> $ \begin{aligned} -2.0 &= \sum_{a \in A} \pi(a s) \left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v(s') \right) \\ &= \pi(n \uparrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-1.0))) + \\ &\quad \pi(e \rightarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-1.0))) + \\ &\quad \pi(s \downarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-1.0))) + \\ &\quad \pi(w \leftarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-1.0))) \\ &= 0.25 \times (-1.0 - 1.0) + \\ &\quad 0.25 \times (-1.0 - 1.0) + \\ &\quad 0.25 \times (-1.0 - 1.0) + \\ &\quad 0.25 \times (-1.0 - 1.0) \\ &= 0.25 \times (-2.0 - 2.0 - 2.0 - 2.0) \\ &= 0.25 \times (-8.0) = -2.0 \end{aligned} $	知乎 @搬砖的旺财
--	--	---	-----------

k = 3	<table border="1"> <tr><td>0.0</td><td>-2.4</td><td>-2.9</td><td>-3.0</td></tr> <tr><td>-2.4</td><td>-2.9</td><td>-3.0</td><td>-2.9</td></tr> <tr><td>-2.9</td><td>-3.0</td><td>-2.9</td><td>-2.4</td></tr> <tr><td>-3.0</td><td>-2.9</td><td>-2.4</td><td>0.0</td></tr> </table>	0.0	-2.4	-2.9	-3.0	-2.4	-2.9	-3.0	-2.9	-2.9	-3.0	-2.9	-2.4	-3.0	-2.9	-2.4	0.0	<p>绿色的-2.4:</p> $ \begin{aligned} -2.4 &= \sum_{a \in A} \pi(a s) \left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v(s') \right) \\ &= \pi(n \uparrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-1.75))) + \\ &\quad \pi(e \rightarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-2.0))) + \\ &\quad \pi(s \downarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-2.0))) + \\ &\quad \pi(w \leftarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times 0.0)) \\ &= 0.25 \times (-1.0 - 1.75) + \\ &\quad 0.25 \times (-1.0 - 2.0) + \\ &\quad 0.25 \times (-1.0 - 2.0) + \\ &\quad 0.25 \times (-1.0 + 0.0) \\ &= 0.25 \times (-2.75 - 3.0 - 3.0 - 1.0) \\ &= 0.25 \times (-9.75) = -2.4375 \end{aligned} $ <p>蓝色的-3.0:</p>	 <p>根据该价值函数已经得到了最佳策略。</p> <p>知乎 @搬砖的旺财</p>
0.0	-2.4	-2.9	-3.0																
-2.4	-2.9	-3.0	-2.9																
-2.9	-3.0	-2.9	-2.4																
-3.0	-2.9	-2.4	0.0																

		$ \begin{aligned} -3.0 &= \sum_{a \in A} \pi(a s) \left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v(s') \right) \\ &= \pi(n \uparrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-2.0))) + \\ &\quad \pi(e \rightarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-2.0))) + \\ &\quad \pi(s \downarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-2.0))) + \\ &\quad \pi(w \leftarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-2.0))) \\ &= 0.25 \times (-1.0 - 2.0) + \\ &\quad 0.25 \times (-1.0 - 2.0) + \\ &\quad 0.25 \times (-1.0 - 2.0) + \\ &\quad 0.25 \times (-1.0 - 2.0) \\ &= 0.25 \times (-3.0 - 3.0 - 3.0 - 3.0) \\ &= 0.25 \times (-12.0) = -3.0 \end{aligned} $ <p>橘黄色的-2.9:</p> $ \begin{aligned} -2.9 &= \sum_{a \in A} \pi(a s) \left(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v(s') \right) \\ &= \pi(n \uparrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-2.0))) + \\ &\quad \pi(e \rightarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-2.0))) + \\ &\quad \pi(s \downarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-1.7))) + \\ &\quad \pi(w \leftarrow \bullet) \times (-1.0 + 1.0 \times (1.0 \times (-2.0))) \\ &= 0.25 \times (-1.0 - 2.0) + \\ &\quad 0.25 \times (-1.0 - 2.0) + \\ &\quad 0.25 \times (-1.0 - 1.7) + \\ &\quad 0.25 \times (-1.0 - 2.0) \\ &= 0.25 \times (-3.0 - 3.0 - 2.7 - 3.0) \\ &= 0.25 \times (-11.7) = -2.925 \end{aligned} $																	
k = 1 0	<table border="1"> <tr><td>0.0</td><td>-6.1</td><td>-8.4</td><td>-9.0</td></tr> <tr><td>-6.1</td><td>-7.7</td><td>-8.4</td><td>-8.4</td></tr> <tr><td>-8.4</td><td>-8.4</td><td>-7.7</td><td>-6.1</td></tr> <tr><td>-9.0</td><td>-8.4</td><td>-6.1</td><td>0.0</td></tr> </table>	0.0	-6.1	-8.4	-9.0	-6.1	-7.7	-8.4	-8.4	-8.4	-8.4	-7.7	-6.1	-9.0	-8.4	-6.1	0.0		
0.0	-6.1	-8.4	-9.0																
-6.1	-7.7	-8.4	-8.4																
-8.4	-8.4	-7.7	-6.1																
-9.0	-8.4	-6.1	0.0																
k = ∞	<table border="1"> <tr><td>0.0</td><td>-14.</td><td>-20.</td><td>-22.</td></tr> <tr><td>-14.</td><td>-18.</td><td>-20.</td><td>-20.</td></tr> <tr><td>-20.</td><td>-20.</td><td>-18.</td><td>-14.</td></tr> <tr><td>-22.</td><td>-20.</td><td>-14.</td><td>0.0</td></tr> </table>	0.0	-14.	-20.	-22.	-14.	-18.	-20.	-20.	-20.	-20.	-18.	-14.	-22.	-20.	-14.	0.0		
0.0	-14.	-20.	-22.																
-14.	-18.	-20.	-20.																
-20.	-20.	-18.	-14.																
-22.	-20.	-14.	0.0																

