高级人工智能-强化学习补充部分：棋类游戏的贝尔曼方程

1. 贝尔曼方程

$V_\pi(s)=\sum_\pi \pi(a|s)\sum_{s',r} P(s',r|s,a)(r+\gamma V_\pi(s'))$

这个过程是：首先从策略的分布π中采样一个行动a出来，然后在状态s执行a会以概率p转移到另一个状态s'，并且立即获得报酬r。于是，一个状态的估值由执行a转移到s'获得 的立即报酬r(s,a,s')以及s'的估值决定（乘上折扣因子）。

在非常一般的模型中，执行a会以不同的概率转移到不同的后续状态s'，而且获得的报酬r(s,a,s')也不是一个固定值，可能是一个分布。注：报酬与s a s'三个变量有关，所以表示成r(s,a,s')。

许多情况下，对于$P(s',r|s,a)$和$r(s,a,s')$的确定查表就够了。

表 4.3　转移概率和报酬

状态 i	可用行动 a	转移概率 $q(j\|i,a)$		报酬 $r(i,a)$
		$j=1$	$j=2$	
1	1	0.5	0.5	5
	2	0	1	10
2	1	0	1	-1

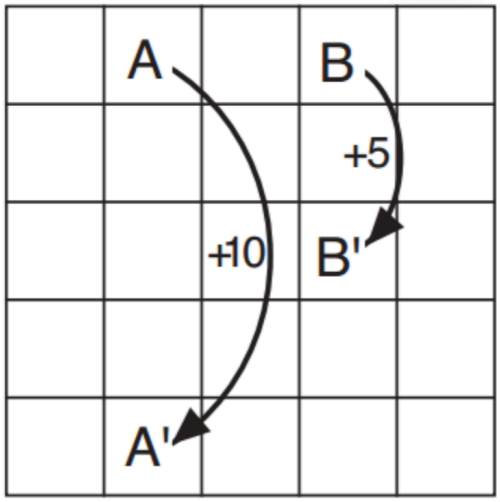
这就是一个MDP的表。在状态1执行行动1时，转移到1和2的概率都是0.5,并且获得的报酬都是5。这个表没有反映出策略的分布来。而且表中设定的报酬只与s a有关。（来源：本学期运筹学通论课程PPT）

2. 棋类游戏的贝尔曼方程组

状态估值函数有两个很简单的解法：当成线性方程组求解，或者动态规划的迭代法。这里给大家列出棋类游戏初始状态的贝尔曼方程组（未经过策略提升）。

- 游戏规则：正常移动一次的奖励为0；出界时回到当前位置，奖励为-1；A位置选择任意方向都到达A‘，奖励为+10；B位置选择任意方向都到达B’，奖励为+5；折扣率$\gamma = 0.9$。
- 假定策略为：从每个格子等概率选择四个移动方向（行为）

显示菜单



游戏规则

Actions

策略

3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

策略对应的状态估值函数

首先把棋盘编号。左上角是1号，按照从左往右、从上往下递增的方式编号，所以右下角是25号。A是2号，B是4号，A'是22号，B‘是14号。
初始策略是上下左右移动各1/4，也就是对任意的s,都有:

$$\pi(\uparrow|s) = \pi(\downarrow|s) = \pi(\leftarrow|s) = \pi(\rightarrow|s) = \frac{1}{4}$$

注意：这里的$P(s', r|s, a)$恒等于1。也就是说，在棋盘上向上、下、左、右移动就一定到达上面、下面、左面、右面的格子（一般情况）；如果向上移动出界，那么一定回到原来的格子，就是不动；如果是从A出发的，就一定到达A'。如果$P(s', r|s, a)$不等于1，意思就是执行了向上移动后，有一定概率到达上面以外的格子。而我们的游戏设定是不存在这种转移到其他s'的可能性。而且这里面报酬r也是固定值，不是分布。

带入π，以及奖励、折扣因子等，由于P=1就不写在里面了。贝尔曼方程组如下：(按照上，右，下，左的行动顺序)

$$V1 = \frac{1}{4}(-1 + 0.9 * V1) + \frac{1}{4}(0 + 0.9 * V2) + \frac{1}{4}(0 + 0.9 * V6) + \frac{1}{4}(-1 + 0.9 * V1)$$

$$V2(A) = \frac{1}{4}(10 + 0.9 * V22) + \frac{1}{4}(10 + 0.9 * V22) + \frac{1}{4}(10 + 0.9 * V22) + \frac{1}{4}(10 + 0.9 * V22)$$

$$V3 = \frac{1}{4}(-1 + 0.9 * V3) + \frac{1}{4}(+0.9 * V4) + \frac{1}{4}(0 + 0.9 * V8) + \frac{1}{4}(0 + 0.9 * V2)$$

$$V4(B) = \frac{1}{4}(5 + 0.9 * V14) + \frac{1}{4}(5 + 0.9 * V14) + \frac{1}{4}(5 + 0.9 * V14) + \frac{1}{4}(5 + 0.9 * V14)$$

$$V5 = \frac{1}{4}(-1 + 0.9 * V5) + \frac{1}{4}(-1 + 0.9 * V5) + \frac{1}{4}(0 + 0.9 * V10) + \frac{1}{4}(0 + 0.9 * V4)$$

$$V6 = \frac{1}{4}(0 + 0.9 * V1) + \frac{1}{4}(0 + 0.9 * V7) + \frac{1}{4}(0 + 0.9 * V11) + \frac{1}{4}(-1 + 0.9 * V6)$$

$V7 = \frac{1}{4}(0 + 0.9 * V2) + \frac{1}{4}(0 + 0.9 * V8) + \frac{1}{4}(0 + 0.9 * V12) + \frac{1}{4}(0 + 0.9 * V6)$

$V8 = \frac{1}{4}(0 + 0.9 * V3) + \frac{1}{4}(0 + 0.9 * V9) + \frac{1}{4}(0 + 0.9 * V13) + \frac{1}{4}(0 + 0.9 * V7)$

$V9 = \frac{1}{4}(0 + 0.9 * V4) + \frac{1}{4}(0 + 0.9 * V10) + \frac{1}{4}(0 + 0.9 * V14) + \frac{1}{4}(0 + 0.9 * V8)$

$V10 = \frac{1}{4}(0 + 0.9 * V5) + \frac{1}{4}(-1 + 0.9 * V10) + \frac{1}{4}(0 + 0.9 * V15) + \frac{1}{4}(0 + 0.9 * V9)$

$V11 = \frac{1}{4}(0 + 0.9 * V6) + \frac{1}{4}(0 + 0.9 * V12) + \frac{1}{4}(0 + 0.9 * V16) + \frac{1}{4}(-1 + 0.9 * V11)$

$V12 = \frac{1}{4}(0 + 0.9 * V7) + \frac{1}{4}(0 + 0.9 * V13) + \frac{1}{4}(0 + 0.9 * V17) + \frac{1}{4}(0 + 0.9 * V11)$

$V13 = \frac{1}{4}(0 + 0.9 * V8) + \frac{1}{4}(0 + 0.9 * V14) + \frac{1}{4}(0 + 0.9 * V18) + \frac{1}{4}(0 + 0.9 * V12)$

$V14 = \frac{1}{4}(0 + 0.9 * V9) + \frac{1}{4}(0 + 0.9 * V15) + \frac{1}{4}(0 + 0.9 * V19) + \frac{1}{4}(0 + 0.9 * V13)$

$V15 = \frac{1}{4}(0 + 0.9 * V10) + \frac{1}{4}(-1 + 0.9 * V15) + \frac{1}{4}(0 + 0.9 * V20) + \frac{1}{4}(0 + 0.9 * V14)$

$V16 = \frac{1}{4}(0 + 0.9 * V11) + \frac{1}{4}(0 + 0.9 * V17) + \frac{1}{4}(0 + 0.9 * V21) + \frac{1}{4}(-1 + 0.9 * V16)$

$V17 = \frac{1}{4}(0 + 0.9 * V12) + \frac{1}{4}(0 + 0.9 * V18) + \frac{1}{4}(0 + 0.9 * V22) + \frac{1}{4}(0 + 0.9 * V16)$

$V18 = \frac{1}{4}(0 + 0.9 * V13) + \frac{1}{4}(0 + 0.9 * V19) + \frac{1}{4}(0 + 0.9 * V23) + \frac{1}{4}(0 + 0.9 * V17)$

$V19 = \frac{1}{4}(0 + 0.9 * V14) + \frac{1}{4}(0 + 0.9 * V120) + \frac{1}{4}(0 + 0.9 * V24) + \frac{1}{4}(0 + 0.9 * V18)$

$V20 = \frac{1}{4}(0 + 0.9 * V15) + \frac{1}{4}(-1 + 0.9 * V20) + \frac{1}{4}(0 + 0.9 * V25) + \frac{1}{4}(0 + 0.9 * V19)$

$V21 = \frac{1}{4}(0 + 0.9 * V16) + \frac{1}{4}(0 + 0.9 * V22) + \frac{1}{4}(-1 + 0.9 * V21) + \frac{1}{4}(-1 + 0.9 * V21)$

$V22 = \frac{1}{4}(0 + 0.9 * V17) + \frac{1}{4}(0 + 0.9 * V23) + \frac{1}{4}(-1 + 0.9 * V22) + \frac{1}{4}(0 + 0.9 * V21)$

$V23 = \frac{1}{4}(0 + 0.9 * V18) + \frac{1}{4}(0 + 0.9 * V24) + \frac{1}{4}(-1 + 0.9 * V23) + \frac{1}{4}(0 + 0.9 * V22)$

$V24 = \frac{1}{4}(0 + 0.9 * V19) + \frac{1}{4}(0 + 0.9 * V25) + \frac{1}{4}(-1 + 0.9 * V24) + \frac{1}{4}(0 + 0.9 * V23)$

$V25 = \frac{1}{4}(0 + 0.9 * V20) + \frac{1}{4}(-1 + 0.9 * V25) + \frac{1}{4}(-1 + 0.9 * V25) + \frac{1}{4}(0 + 0.9 * V25)$

抓几个典型说明一下：
$V1 = \frac{1}{4}(-1 + 0.9 * V1) + \ldots$, 第一项的意思是从格子1以1/4的概率向上移动后，一定回到自己，立即报酬为−1，此时s'=s。

$V2(A) = \frac{1}{4}(10 + 0.9 * V22) + \ldots$, 第一项的意思是从格子A以1/4的概率选择向上移动后，一定到达A'，立即报酬为10，此时s'=A'.

上面那堆25个方程就是线性方程组$AV = b$。系数矩阵A里，每行最多只有五个变量系数不为0。
$V = (V1, \ldots, V25)^T$

完毕.