《强化学习》课程之第三讲（2021年春季研究生）

数学建模

苏州大学计算机科学与技术学院
主讲：刘 全

目　录

引言

马尔可夫决策过程（MDP）：

- 强化学习的数学理论基础；
- 以概率形式对强化学习任务进行建模；
- 对强化学习过程中出现的状态、动作、状态转移概率和奖赏等概念进行抽象表达。

3.1 马尔可夫决策过程（1）

➢ **马尔可夫性质：**

在某一任务中，如果**Agent**从环境中得到的下一状态仅依赖于当前状态，而不考虑历史状态，即：

$$P[S_{t+1} \mid S_t] = P[S_{t+1} \mid S_1, \ldots, S_t]$$

那么该任务就满足马尔可夫性质。

➢ **马尔可夫过程（Markov Process，MP）：**

由二元组 $(\mathcal{S}, \mathcal{S})$ 中的(S_t, S_{t+1})组成的马尔可夫链，该链中的所有状态都满足马尔可夫性质。

3.1 马尔可夫决策过程（2）

- **马尔可夫奖赏过程（Markov Reward Process，MRP）：**

  由三元组$(\mathcal{S}, P, \mathcal{R})$组成的马尔可夫过程。根据概率，状态自发地进行转移，其状态转移概率P与动作无关，记为：

  $$P[S_{t+1} = s' \mid S_t = s]$$

- **马尔可夫决策过程（Markov Decision Process，MDP）：**

  由四元组$(\mathcal{S}, \mathcal{A}, P, \mathcal{R})$组成的马尔可夫过程，状态依靠动作进行转移。马尔可夫决策过程分为：

  - 有穷马尔可夫决策过程；
  - 无穷马尔可夫决策过程。

3.1 马尔可夫决策过程（3）

（1）状态（state）或观测值（observation）

马尔可夫决策过程由四元组组成：

$$(\mathcal{S}, \mathcal{A}, P, \mathcal{R})$$

$\mathcal{S}$：用来表示不包含终止状态的状态空间；

$\mathcal{S}^{+}$：用来表示包含终止状态的状态空间；

s：用来表示状态空间中的某一状态。通常用向量来表示，可分为离散状态和连续状态两种类型。

3.1 马尔可夫决策过程（4）

（2）动态（action）

$\mathcal{A}$: 表示动作空间；

$\mathcal{A}(s)$: 表示状态**s**的动作空间；

a: 表示动作空间中的某一个动作。

- 通常用向量来表示可分为：

  离散动作和连续动作

  两种类型。

3.1 马尔可夫决策过程（5）

（3）状态转移（state transition）

p： 表示状态转移概率，即在状态$\boldsymbol{s}$下，执行动作$\boldsymbol{a}$转移到$\boldsymbol{s}'$的概率。

- 可以表示为如下两种形式：

$$p(s',r \mid s,a) = P[S_{t+1}=s', R_{t+1}=r \mid S_t=s, A_t=a] \quad , \quad \sum_{s'}\sum_{r} p(s',r \mid s,a)=1$$

$$p(s' \mid s,a) = P[S_{t+1}=s' \mid S_t=s, A_t=a] \quad , \quad \sum_{s'} p(s' \mid s,a)=1$$

确定环境： $p(s',r \mid s,a)=1$

随机环境： $p(s',r \mid s,a) \neq 1$

3.1 马尔可夫决策过程（6）

（4）奖赏（reward）

$\mathcal{R}$：　表示奖赏空间；

$r(s,a,s')$：　表示**Agent**在状态***s***下，执行动作***a***转移到***s'***所获得的期望奖赏，可分为离散奖赏和连续奖赏两种。

➢　奖赏公式可以表示为：

$$r(s,a,s') = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a, S_{t+1} = s'] = \sum_r r \frac{p(s',r \mid s,a)}{p(s' \mid s,a)}$$

或者：

$$r(s,a) = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a] = \sum_r r \sum_{s'} p(s',r \mid s,a)$$

3.1 马尔可夫决策过程（7）

对于奖赏，可以从两个方面进行理解：

- 先获得奖赏再进入下一状态：奖赏R_{t+1}与当前状态S_t和动作A_t相关；
- 先进入下一状态再获得奖赏：奖赏R_{t+1}与当前状态S_t、动作A_t和下一状态S_{t+1}相关，这也是奖赏用R_{t+1}表示的一个重要原因。

3.1 马尔可夫决策过程（8）

例3.1 确定环境下扫地机器人任务的MDP数学建模

考虑图中描述的确定环境**MDP问题：**

一个扫地机器人，在躲避障碍物的同时，一方面需要到指定的位置收集垃圾，另一方面可以到指定位置给电池充电。

20	21	22	23	24
15	16	17	18	19
10	11	12	13	14
5	6	7	8	9
0	1	2	3	4

3.1 马尔可夫决策过程（9）

扫地机器人任务的MDP数学建模如下：

- **状态空间：**

离散化为**24**个不同的状态（除去[3, 3]），用集合表示为：

$$\mathcal{S}^+=\left\{\begin{array}{l} S_0{:}[1,1],\ S_1{:}[2,1],\ S_2{:}[3,1],\ \cdots,\ S_{11}{:}[2,3], S_{13}{:}[4,3],\ \cdots, \\ S_{19}{:}[5,4],\cdots,\ S_{23}{:}[4,5],\ S_{24}{:}[5,5] \end{array}\right\}$$

- **动作空间：**

离散化为上、下、左、右**4**个不同的动作，用集合表示为：

$$\mathcal{A}=\left\{上{:}[0,1],\ 下{:}[0,-1],\ 左{:}[-1,0],\ 右{:}[1,0]\right\}$$

3.1 马尔可夫决策过程（10）

- **状态转移函数：**

  - ✓ 映射为下一个状态：

$$f(s,a)=\begin{cases} s+a, & s\neq[1,1] \text{ 且 } s\neq[5,4] \text{ 且 } s+a\neq[3,3] \\ s, & \text{其他} \end{cases}$$

  - ✓ 映射为下一个状态的概率：

$$p(s,a,s')=\begin{cases} 1, & (s+a=s' \text{ 且 } s+a\neq[3,3]) \\ & \text{或}\left(\left(s=[1,1] \text{ 或 } s=[5,4] \text{ 或 } s+a=[3,3]\right) \text{ 且 } s=s'\right) \\ 0, & \text{其他} \end{cases}$$

3.1 马尔可夫决策过程（11）

- 奖赏函数：
  - ✓ 到达状态 $S_{19}=[5,4]$，可以捡到垃圾，得到+3的奖赏；
  - ✓ 到达状态 $S_0=[1,1]$，充电，得到+1的奖赏；
  - ✓ 机器人采取动作向坐标[3, 3]处移动时，会撞到障碍物，保持原地不动，并得到-10的奖赏；
  - ✓ 其他情况，奖赏均为0。

$$r(s,a)=\begin{cases} +3, & \text{如果} s \neq [5,4] \text{ 且 } s+a=[5,4] \\ +1, & \text{如果} s \neq [1,1] \text{ 且 } s+a=[1,1] \\ -10, & \text{如果} s+a=[3,3] \\ 0, & \text{其他} \end{cases}$$

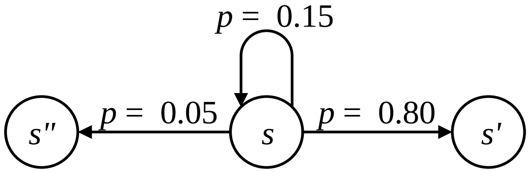
3.1 马尔可夫决策过程（12）

例3.2 随机环境下扫地机器人任务的MDP数学建模

重新考虑图中描述的随机环境

MDP问题：

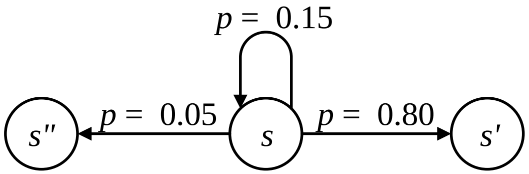
假设由于地面的问题，采取某一动作后，状态转换不再确定。当采取某一动作试图向某一方向移动时，机器人成功移动的概率为**0.80**，保持原地不动的概率为**0.15**，移动到相反方向的概率为**0.05**。



3.1 马尔可夫决策过程（13）

在随机环境下，状态空间、动作空间与确定环境是完全相同的，其随机性主要体现在状态转移函数和奖赏函数上。根据任务的随机性，状态转移只能用概率来表示。

➢ 状态转移函数



$$p(s,a,s')=\begin{cases}0.80, & \text{如果} s+a=s' \text{ 且 } s\neq s'\\ 0.15, & \text{如果} s=s' \text{ 且 } s\neq[1,1] \text{ 且 } s\neq[5,4]\\ 0.05, & \text{如果} s-a=s' \text{ 且 } s\neq s'\end{cases}$$

3.1 马尔可夫决策过程（14）

➢ 奖赏函数

在随机环境下，奖赏的获取不单纯受(s,a)的影响，还与下一状态 s' 相关。

$$
r(s,a,s')=\begin{cases}+3, & \text{如果} s\neq[5,4] \text{ 且 } s'=[5,4] \\ +1, & \text{如果} s\neq[1,1] \text{ 且 } s'=[1,1] \\ -10, & \text{如果} s+a=[3,3] \text{ 且 } s=s' \\ 0, & \text{其他}\end{cases}
$$

目 录

3.2 基于模型和无模型（1）

从状态转移概率p是否已知的角度，强化学习可以分为基于模型（**model-based**）强化学习和无模型（**model-free**）强化学习两种：

- 基于模型：状态转移概率p已知，能够通过建立完备的环境模型来模拟真实反馈。相关算法如：动态规划法。
- 无模型：状态转移概率p未知，**Agent**所处的环境模型是未知的。相关算法：蒙特卡洛法、时序差分法、值函数近似以及策略梯度法。

3.2 基于模型和无模型（2）

- **基于模型的优缺点：**

优点：

- ✓ 能够基于模拟经验数据直接模拟真实环境；
- ✓ 具备推理能力，能够直接评估策略的优劣性；
- ✓ 能够与监督学习算法相结合，来求解环境模型。

3.2 基于模型和无模型（3）

- **基于模型的优缺点：**

缺点：

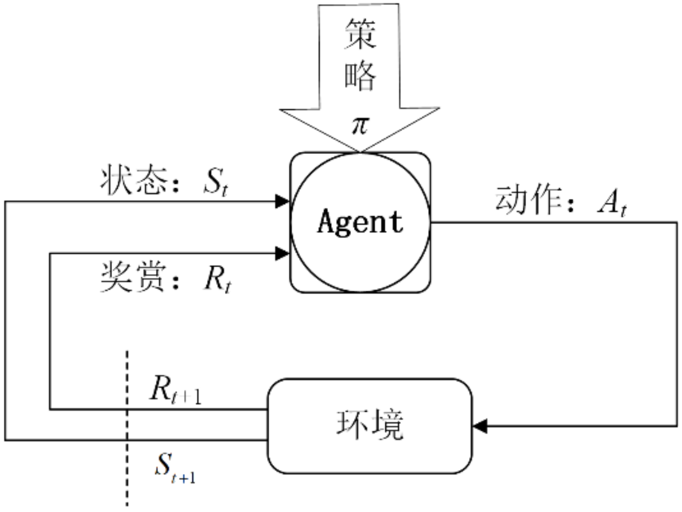
存在二次误差。两次近似误差具体体现在：

- ✓ 第一次近似误差：基于真实经验对模型进行学习，得到的模型仅仅是Agent对环境的近似描述。
- ✓ 第二次近似误差：基于模拟模型对值函数或策略进行学习时，存在学习误差。

目 录

3.3 求解强化学习任务（1）

在t时刻，**Agent**从环境中得到当前状态S_t，根据策略π执行动作A_t，并返回奖赏R_{t+1}和下一状态S_{t+1}。**Agent**通过不断地与环境交互进行学习，并在学习过程中不断更新策略，从而经过多次学习后，得到解决问题的最优策略。。



基于MDP的强化学习基本框架

3.3 求解强化学习任务（2）

3.3.1 策略

强化学习的目的就是：在MDP中搜索到最优策略。策略表示状态到动作的映射，即在某一状态下采取动作的概率分布。

与状态转移概率不同，策略概率通常是人为设定的。根据概率分布形式，策略可以分为确定策略和随机策略两种。

3.3 求解强化学习任务（3）

- 在**确定策略**下，**Agent**在某一状态下只会执行固定一个动作。可以表示为：

$$a = \pi(s)$$

- 在**随机策略**下，**Agent**在一个状态下可能会执行多种动作，随机策略将状态映射为执行动作的概率。可以表示为：

$$\pi(a \mid s) = P(a \mid s) = P(\boldsymbol{A}_t = a \mid \boldsymbol{S}_t = s)$$

3.3 求解强化学习任务（4）

MDP应用一个策略产生序列的方法：

- 从初始状态分布中产生一个初始状态 $\boldsymbol{S}_i = \boldsymbol{S}_0$；
- 根据策略 $\pi(a \mid \boldsymbol{S}_i)$，给出采取的动作 $\boldsymbol{A}_i$ ，并执行该动作 $\boldsymbol{A}_i$；
- 根据奖赏函数和状态转移函数得到奖赏 R_{i+1} 和下一个状态 $\boldsymbol{S}_{i+1}$ ；

  $$\boldsymbol{S}_i = \boldsymbol{S}_{i+1}$$

- 不断重复第（2）步到第（4）步的过程，产生一个序列：

  $$\boldsymbol{S}_0, \boldsymbol{A}_0, R_1, \boldsymbol{S}_1, \boldsymbol{A}_1, R_2, \boldsymbol{S}_2, \boldsymbol{A}_2, R_3, \boldsymbol{S}_3, \cdots$$

- 如果任务是情节式的，序列将终止于状态 $\boldsymbol{S}_{\mathrm{goal}}$ ；如果任务是连续式的，序列将无穷延续。

3.3 求解强化学习任务（5）

强化学习任务的两种随机性：

- **策略随机性：** 人为设定的。

$$\pi(a \mid s)$$

- **状态转移的随机性：** 任务本身所固有的特性。

$$p(s', r \mid s, a)$$

3.3 求解强化学习任务（6）

3.3.2 奖赏与回报

- **Agent会依据该策略得到一个状态-动作序列，其形式为：**

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, S_3, \cdots$$

- **定义马尔可夫决策过程的回报如下：**

$$G_t = R_{t+1} + R_{t+2} + \ldots + R_{\mathrm{T}}$$

实际情况中，需要引入折扣率γ，用于对未来奖赏赋予折扣，则回报定义如下：

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \quad , \quad \gamma \in [0,1]$$

$$\boxed{\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \\ &= R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \ldots) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}}$$

3.3 求解强化学习任务（7）

例3.3 设折扣率$\gamma = 0.2$，$\mathrm{T} = 4$， 奖赏序列为：

$$R_1 = 2, R_2 = 1, R_3 = 5, R_4 = 4$$

计算各时刻的回报：$G_0, G_1, \cdots, G_4$

$$G_4 = 0$$

$$G_3 = R_4 + \gamma * G_4 = 4 + 0.2 * 0 = 4$$

$$G_2 = R_3 + \gamma * G_3 = 5 + 0.2 * 4 = 5.8$$

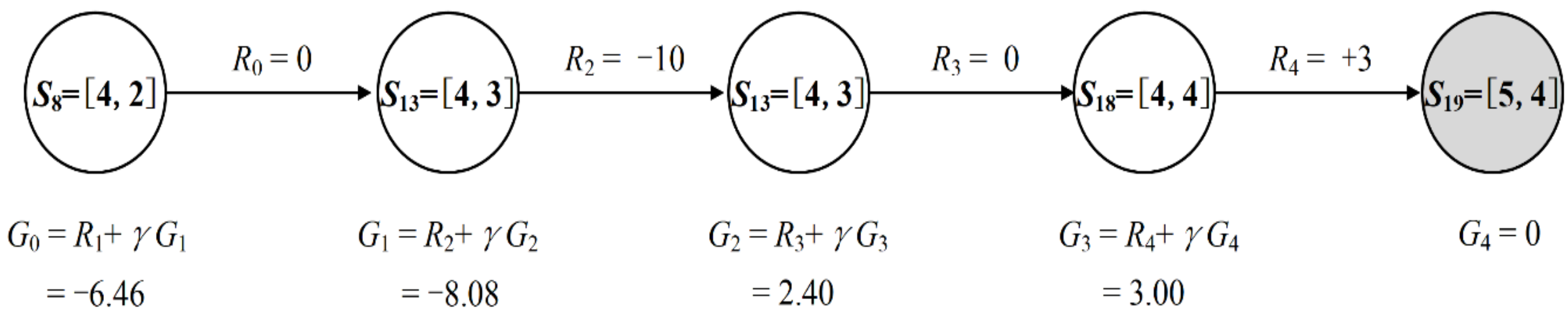
$$G_1 = R_2 + \gamma * G_2 = 1 + 0.2 * 5.8 = 2.16$$

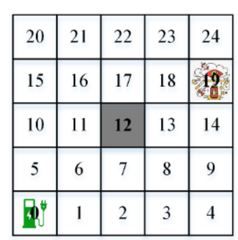
$$G_0 = R_1 + \gamma * G_1 = 2 + 0.2 * 2.16 = 2.432$$

3.3 求解强化学习任务（8）

例3.4 扫地机器人任务

选取机器人的一段移动轨迹，令折扣率为0.8，计算轨迹中每个状态的折扣回报。





3.3 求解强化学习任务（9）

3.3.3 值函数与贝尔曼方程

- 状态值函数（**state-value function**）：

状态值函数 $v_{\pi}(s)$ 表示遵循策略 π ，状态**s**的价值。

可表示为：

$$v_{\pi}(s) = \mathbb{E}_{\pi}(G_t \mid S_t = s)$$

- 动作值函数（**action-value function**）

动作值函数 $q_{\pi}(s,a)$ 表示遵循策略 π ，状态**s**采取

动作***a***的价值。可表示为：

$$q_{\pi}(s,a) = \mathbb{E}_{\pi}(G_t \mid S_t = s, A_t = a)$$

3.3 求解强化学习任务（10）

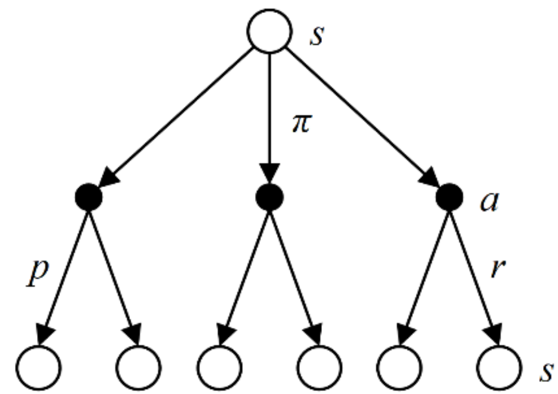
- 状态值函数的贝尔曼方程

动作值函数是在状态值函数的基础上考虑了执行动作$\boldsymbol{a}$所产生的影响。于是可以构建值函数的递归关系：

$$\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi(G_t \mid S_t = s) \\
&= \mathbb{E}_\pi(R_{t+1} + \gamma G_{t+1} \mid S_t = s) \\
&= \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma \mathbb{E}_\pi(G_{t+1} \mid S_{t+1} = s')\right] \\
&= \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma v_\pi(s')\right]
\end{aligned}$$

3.3 求解强化学习任务（11）

根据状态值函数贝尔曼方程，可以构建状态值函数更新图，空心圆表示状态，实心圆表示动作。由图可知，状态值函数与动作值函数满足如下关系式：



$$v_\pi(s) = \sum_a \pi(a \mid s) q_\pi(s, a)$$

3.3 求解强化学习任务（12）

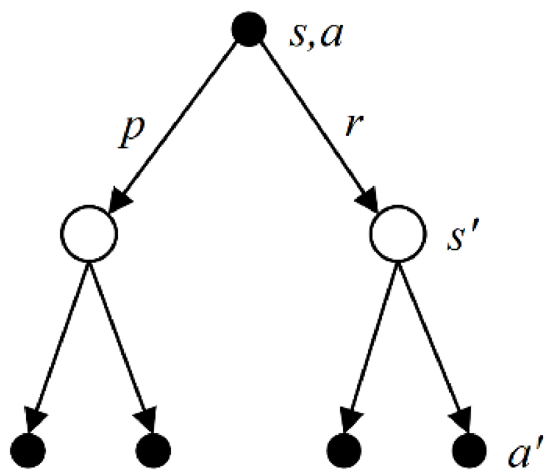
➢ 动作值函数的贝尔曼方程

与状态值函数的贝尔曼方程推导方式类似，同理可以得到动作值函数的贝尔曼方程：

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi(G_t \mid S_t = s, A_t = a) \\
&= \mathbb{E}_\pi(R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a) \\
&= \mathbb{E}_\pi(R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a) \\
&= \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma v_\pi(s')\right] \\
&= \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s',a')\right]
\end{aligned}
$$

3.3 求解强化学习任务（13）

根据动作值函数的贝尔曼方程，可以构建动作值函数更新图：

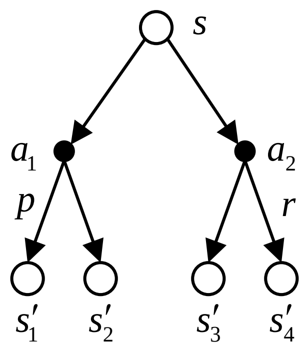


由左图可知，动作值函数与状态值函数满足如下关系式：

$$q_{\pi}(s,a) = r + \gamma \sum_{s'} p(s' \mid s,a) v_{\pi}(s')$$

3.3 求解强化学习任务（14）

例3.5 已知s_1' 、s_2'、s_3' 、s_4'的状态值，利用状态值函数的贝尔曼方程，表示 s 的状态值。



$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}_\pi(G_t \mid S_t = s) \\
&= \pi(a_1 \mid s)\, p(s_1', r_1 \mid s, a_1)\left[r_1 + \gamma v_\pi(s_1')\right] \\
&\quad + \pi(a_1 \mid s)\, p(s_2', r_2 \mid s, a_1)\left[r_2 + \gamma v_\pi(s_2')\right] \\
&\quad + \pi(a_2 \mid s)\, p(s_3', r_3 \mid s, a_2)\left[r_3 + \gamma v_\pi(s_3')\right] \\
&\quad + \pi(a_2 \mid s)\, p(s_4', r_4 \mid s, a_2)\left[r_4 + \gamma v_\pi(s_4')\right]
\end{aligned}
$$

3.3 求解强化学习任务（15）

例3.6 确定环境扫地机器人任务

确定情况下扫地机器人任务中，采用的随机策略为：

$$\pi\left(a \mid S_i\right) = 1/\left|\mathcal{A}(S_i)\right|,\quad a \in \mathcal{A}(S_i)$$

$\left|\mathcal{A}(S_i)\right|$ 表示状态 S_i 可以采取的动作数。

在折扣率$\gamma = 0.8$的情况下，求扫地机器人任务中每个状态的状态值。

3.3 求解强化学习任务（16）

首先，列出贝尔曼方程：

$$v_\pi(S_i) = \sum_{a \in \mathcal{A}(S_i)} \pi(a \mid S_i)\, p(S_i, a, s')\left(r(S_i, a) + \gamma v_\pi(s')\right)$$

根据贝尔曼方程，可以列出方程组：

$$\begin{cases}
v_\pi(S_1) = \frac{1}{3}\times(0+0.8\times v_\pi(S_6)) + \frac{1}{3}\times(1+0.8\times 0) + \frac{1}{3}\times(0+0.8\times v_\pi(S_2)) \\
v_\pi(S_2) = \frac{1}{3}\times(0+0.8\times v_\pi(S_7)) + \frac{1}{3}\times(0+0.8\times v_\pi(S_1)) + \frac{1}{3}\times(0+0.8\times v_\pi(S_3)) \\
\qquad\qquad \vdots \\
v_\pi(S_{11}) = \frac{1}{4}\times(0+0.8\times v_\pi(S_{16})) + \frac{1}{4}\times(0+0.8\times v_\pi(S_6)) + \\
\qquad\qquad \frac{1}{4}\times(0+0.8\times v_\pi(S_{10})) + \frac{1}{4}\times(-10+0.8\times v_\pi(S_{11})) \\
\qquad\qquad \vdots \\
v_\pi(S_{13}) = \frac{1}{4}\times(0+0.8\times v_\pi(S_{18})) + \frac{1}{4}\times(0+0.8\times v_\pi(S_8)) + \\
\qquad\qquad \frac{1}{4}\times(-10+0.8\times v_\pi(S_{13})) + \frac{1}{4}\times(0+0.8\times v_\pi(S_{14})) \\
\qquad\qquad \vdots \\
v_\pi(S_{23}) = \frac{1}{3}\times(0+0.8\times v_\pi(S_{18})) + \frac{1}{3}\times(0+0.8\times v_\pi(S_{22})) + \frac{1}{3}\times(0+0.8\times v_\pi(S_{24})) \\
v_\pi(S_{24}) = \frac{1}{2}\times(3+0.8\times 0) + \frac{1}{2}\times(0+0.8\times v_\pi(S_{23}))
\end{cases}$$

3.3 求解强化学习任务（17）

求解方程组，得到各个状态的状态值：

-1.111	-1.359	-1.615	-0.329	1.368
-1.417	-2.372	-4.369	-0.987	0.000
-1.831	-4.716		-3.987	-0.300
-0.731	-2.162	-4.649	-2.160	-0.887
0.000	-0.716	-1.772	-1.280	-0.867

3.3 求解强化学习任务（18）

3.3.4 最优策略与最优值函数

利用强化学习方法解决任务的关键在于：搜索出**MDP**中的最优策略。

- 最优策略就是使得值函数最大的策略。在有穷**MDP**中，由于状态空间和动作空间都是有穷的，所以策略也是有穷的。
- 更优策略π'，执行该策略时，所有状态的期望回报都大于或等于执行π策略的期望回报。也就是说，对于所有$s \in \mathcal{S}$，$\pi' \geq \pi$都存在$v_{\pi'}(s) \geq v_{\pi}(s)$。

3.3 求解强化学习任务（19）

- **最优状态值函数**定义为：最优策略可能不止一个，它们共享相同的状态值函数。

$$v_*(s) = v_{\pi_*}(s) = \max_{\pi} v_{\pi}(s) \quad , \quad s \in \mathcal{S}$$

- **最优动作值函数**定义为：在状态s处，执行动作a，并在随后的过程中采取最优策略π_*得到的期望回报，也就是在状态-动作对(s,a)处能够获得的最大价值。

$$q_*(s,a) = q_{\pi_*}(s,a) = \max_{\pi} q_{\pi}(s,a) \quad , \quad s \in \mathcal{S}, a \in \mathcal{A}$$

3.3 求解强化学习任务（20）

➢ 贝尔曼最优方程

基于状态值的贝尔曼最优方程：

$$
\begin{aligned}
v_*(s) &= \max_{a\in\mathcal{A}(s)} q_{\pi_*}(s,a) \\
&= \max_a \mathbb{E}_{\pi_*}(R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a) \\
&= \max_a \mathbb{E}(R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a) \\
&= \max_a \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma v_*(s')\right]
\end{aligned}
$$

3.3 求解强化学习任务（21）

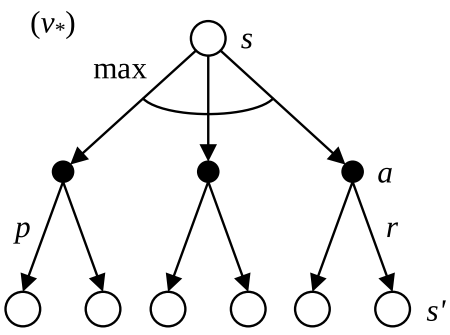
➢ 贝尔曼最优方程

基于动作值的贝尔曼最优方程：

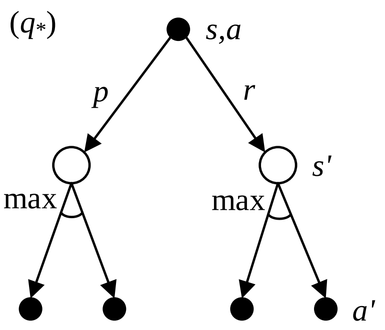
$$
\begin{aligned}
q_*(s,a) &= \mathbb{E}(R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a) \\
&= \mathbb{E}(R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a) \\
&= \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma \max_{a'} q_*(s', a') \right]
\end{aligned}
$$

3.3 求解强化学习任务（22）

➢ 贝尔曼最优方程



基于状态值的贝尔曼最优方程更新图



基于动作值的贝尔曼最优方程更新图

3.3 求解强化学习任务（23）

➢ 贝尔曼最优方程

Agent通过与环境不断地交互获得的信息来更新策略，以最终获得最优值函数。一旦获得最优状态值函数$\boldsymbol{v}_*$或最优动作值函数$\boldsymbol{q}_*$，Agent便能得到最优策略。

Agent可以直接选择最大动作值函数所对应的动作，这一方法也称为贪心动作选择方法，其表达式为：

$$A_t = \arg\max_a q_t(s,a)$$

3.3 求解强化学习任务（24）

例3.7 求解确定环境下扫地机器人任务的最优状态值函数，并给出最优策略。设折扣率 $\gamma = 0.8$。

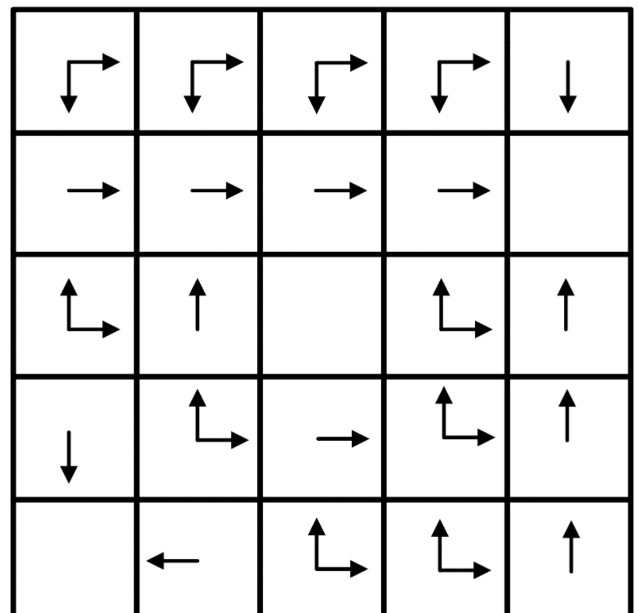
可以显式地给出在该扫地机器人任务，最优贝尔曼方程：

$$v_*(S_i) = \max_a p(S_i, a, s')\left(r(S_i, a) + \gamma v_*(s')\right)$$

3.3 求解强化学习任务（25）

利用第**4**章的值迭代算法，可以求得最优状态值和最优策略：

1.229	1.536	1.920	2.400	3.000
1.536	1.920	2.400	3.000	0.000
1.229	1.536		2.400	3.000
1.000	1.229	1.536	1.920	2.400
0.000	1.000	1.229	1.536	1.920



目 录

3.4 探索与利用（1）

- 强化学习的一大矛盾：探索与利用的平衡
  - ✓ **Agent**秉持利用机制（**exploitation**），为了得到最大回报，需要始终采用最优动作，即根据当前的值函数选择最优动作，最大限度地提升回报。
  - ✓ **Agent**需要探索机制（**exploration**），摒弃基于值函数的贪心策略，找到更多可能的动作来获得更好的策略，探索更多的可能性。

3.4 探索与利用（2）

- **行为策略（behavior policy）**：用于产生采样数据的策略，具备探索性，能够覆盖所有情况，通常采用ε-柔性策略；
- **目标策略（target policy）**：强化学习任务中待求解的策略，也就是待评估和改进的策略，一般不具备探索性，通常采用确定性贪心策略。

3.4 探索与利用（3）

- **同策略（on-policy）**：行为策略和目标策略相同。通过ε贪心策略平衡探索和利用，在保证初始状态-动作对(S_0, A_0)不变的前提下，确保每一组(*s*,*a*)都有可能被遍历到。常用算法为Sarsa和Sarsa(λ)算法。
- **异策略（off-policy）**：行为策略和目标策略不同。将探索与利用分开，在行为策略中贯彻探索原则：采样数据，得到状态-动作序列；在目标策略中贯彻利用原则：更新值函数并改进目标策略，以得到最优目标策略。常用算法为Q-learning和DQN算法。

目 录

3.5 小结（1）

- 本章主要介绍了强化学习的基础数学理论，以马尔可夫决策过程描述了Agent与环境的交互。状态是Agent选择动作的基础，通过动作的选择，完成状态的转移，并以奖赏评判Agent动作选择的优劣。
- 有限的状态、动作和收益共同构成了有限马尔可夫决策过程，回报刻画了Agent能获得的全部未来奖赏，对于不同的任务，未来状态的奖赏会有不同的折扣，而Agent的任务就是最大化回报。动作的选择依赖于Agent所采取的策略，而强化学习的目的就是获得最优策略。

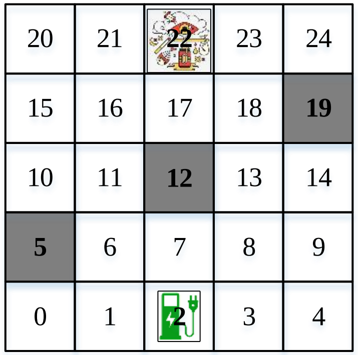
3.5 小结（2）

- 引入状态值和动作状态值来描述回报，通过贝尔曼最优方程将马尔可夫决策过程表达抽象化，从而可以相对容易地求解得到最优价值函数。在强化学习问题中，定义环境模型和明确最优值函数是计算最优策略的基础，在后续章节中，将进一步讨论如何求解最优策略。

3.6 习题（1）

1. 举例说明基于模型与无模型强化学习的异同点。

2. 分别给出如图3.12所示的确定环境和随机环境下扫地机器人任务的MDP数学模型。与例3.1和3.2相比，主要有两方面的变化：（1）图3.12中障碍物、充电桩及垃圾位置不同；（2）在任何状态下都有上、下、左、右4个不同的动作，当采取冲出边界的动作时，机器人保持原地不同。其他参数等设置与例3.1、3.2相同。



3.6 习题（2）

3.考虑一个折扣因子为 γ的连续式任务，其奖赏序列为：R_1，$R_2 = R_3 = \cdots =$ ，计算G_0 ,G_1的值。

4.（编程）通过解方程组计算：在确定环境、等概率策略下，扫地机器人在折扣率$\gamma = 0.8$的情况下，每个状态的状态值。

5. 简述同策略与异策略强化学习的异同点。

The End