强化学习（RL）

(RL)

第四章。

强化学习的蒙特卡洛方法

赛义德-赛义德万德，博士。

内容

在本章中。

- ✓ 基于模型的学习和无模型的学习
- ✓ 首次访问法MC预测
- ✓ 每次访问的方法MC预测
- ✓ 蒙特卡洛探索启动（MC-ES）。
- ✓ 贪婪的MC Epsilon (没有探索性的启动)

本章的目的。

- ✓ 理解基于模型和无模型算法的区别。学习基于预测和控制的蒙特卡洛不同方法。

蒙特卡洛算法（无模型RL）。

想法

° ✓ 在政策和价值迭代算法中，假设包括能够获得**完整环境模型**的代理人。

- 过渡动力学（每个行动后的可能状态）。
- 不需要互动，而且计算结果是可预测的（包括奖励）。

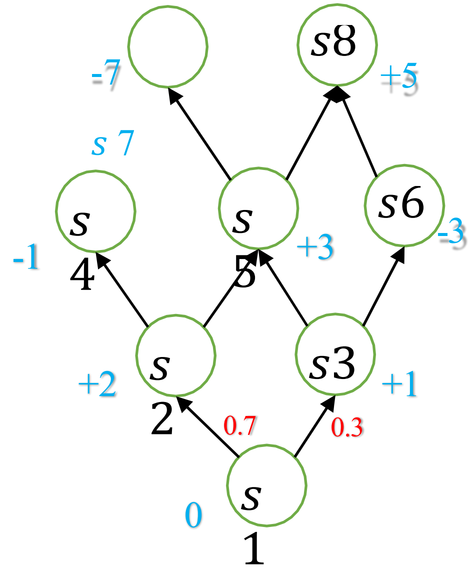
这在许多情况和环境下是不切实际的

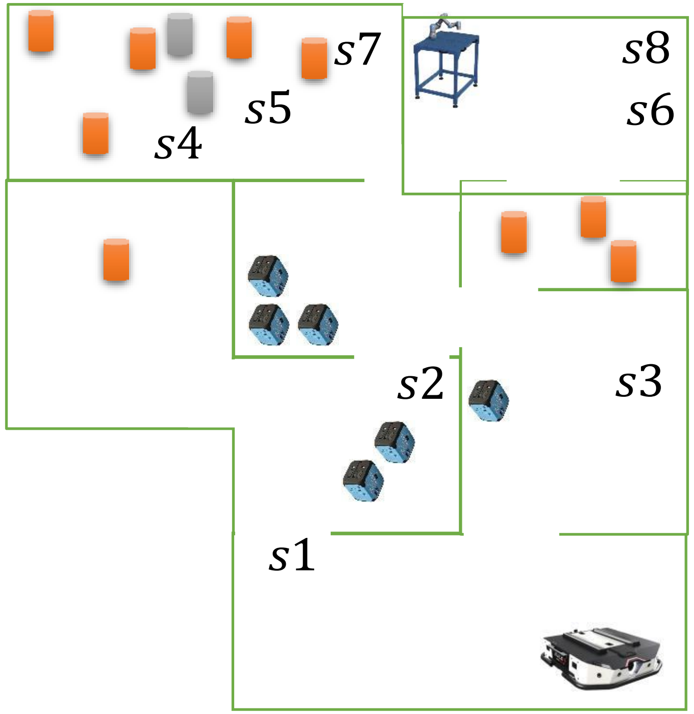
蒙特卡洛算法（无模型RL）。

想法

。

- ✓ **代理人也**许不能掌握环境的所有信息，我们需要算法来解决这个问题。
- ✓ 因此，我们需要在环境中进行互动的代理来学习政策





蒙特卡洛算法

蒙特卡洛术语的含义。

- 任何具有显著随机成分的估计方法。

MC的定义

✓ 蒙特卡洛方法是基于**经验抽样的**工作

✓ 经验抽样是指从实际或模拟的与环境的**互动**中获得的状态、行动和奖励的**序列。**

- 在许多情况下，在真正的尝试之前是不可预测的。

✓ 只生成样本过渡，而不是像动态编程中那样生成所有的完整概率分布。

蒙特卡洛算法

MC的定义

- ✓ **蒙特卡洛**方法是**解决**基于平均化样本回报的**强化学习问题的方法**。
- ✓ 因此，蒙特卡洛方法在**逐集的意义上**可以是**增量的**，但在逐步（在线）的意义上不是。

马尔科夫决策过程(MDPs)

什么是MDP的定义？

- ✓ 有一组状态S，行动A
- ✓ 奖励模式

$$\cancel{R_t = R(S_t, A_t)}$$

- ✓ 过渡模式。

$$\cancel{P(S_t \mid S_{t-1}, a_{t-1})}$$

- ✓ 折扣系数（γ），它在[0,1]之间。
- ✓ 地平线（h）（集，或时间步骤）。

我们在MC算法中没有这些算法

蒙特卡洛算法(RL)

假设

- ✓ 我们假设经验被划分为若干个事件（最终终止）。
- ✓ **在互动的基础上**学习和**更新政策**
- ✓ 代理人**不知道**环境和环境模型的情况
- ✓ 代理人**需要互动**和尝试

基于模型的学习与无模型的学习

模型

✓ RL中的模型严格来说是指代理是否**通过环境**行动来使用**学习**。

基于模型的学习

- 在基于模型的RL中，代理可以获得环境的模型。
- 其优点是，这使得代理人可以通过提前预测来提前计划

无模型的学习

- 在无模型学习中，代理人没有机会获得有关的模型。

环境（**预测**状态转换**和执行以获得**奖励）。

蒙特卡洛算法

所有蒙特卡洛方法的基本理念

- ✓ 蒙特卡洛方法的目的是学习**状态值**或**行动值**功能（根据我们使用的方法）。
- ✓ 访问该州后观察到的回报的简单平均值
- ✓ 观察到更多的回报，平均数应该收敛到预期值。
- ✓ 具有终端状态问题的偶发性RL问题的蒙特卡洛方法

快速提醒minder

状态值的功能是什么。

- 一个国家的价值就是预期收益

行动-价值函数是什么。

- 状态下的每个行动的价值是预期收益
- 被称为Q值

预期的回报是什么。

- 从该状态开始的未来预期累积折现报酬

蒙特卡洛算法-探索的开始

不同版本的蒙特卡洛算法

- ✓ **首访法**（MC预测法）
  - 第一次访问的MC方法估计$v\pi(s)$为**第一次访问s状态后的平均回报。**
  - 基于状态值（价值函数的估计）。
- ✓ **每次访问的方法**（MC预测）
  - 每次访问的MC方法估计$v\pi(s)$为**所有访问状态s的平均值。**
  - 基于状态值（价值函数的估计）。
- ✓ **探索启动方法**（MC控制）。
  - 基于国家行动对
- ✓ **贪婪的MC Epsilon** (MC控制)
  - 没有探索的开始
- ✓ …

蒙特卡洛算法-探索的开始

RL和Monte Carlo方法中控制和预测之间的差异

RL 预测

- ✓ RL中的预测任务是指当**策略$\boldsymbol{\pi}$已经给出**，我们需要衡量它的表现如何。
- ✓ 这意味着行动已经固定（π（a，s））。
- ✓ 只预测任何状态下的预期总回报

RL控制

- ✓ 在RL的控制任务中，政策不是固定的，目标是找到最佳政策。
- ✓ 这意味着找到$\pi(a, s)$ 使预期报酬最大化

蒙特卡洛算法

首次访问法MC预测

输入：政策π，事件的数量n_ep

输出：值函数V（如果n_ep足够大$V \approx v\pi$；意味着通过采样预测是准确的）初始化。对于所有的

$s \in S$，返回（s）=0

初始化：$N(s)$=0，适用于所有$s \in S$的情况。

for episode e = 1 **to** n_ep **do**

使用政策$s0$，$a0$，$r0$），（$s1$，$a1$，$r1$），...，（$s_{T\text{-}1}$，aT-1，r_T）产生一个情节（π（T步骤数））。

$G = 0$

对于时间步长t = T - 1**到0做**（每一集的状态）。

$G = \gamma G + Rt{+}1$

如果搜索(St)==false，**那么**(从头开始搜索，看St是否在该集不存在了(完成→第一次出现St) Returns(St) = Returns(St) + G (每集将收集每个St的回报)

$N(St) = N(St) + 1$

end if

( ) **end for**

结束

V s = Returns$(s)/N(s)$，对于所有$s \in S$而言

返回 (V)

函数 bool search (state St)

对于$state$ = 0 to t - 1 in episode do

ifepisode (state) == St 则**返回**（true）（在生成的情节步骤中逐一搜索St）。

其他情况下，返回（假）。

蒙特卡洛算法--首次访问(预测)

例子

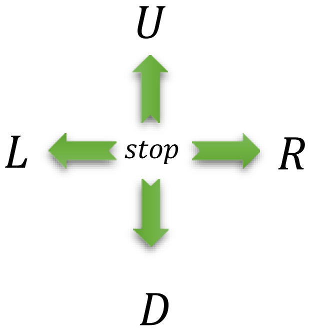
产生一个插曲（T=3）。

$E1: (s4, U, r4) \rightarrow (s1, R, s) \rightarrow (s2, L, r2) \rightarrow (s1, s, s)$
$E2: (s3, U, r3) \rightarrow (s0, sto, r0)$
$E3: (s5, L, r5) \rightarrow (s4, U, r4) \rightarrow (s1, R, r1) \rightarrow (s, s, s)$

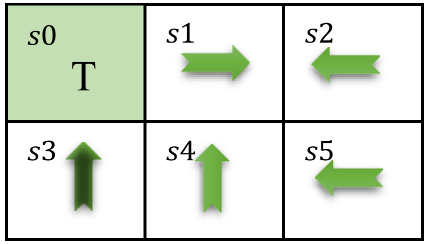
终端

环境

$s0$	$s1$	$s2$
$s3$	$s4$	$s5$



鉴于
政策 π



奖励

+50	-1	-3
-1	-2	-4

蒙特卡洛算法--首次访问(预测)

例子

剧集。

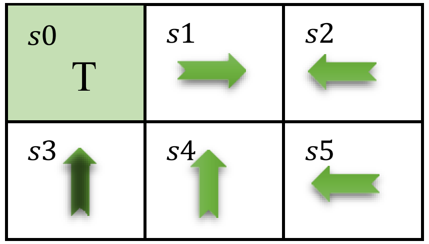
$E1: (s4, U, r4) \rightarrow (s1, R, s) \rightarrow (s2, L, r2) \rightarrow (s1, s, s)$

$E2: (s3, U, r3) \rightarrow (s0, sto, r0)$

$E3: (s5, L, r5) \rightarrow (s4, U, r4) \rightarrow (s1, R, r1) \rightarrow (s, s, s)$

鉴于

政策 π



奖励

+50	-1	-3
-1	-2	-4

第一次访问蒙特卡洛。

✓ 首次访问估计值(Value | State: st)为首次访问$\boldsymbol{st}$州后的平均回报。

每一次访问蒙特卡洛。

✓ 它将(价值|国家：st)估计为每次访问国家$\boldsymbol{st}$的平均回报。

蒙特卡洛算法--首次访问(预测)

例子

$G = \gamma G + Rt{+}1$

鉴于
政策 π

$s0$ T	$s1$ →	$s2$ ←
$s3$ ↑	$s4$ ↑	$s5$ ←

剧集。

$E1: (s4, U, r4) \rightarrow (s, \boldsymbol{R}, r) \rightarrow (s2, L, r2) \rightarrow (s1, s, r1)$

$E2: (s3, U, r3) \rightarrow (s0, sto, r0)$

$E3: (s5, L, r5) \rightarrow (s4, U, r4) \rightarrow (s, \boldsymbol{R}, r) \rightarrow (s2, s, r2)$

奖励

+50	-1	-3
-1	-2	-4

首次访问蒙特卡洛。

✓ 将第一次访问（St）后的所有奖励加起来。**这里$s1$**（为简单起见γ=1).

$_1 E: G = ($ E_2 :Null 对于(0 × 1) + (-3))+ (-1) $\rightarrow G = -4$

E_3 ： $G = -3$

$$V(s1) = \frac{-4 - 3}{2} = -4.5$$

价值函数V

T	**-4.5**	0
0	0	0

注意：如果某一集没有出现过 **s1**，它不会被考虑在平均数中

蒙特卡洛算法--首次访问(预测)

例子

剧集。

$E1: (s4, U, r4) \rightarrow (s1, R, r1) \rightarrow (s, \mathbf{L}, r) \rightarrow (s1, s, r1)$

$E2: (s3, U, r3) \rightarrow (s0, sto, r0)$

$E3: (s5, L, r5) \rightarrow (s4, U, r4) \rightarrow (s1, R, r1) \rightarrow (s, r, \boldsymbol{stop})$

鉴于

政策 π

$s0$ T	$s1$ →	$s2$ ←
$s3$ ↑	$s4$ ↑	$s5$ ←

奖励

+50	-1	-3
-1	-2	-4

第一次访问蒙特卡洛。

✓ 计算 $\boldsymbol{s2}$

$\rightarrow G = \text{-}1$

对于情节E_1 : $G=$ (-1)

对于情节E_2。 无

对于情节E_3 : $G = 0 \rightarrow G = 0$

0

蒙特卡洛算法-首次访问(预测)

$$V_{(s2)} = \frac{-1 + 0}{2} = -0.5$$

价值函数V

T	-4.5	**-0.5**
0	0	0

蒙特卡洛算法--首次访问(预测)

例子

剧集。

$E1$: $(s4, U, r4) \rightarrow (s1, R, s) \rightarrow (s2, L, r2) \rightarrow (s1, s, s)$

$E2$: $(\boldsymbol{s3}, \boldsymbol{U}, \boldsymbol{r3}) \rightarrow (s0, op, r0)$

$E3$: $(s5, L, r5) \rightarrow (s4, U, r4) \rightarrow (s1, R, r1) \rightarrow (s, s, s)$

首次访问蒙特卡洛。

✓ $\boldsymbol{s3}$的计算结果

对于情节E_1。 Null

For episode E_2 : $G = 50$

For episode E_3 :无

50

$(s3)$

鉴于

政策 π

$s0$ T	$s1$ →	$s2$ ←
$s3$ ↑	$s4$ ↑	$s5$ ←

奖励

+50	-1	-3
-1	-2	-4

价值函数V

T	-4.5	-0.5
50	0	0

蒙特卡洛算法--首次访问(预测)

蒙特卡洛算法--首次访问(预测)

例子

剧集。

$E1: (\boldsymbol{s4}, \boldsymbol{U}, \boldsymbol{r4}) \rightarrow (s, R, r) \rightarrow (s2, L, r2) \rightarrow (s, r, r)$
$E2: (s3, U, r3) \rightarrow (s0, sto, r0)$
$E3: (s5, L, r5) \rightarrow (s, \boldsymbol{U}, r) \rightarrow (s1, R, r1) \rightarrow (s2, s, r2)$

终端

鉴于政策 π

$s0$ T	$s1$	$s2$
$s3$	$s4$	$s5$

奖励

+50	-1	-3
-1	-2	-4

第一次访问蒙特卡洛。

✓ 计算 $\boldsymbol{s4}$

对于情节E_1 ： $G=$ (-1) + (-3) + (-1) $\rightarrow G =$ -5

对于情节E_2。 无

对于情节E_3 : $G=$ (-1) + (-3) $\rightarrow G =$ -4

-4 - 5

价值函数V

T	-4.5	-0.5
50	**-4.5**	0

蒙特卡洛算法 -- 首次访问(预测)

Monte-Carlo Algorithm - First Visit (Prediction)

蒙特卡洛算法--首次访问(预测)

例子

剧集。

$E1: (s4, U, r4) \rightarrow (s1, R, s) \rightarrow (s2, L, r2) \rightarrow (s1, s, s)$
$E2: (s3, U, r3) \rightarrow (s0, sto, r0)$
$E3: (s, \mathbf{L}, r) \rightarrow (s4, U, r4) \rightarrow (s1, R, r1) \rightarrow (s2, s, r2)$

终端

鉴于
政策 π

$s0$ T	$s1$	$s2$
$s3$	$s4$	$s5$

价值函数V
奖励

T	-4.5	-0.5
50	-4.5	**-6**

-6

首次访问蒙特卡洛。

(5)

✓ 计算 $\boldsymbol{s5}$

对于情节E_1。 无

对于情节E_2。 无

对于情节E_3 : $G=$ (-2) + (-1) + (-3)

-1	-2	-4

蒙特卡洛算法--首次访问(预测)

$\rightarrow G = \text{-}6$

$$V = \frac{}{1} = \text{-}6$$

蒙特卡洛算法，每次访问（预测）。

每次访问的方法MC预测

输入：政策π，事件的数量n_ep

输出：值函数V（如果n_ep足够大$V \approx v\pi$；意味着通过采样预测是准确的）初始化。对于所有的$s \in S$，返回（s）=0

初始化：$N(s)$=0，适用于所有$s \in S$的情况。

for episode e = 1 **to** n_ep **do**

使用政策$s0$，$a0$，$r0$），（$s1$，$a1$，$r1$），…，（sT-1，aT-1，rT）产生一个情节（π（T步骤数））。

$G = 0$

对于时间步长t = T - 1**到**0的插曲e**做**（插曲的每个状态）。

$G = \gamma G + Rt+1$

Returns(St) = Returns(St) + G (每一集将收集每个St的回报。St可以多次访问，并将其下面的状态添加到情节中，将再次进行汇总)

$N(St)$= $N(St)$+1（每访问一次S t，其访问时间将增加。因此，如果一个状态不在一个情节中，它的计数器将自动不增加)

（）结束时，

结束时

V s = Returns$(s)/N(s)$，对于所有$s \in$ S而言

返回 (V)

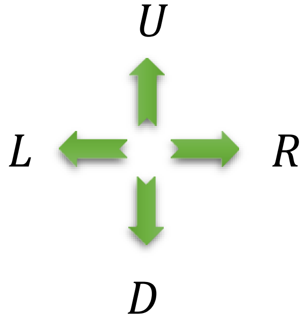
蒙特卡洛算法，每次访问（预测）。

同样的例子

环境

终端

$s0$	$s1$	$s2$
$s3$	$s4$	$s5$

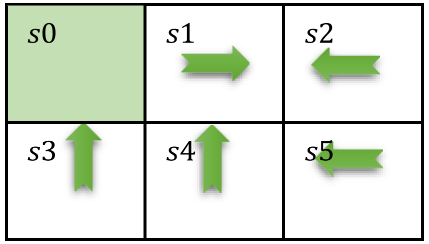


生成一集（$T=4$）。(我们做了3次，看看)

$E1: (s4, U, r4) \rightarrow (s1, R, s) \rightarrow (r, L, s) \rightarrow (s1, R, s) \rightarrow (r, r, r)$

$E2: (s3, U, r3) \rightarrow (s0, sto, r0)$

$E3: (s5, s, r5) \rightarrow (s4, U, r4) \rightarrow (r, R, r) \rightarrow (s2, s, r2) \rightarrow (r, s, s)$



给定政策 π

+50	-1	-3
-1	-2	-4

奖励

第一次访问蒙特卡洛。

- ✓ 首次访问估计值(Value | State: st)为首次访问 st后的平均回报。

每一次访问蒙特卡洛。

- ✓ 它将(价值|国家：st)估计为每次访问国家$\boldsymbol{st}$的平均回报。

蒙特卡洛算法，每次访问（预测）。

例子

$$G = \gamma G + Rt{+}1$$

终端

$s0$ T	$s1$	$s2$
$s3$	$s4$	$s5$

奖励

+50	-1	-3
-1	-2	-4

剧集。

$E1$: $(s4, U, r4) \to (s, \boldsymbol{R}, \boldsymbol{s1}) \to (s2, L, \boldsymbol{r1}) \to (s, \boldsymbol{R}, \boldsymbol{s1}) \to (s2, s, \boldsymbol{r1})$
$E2$: $(s3, U, r3) \to (s0, sto, r0)$
$E3$: $(s5, r, r5) \to (s4, U, r4) \to (\boldsymbol{s1}, \boldsymbol{R}, \boldsymbol{r1}) \to (s2, r, r2) \to (\boldsymbol{s1}, \boldsymbol{stop}, \boldsymbol{r1})$

每一次访问蒙特卡洛。

✓ 将第一次访问（St）后的所有奖励加起来。**这里$s1$**（为简单起见γ= 1).

对于情节$\boldsymbol{E1}$：G=（ (1×0) +(-3)) → (-3) G = -3

。

对于情节$\boldsymbol{E1}$：G = (-3) + (-1) + = 0 (-1) → G = -7

对于情节E_2。 无

$$V(s1) = \frac{(-3) + (-7) + (-4) + (0)}{} = -3.5$$

蒙特卡洛算法，每次访问（预测）。

Monte Carlo Algorithm, Every Visit (Prediction)

对于情节$E3$：$G = (0) \rightarrow G$ 4

对于情节$E3$：$G = (-3) + \rightarrow G = -4$

再次注意：如果一集没有出现 $s1$，它将不会被考虑在平均数中。

鉴于

政策 π

蒙特卡洛算法，每次访问（预测）。

例子

终端

$s0$ T	$s1$	$s2$
$s3$	$s4$	$s5$

奖励

+50	-1	-3
-1	-2	-4

剧集。

$E1$: $(s4, U, r4) \rightarrow (s1, R, r) \rightarrow (\boldsymbol{s2}, \boldsymbol{L}, \boldsymbol{r2}) \rightarrow (s1, L, r) \rightarrow (\boldsymbol{s2}, \boldsymbol{stop}, \boldsymbol{r2})$

$E2$: $(s3, U, r3) \rightarrow (s0, sto, r0)$

$E3$: $(s5, L, r5) \rightarrow (s4, U, r4) \rightarrow (s, R, r) \rightarrow (op, \boldsymbol{L}, r) \rightarrow (s, op, r)$

第一次访问蒙特卡洛。

✓ 计算 $\boldsymbol{s2}$

对于情节E_1 : $G=$ (0) → G = 0

对于情节E_1 ： $G=$ (-1) + (-3) → G = -4

对于情节E_2。 无

价值函数V

0 0

蒙特卡洛算法，每次访问（预测）。

Monte Carlo Algorithm, Every Visit (Prediction)

对于情节$E_3: G = -1 \to G = -1$

$$V(s2) = \frac{0 + (-4) + (-1)}{3} = -1.66$$

鉴于
政策 π

T	**-3.5**	**-1.66**
0	0	0

蒙特卡洛算法，每次访问（预测）。

例子

剧集。

$E1: (s4, U, r4) \rightarrow (s1, R, r1) \rightarrow (s, r, r) \rightarrow (s1, r, r1) \rightarrow (s, s, s)$

$E2: (\boldsymbol{s3}, \boldsymbol{U}, \boldsymbol{r3}) \rightarrow (s0, op, r0)$

$E3: (s5, s, r5) \rightarrow (s4, U, r4) \rightarrow (r, R, r) \rightarrow (s2, s, r2) \rightarrow (r, r, s)$

首次访问蒙特卡洛。

✓ $\boldsymbol{s3}$的计算结果

对于情节E_1。Null

For episode E_2 : $G = 50$ $\quad \frac{50}{1} = 50$

For episode E_3 :无

$V(s3) =$

终端

鉴于政策 π

$s0$ T	$s1$ →	$s2$ ←
$s3$ ↑	$s4$ ↑	$s5$ ←

奖励

+50	-1	-3
-1	-2	-4

价值函数V

T	-3.5	-1.66
50	0	0

蒙特卡洛算法，每次访问（预测）。

例子

剧集。

$E1$: $(\boldsymbol{s4}, \boldsymbol{U}, \boldsymbol{r4}) \rightarrow (s1, R, s) \rightarrow (r, op, op) \rightarrow (s1, op, s) \rightarrow (r, s, r)$

$E2$: $(s3, U, r3) \rightarrow (s0, sto, r0)$

$E3$: $(s5, L, r5) \rightarrow (s, \boldsymbol{U}, r) \rightarrow (op, R, r) \rightarrow (s2, L, r2) \rightarrow (op, op, s)$

终端

$s0$ T	$s1$	$s2$
$s3$	$s4$	$s5$

奖励

+50	-1	-3
-1	-2	-4

第一次访问蒙特卡洛。

✓ 对$\boldsymbol{s4}$的计算结果

对于情节E_1 : $G=$ (-1) + (-3) + (-1) + (-3) $\rightarrow G$ = -8

对于情节E_2。 无

对于情节E_3 : $G=$ (-1) + (-3) + (-1)

$(s4)$

价值函数V

T	-3.5	-1.66
50	**-6.5**	0

0 -8 - 5

Monte-Carlo Algorithm, Every Visit (Prediction)

蒙特卡洛算法，每次访问（预测）。

$\rightarrow G_5$

$V = \frac{\quad}{2} =$ -6.5

鉴于

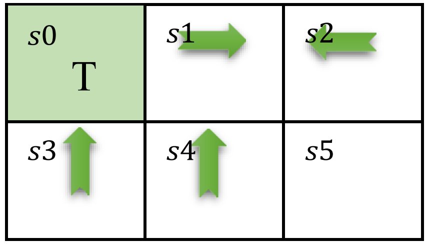
政策 π

蒙特卡洛算法，每次访问（预测）。

例子

终端

鉴于
政策 π



价值函数V 奖励

T	-3.5	-1.66
50	-6.5	**-7**

-7

剧集。

$E1: (s4, U, r4) \rightarrow (s1, R, r1) \rightarrow (s, r, r) \rightarrow (s1, r, r1) \rightarrow (s, s, s)$

$E2: (s3, U, r3) \rightarrow (s0, sto, r0)$

$E3: (s, \mathbf{L}, r) \rightarrow (s4, U, r4) \rightarrow (op, R, s) \rightarrow (s2, L, r2) \rightarrow (op, r, s)$

第一次访问蒙特卡洛。

✓ 计算 $\boldsymbol{s5}$ (5)

对于情节E_1。 无

对于情节E_2。 无

对于情节E_3 : $G=$ (-2) + (-1) + (-3) + (-1)

-1	-2	-4

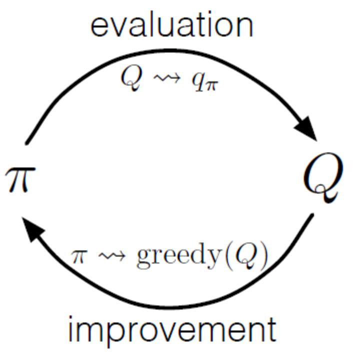
蒙特卡洛算法，每次访问（预测）。

$$\rightarrow G = \text{-}7$$

$$V = \frac{\quad}{1} = \text{-}7$$

蒙特卡洛算法-探索的开始（控制）。

探索起步法



蒙特卡洛探索启动（MC-ES），用于估计$\pi \approx \pi*$。

Initialize:
  $\pi(s) \in \mathcal{A}(s)$ (arbitrarily), for all $s \in \mathcal{S}$
  $Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
  $Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Loop forever (for each episode):
  Choose $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ randomly such that all pairs have probability > 0
  Generate an episode from S_0, A_0, following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
  $G \leftarrow 0$
  Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
    $G \leftarrow \gamma G + R_{t+1}$
    Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:
      Append G to $Returns(S_t, A_t)$
      $Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$
      $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$

蒙特卡洛算法-探索的开始（控制）。

按照步骤进行

a) 为Q值指定任意值，政策先行

重复直到收敛（不改变政策）--最佳政策 π^*。

b) 选择环境中的随机状态s和行动a对

c) 通过政策π产生一个插曲（T步）。

- 从选定的状态s开始，运行选定的行动 a

d) 对于每个状态-行动对，计算折现的回报，并创建列表

- 返回是只计算每一集里的所有下一个状态

$$v_{\pi}(s) = E_{\pi}[G(s)], G = \gamma g + R_{t+1}$$

e) 对列表中的值进行平均，并更新状态-行动对的Q值

$$Q(si, aj) = AVG(List(si, aj))$$

f) 更新政策

$$\pi(s) = \arg\max_{a} Q(s, a)$$

蒙特卡洛探索启动算法（例）。

环境路径规划实例

问题：一个机器人需要从起点走到目标点

奖励。目标+100，其他每一步-1

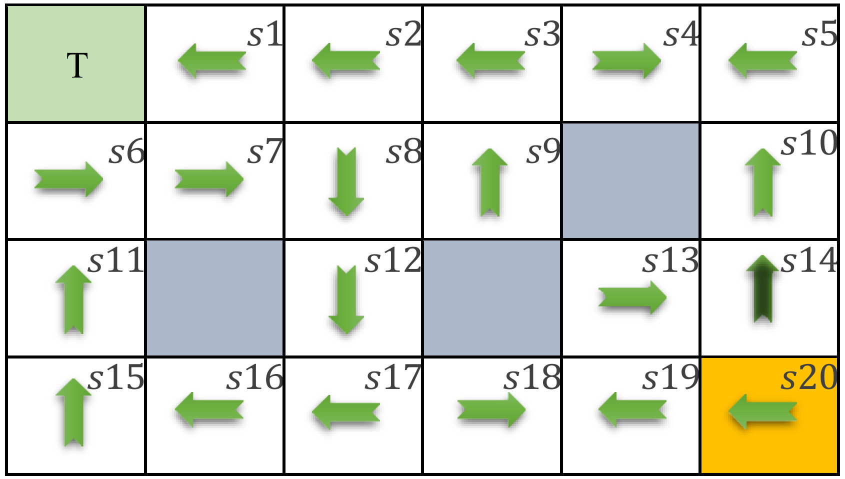
Q值：为简单起见，全部为零

目标	$s0$	$s1$	$s2$	$s3$	$s4$	$s5$
	$s6$	$s7$	$s8$	$s9$		$s10$
	$s11$		$s12$		$s13$	$s14$
	$s15$	$s16$	$s17$	$s18$	$s19$	$s20$

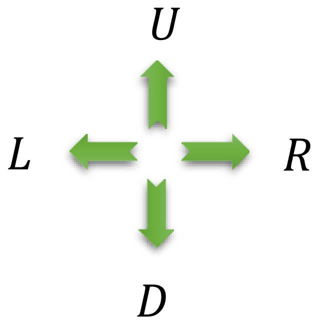
开始

蒙特卡洛算法-探索起点（例）。

a) 我们首先为Q-值分配任意的政策（上、下、左、右）和数值。



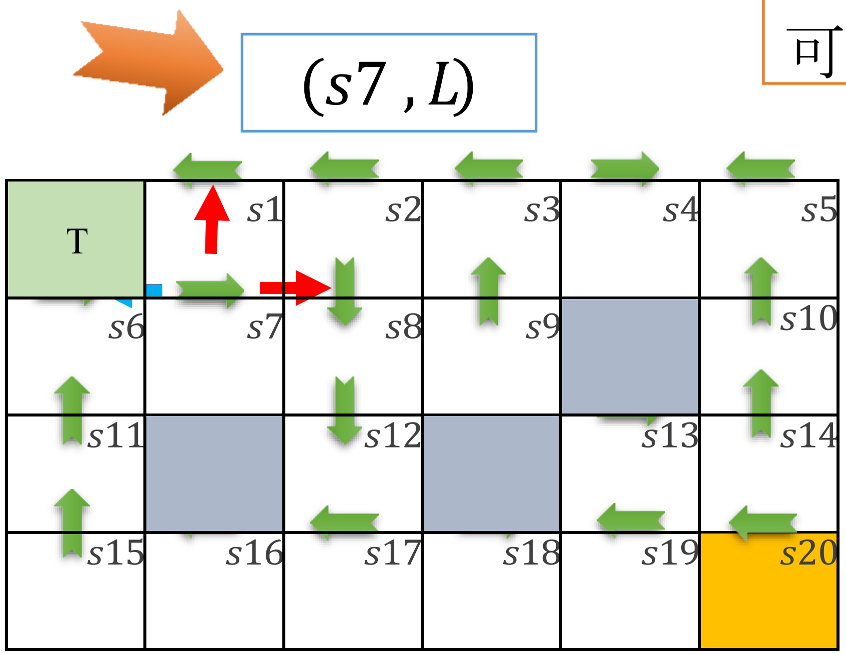
任意的政策



蒙特卡洛算法-探索起点（例）。

b) 选择 随机 状态 s 和 随机 行动 a 对 在 环境中的随机行动

注意：只考虑每个状态下的可能行动



蒙特卡洛算法-探索起点（例）。

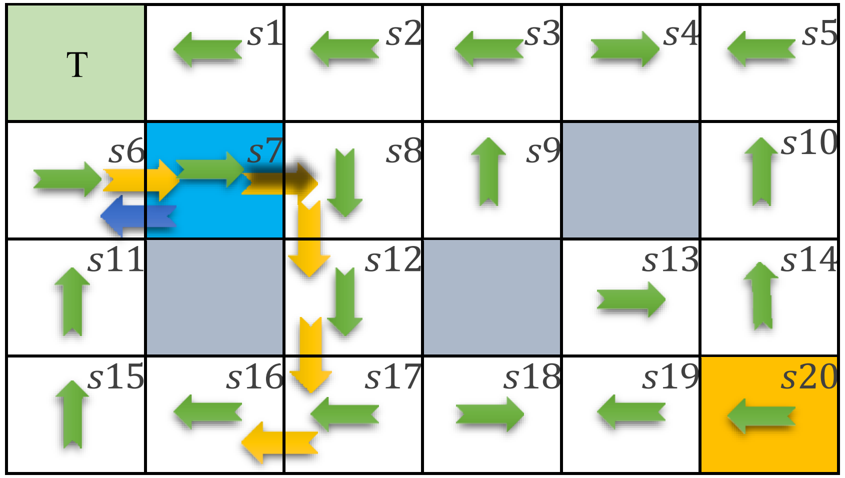
c) 通过政策π产生一个情节（T个步骤，这里是6个）。

➤从选定的随机状态s开始，选择选定的随机动作a。

首先

E1 产生插曲（1）。

$(s7, L)$



$$(s_7, L) \rightarrow (s6, R) \rightarrow (s7, R) \rightarrow (s8, D) \rightarrow (s12, D) \rightarrow (s17, L) \rightarrow (s16, L)$$

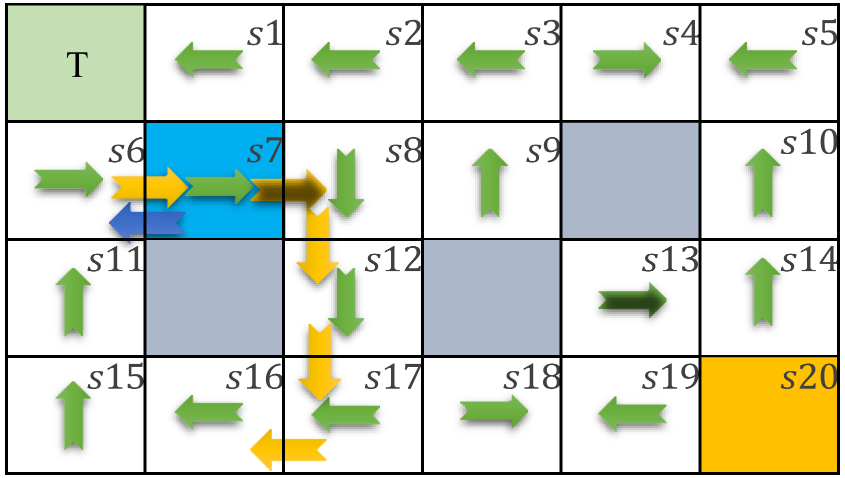
蒙特卡洛算法-探索起点（例）

d) 对于每个状态-行动对，计算折现的回报，并创建列表。

➢ 返回是只计算每一集里的所有下一个状态

折扣后的退货。 $\boxed{G = \gamma G + Rt{+}1}$ $\gamma = 0.9$

$(s_7, L) \rightarrow (s6, R) \rightarrow (s7, R) \rightarrow (s8, D) \rightarrow (s12, D) \rightarrow (s17, L) \rightarrow (s16, L)$



$(s17, L)$ (0.9×0) $=- 1 = -1$

$(s12, \ ) = (0.9 \times -1- 1) = -1.9$

$(D \quad)$ $(\quad)$

$(s8, D)$ $(0.9 \times -1.9-)1 = -2.71$

$(s7, R)$ $(0.9 \times -2.71 \quad)- 1 = -3.439$

$(s6, R)$ (0.9×-3.349) $- 1 = -4.095$

$s7, L$ 0.9×-4.095 $- 1 = -4.685$

Returns_list(s)$_1$, L)=[-1]

7

Returns_list($s12$, D)=[-1.9]

Returns_list(s_8 , D)=[-2.71]

Returns_list(s_7 , R)=[-3.439]

Returns_list(s_6 , R)=[-4.095]

Returns_list(s_7 , L) =[-4.685]

蒙特卡洛算法-探索起点（例）

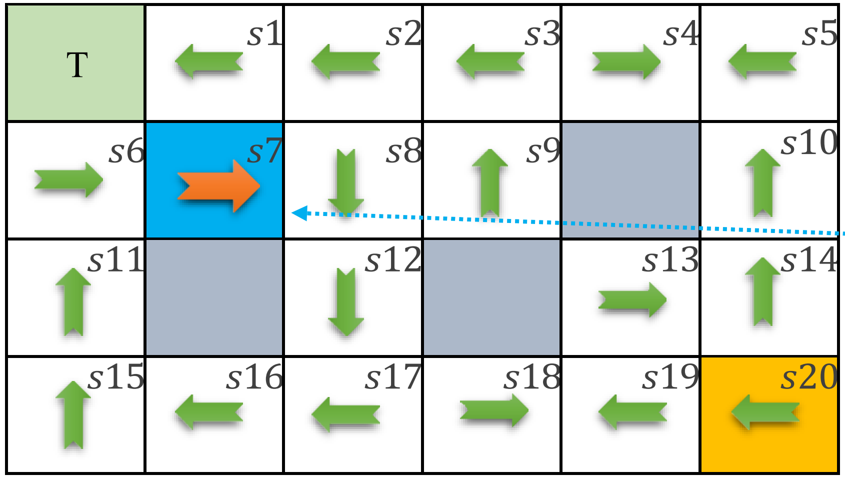
注意：由于我们没有为最后一个状态运行动作，所以不需要包括。

蒙特卡洛算法-探索起点（例）。

f) 更新政策

$$\pi(s) = \arg\max Q(s, a)$$

列表中的平均数。



更新政策

$Q(s_{17}, L)$=[-1]
$Q(s_{12}, L)$=[-1.9]
$Q(s_8, D)$=[-2.71]
$\boldsymbol{Q(s7, R)}$**=[-3.439]**
$Q(s_6, R)$=[-4.095]
$\boldsymbol{Q(s7, L)}$**=[-4.685]**

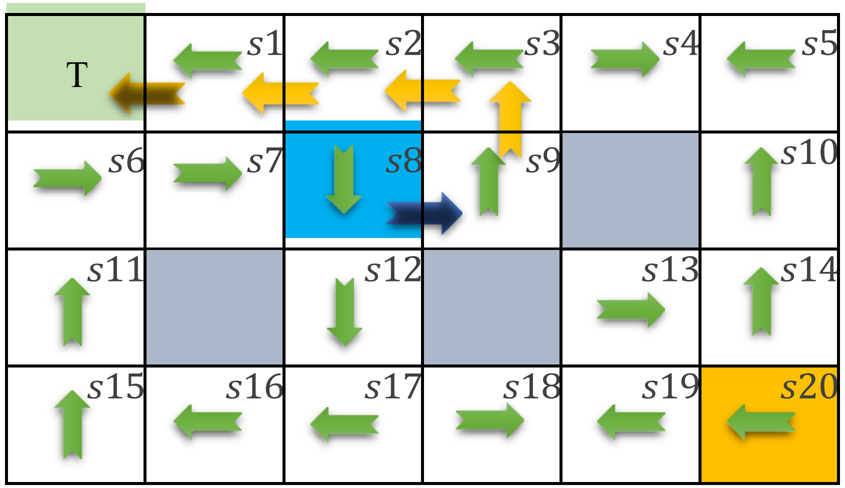
Returns_list($s17$, L)=[-1]。
Returns_list($s12$, D)=[-1.9]。
Returns_list(s_8, D)=[-2.71]
Returns_list(s_7, R)=[-3.439]
Returns_list(s_6, R)=[-4.095]
Returns_list(s_7, L)=[-4.685]

蒙特卡洛算法-探索起点（例）。

E2 通过政策π产生另一集（T步，这里是6步）。

随机 → $(s8\,,R)$

产生插曲（2）。

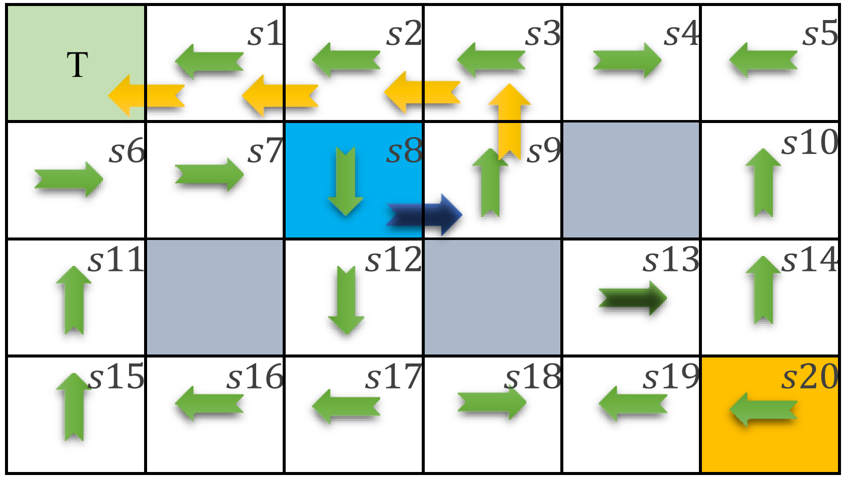


$(\boldsymbol{s8}\,, \boldsymbol{R}) \rightarrow (s9\,, U) \rightarrow (s3\,, L) \rightarrow (s2\,, L) \rightarrow (s1\,, L) \rightarrow ({}_0s\,, stop)$

蒙特卡洛算法-探索起点（例）

折扣后的退货。 $\boxed{G = \gamma G + R_{t+1}}$ $\gamma = 0.9$

$$(s_8, \boldsymbol{R}) \rightarrow (s_9, U) \rightarrow (s_3, L) \rightarrow (s_2, L) \rightarrow (s_1, L) \rightarrow (s_0, stop)$$



$(s_1, L) = (0.9 \times 0 + 100) = 100$

$(s_2, L) = (0.9 \times 100 - 1) = 89$

$(s_3, L) = (0.9 \times 89) - 1 = 79.1$

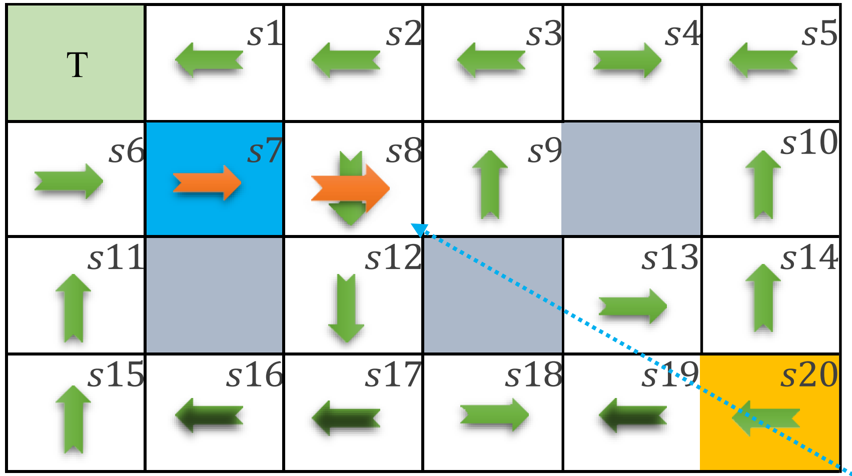
$(s_9, U) = (0.9 \times 79.1) - 1 = 70.19$

$(s_8, R) = (0.9 \times 70.19) - 1 = 62.171$

Returns_list($s17$, L)=[-1]
Returns_list($s12$, D)=[-1.9]
Returns_list(s_8, D)=[-2.71]
Returns_list(s_7, R)=[-3.439]
Returns_list(s_6, R)=[-4.095]
Returns_list(s_7, L)=[-4.685]
Returns_list($s1$, L)=[100]。
Returns_list($s2$, L)=[89]
Returns_list($s3$, L)=[79. 1]
Returns_list($s9$, U)=[70. 19]
Returns_list(s_8, R) =[62. 171]

f) 更新政策

$$\pi(s) = \arg\max Q(s, a)$$



更新政策 ← Q 值 ← 列表中的平均数。

$Q(s_{17}, L)$=[-1]

$Q(s_{12}, D)$=[-1.9]
$Q(s_8, D)$=[-2.71]
$Q(s_7, R)$=[-3.439]
$Q(s_6, R)$=[-4.095]
$Q(s_7, L)$=[-4.685]
$Q(s_1, L)$=[100]
$Q(s_2, L)$=[89]
$Q(s_3, L)$=[79.1]
$Q(s_9, U)$=[70.19]
$Q(s_8, R)$=[62.171]

Returns_list($s17$, L)=[-1]

。

Returns_list($s12$, D)=[-1.9]
Returns_list(s_8, D)=[-2.71]
Returns_list(s_7, R)=[-3.439]
Returns_list(s_6, R)=[-4.095]
Returns_list(s_7, L)=[-4.685]
Returns_list($s1$, L)=[100]
Returns_list($s2$, L)=[89]
Returns_list($s3$, L)=[79.1]
Returns_list($s9$, U) =[70.19]

蒙特卡洛算法 探索起点（例）。

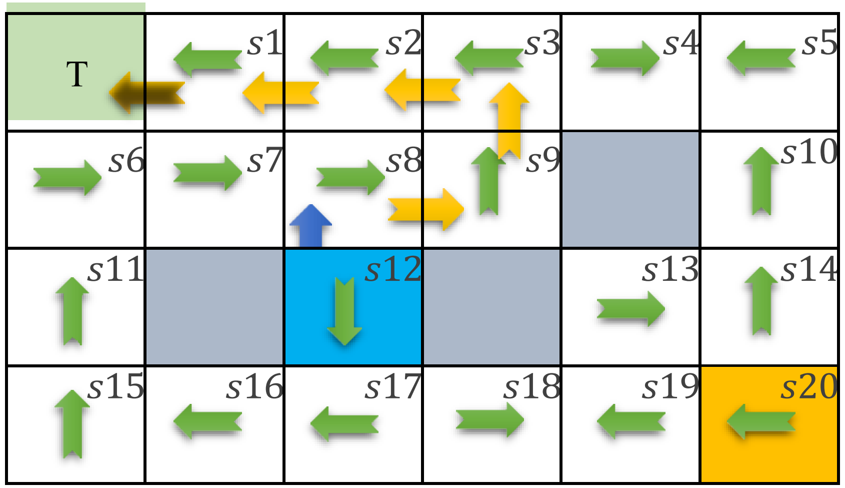
Returns_list(s8, R) = [62. 171]

蒙特卡洛算法-探索起点（例）。

E3 通过政策π产生另一集（T步，这里是6步）。

$(s12\,,U)$

产生插曲（3）。

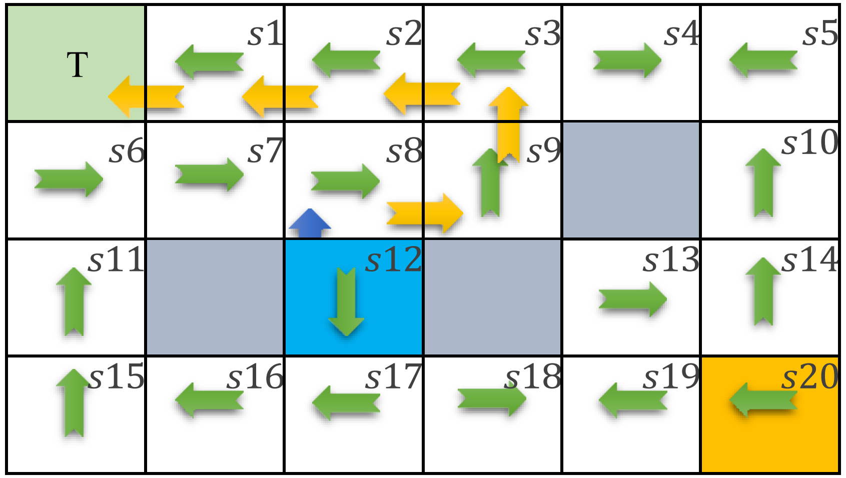


$$(s_{12}\,,\boldsymbol{U}) \rightarrow (s8\,,R) \rightarrow (s9\,,U) \rightarrow (s3\,,L) \rightarrow (s2\,,L) \rightarrow (s1\,,L) \rightarrow ({}_0s\,,stop)$$

蒙特卡洛算法-探索起点（例）。

折扣后的退货。 $\boxed{G = \gamma G + Rt{+}1}$ $\gamma = 0.9$

$(s_{12}, \boldsymbol{U}) \rightarrow (s8, R) \rightarrow (s9, U) \rightarrow (s3, L) \rightarrow (s2, L) \rightarrow (s1, L) \rightarrow (_0 s, stop)$



$(s1, L) = (0.9 \times 0 +)\ 100 = 100$

$(s2, L) = (0.9 \times 100 - 1) = 89$

$(s3, L) = -(1\ 0.9 \times 89)$

$(s9, U) = (0.9 \times 79.1\ ) - 1 = 70.19$

$(s8, R) = -(1\ 0.9 \times 70.19)$

$(s12, U) = -(1\ 0.9 \times 62.171)$

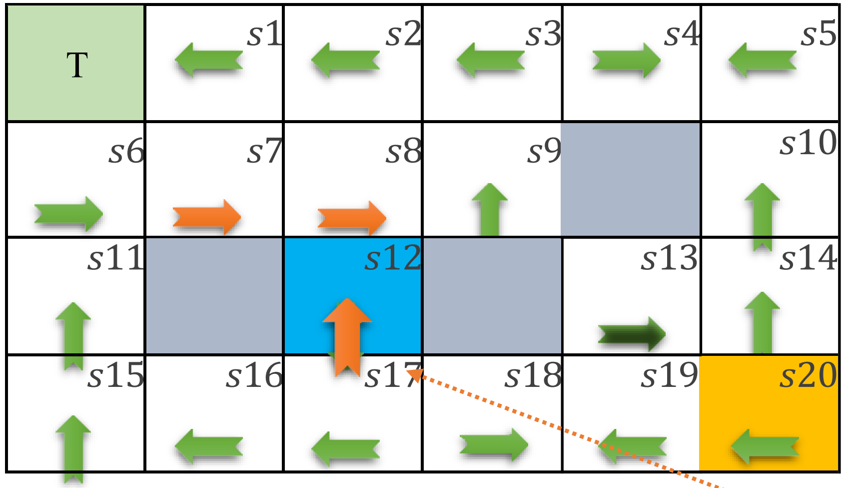
Returns_list(s17, L)=[-1]
Returns_list(s12, D)=[-1.9]
Returns_list(s_8, D)=[-2.71]
Returns_list(s_7, R)=[-3.439]
Returns_list(s_6, R)=[-4.095"
Return_list(s_7, L)=[-4.685]
Return_list(s_1, L)=[100, **100**]
Returns_list(s_2, L)=[89, **89**]
Returns_list(s)$_3$, L)=[79.1, **79.1**]
Returns_list(s_9, U)=[70.19, **70.19**]
Returns_list(s_8, R)=[62.171, **62. 171**]
Returns_list(s12, U)=[54. 953]

蒙特卡洛算法-探索起点（例）

Monte Carlo Algorithm Exploring Starts (Example)

f) 更新政策

$$\pi(s) = argmax\, Q(s, a)$$



更新政策

Q(s_17, L)=[-1]。

Q(s_12, D)=[-1.9]。

Q(s_8, D)=[-2.71]

Q(s_6, R)=[-4.095]

Q(s_7, R)=[-3.439]

Q(s_7, L)=[-4.685]

Q(s_1, L)=[100]。

Q(s_2, L) =[89]

Q(s_3, L)=[79.1]

Q(s9), U)=[70.19]

Q(s8, R)=[62.171]

Q(s12, U)=[54.95]。

列表中的平均数。

Returns_list(s_{17}, L)=[-1]

Returns_list(s_{12}, D) =[-1.9]

Returns_list(s_8, D)=[-2.71]

Returns_list(s_7, R)=[-3.439]

Returns_list(s_6, R)=[-4.095]

Returns_list(s_7, L) =[-4.685]

Returns_list(s1, L)=[100, **100**]。

Returns_list(s_2, L)=[89, **89**]

Returns_list(s_3, L) =[79.1, **79.1**]

Returns_list(s_9, U)=[70.19, **70.19**]

Returns_list(s_8, R)=[62.171, **62.171**]

蒙特卡洛算法-探索起点（例）。

Returns_list(s , U)=[54. , U)=[54. 953]

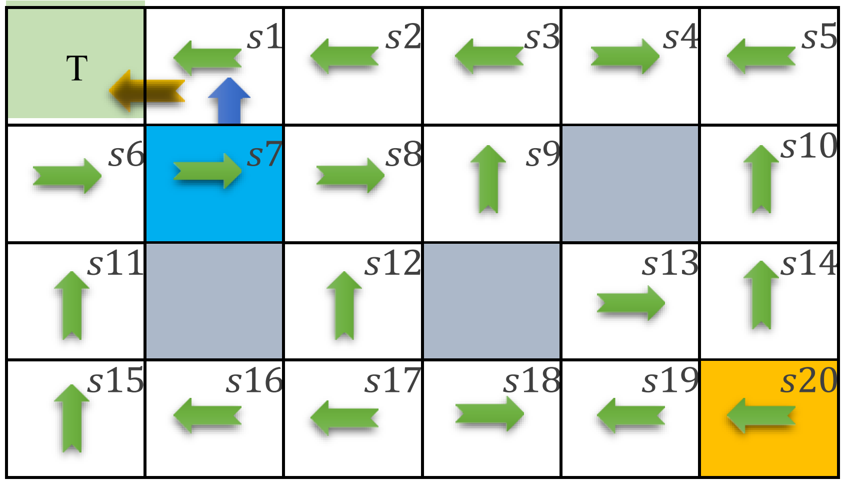
12

蒙特卡洛算法-探索起点（例）。

E4 通过政策π产生另一集（n步，这里是6步）。

$$\boxed{(s7\,,U)}$$

生成一集（4）。

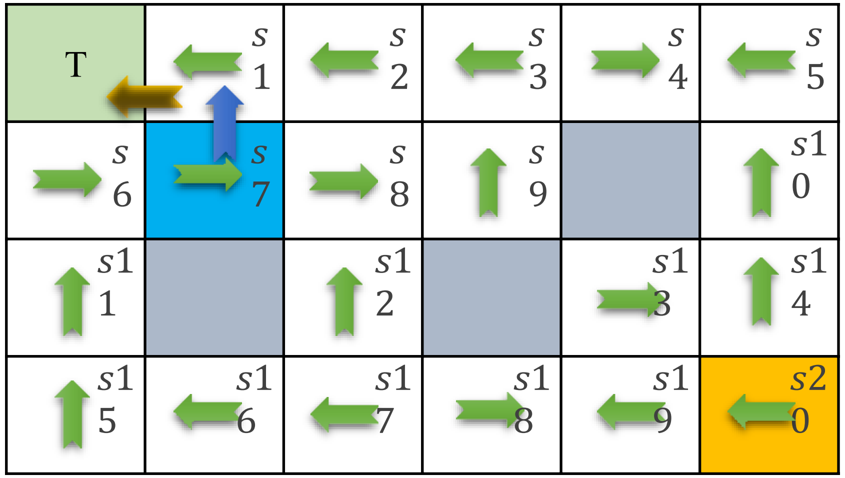


$$(s_7\,, U) \rightarrow (s1\,, L) \rightarrow ({}_0s\,, stop)$$

蒙特卡洛算法-探索起点（例）。

折扣后的退货。 $\boxed{G = \gamma G + Rt{+}1}$ $\gamma = 0.9$

$(s_7, \boldsymbol{U}) \rightarrow (s1, L) \rightarrow ({}_0s, stop)$



$(s1, L) = 100$

$(s7, U) = -1 + (0.9 \times 100)$

Returns_list(s_17, L) =[-1]
Returns_list(s_12, D) =[-1.9]
Returns_list(s_8, D) =[-2.71]
Returns_list(s_7, R) =[-3.439]
Returns_list(s_6, R)=[-4.095]
Returns_list(s_7, L)=[-4.685] Returns_list(s_1 , L) =[100, 100, **100**]
Returns_list(s_2 , L)=[89, 89]
Returns_list(s_3 , L)=[79.1, 79.1]
Returns_list(s_9 , U)=[70.19, 70.19]
Returns_list(s_8 , R)=[62.171, 62.171]
Returns_list(s12, U)=[54.953]
Returns_list(s7, U) =[89]

蒙特卡洛算法-探索起点（例）。

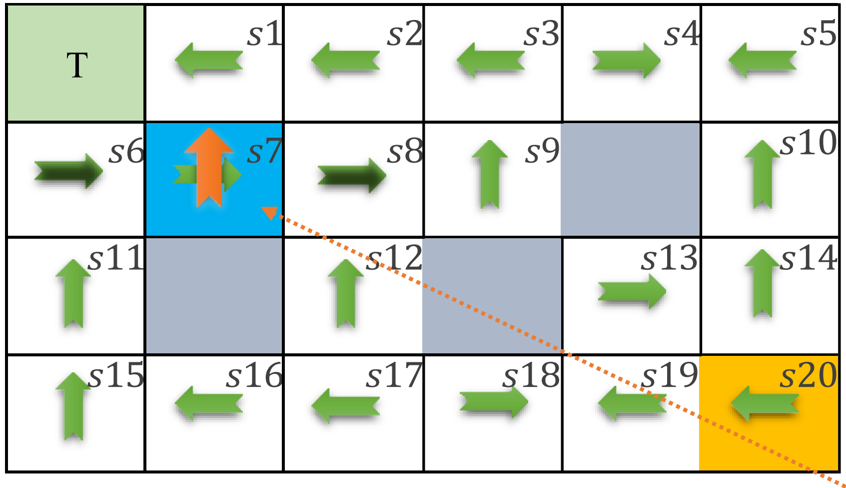
Monte Carlo Algorithm-Exploring Starts (Example)

f) 更新政策

$$\pi(s) = argmax\, Q(s, a)$$

列表中的平均数。

更新政策



Q(s_{17} , L)=[-1]
Q(s_{12} , D)=[-1.9]
Q(s_8 , D)=[-2.71]
Q(s_7 , R)=[-3.439]
Q(s_6 , R)=[-4.095]
Q(s_7 , L)=[-4.685]
Q(s_1 , L)=[100]
Q(s_2 , L)=[89]
Q(s_3 , L)=[79.1]
Q(s_9 , U)=[70.19]。
Q(s_8 , R）=[62.171]
Q(s_{12} ,
Q(*s7*, *U*)=[89]。

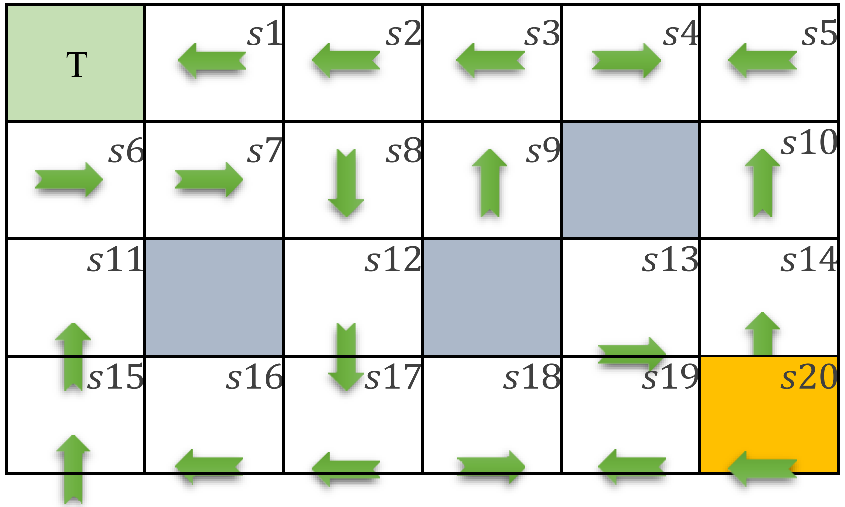
Returns_list($s17$, L)=[-1]
Returns_list($s12$, D)=[-1.9]
Returns_list(s_8 , D)=[-2.71]
Returns_list(s_7 , R)=[-3.439]
Returns_list(s_6 , R)=[-4.095]
Returns_list(s_7 , L) =[-4.685]
Returns_list(s_1 , L) =[100, 100, **100**]
Returns_list(s_2 , L)=[89, 89]
Returns_list(s_3, L)=[79.1, 79.1]
Returns_list($s9$, U)=[70.19, 70.19]。

蒙特卡洛算法-探索起点（例）

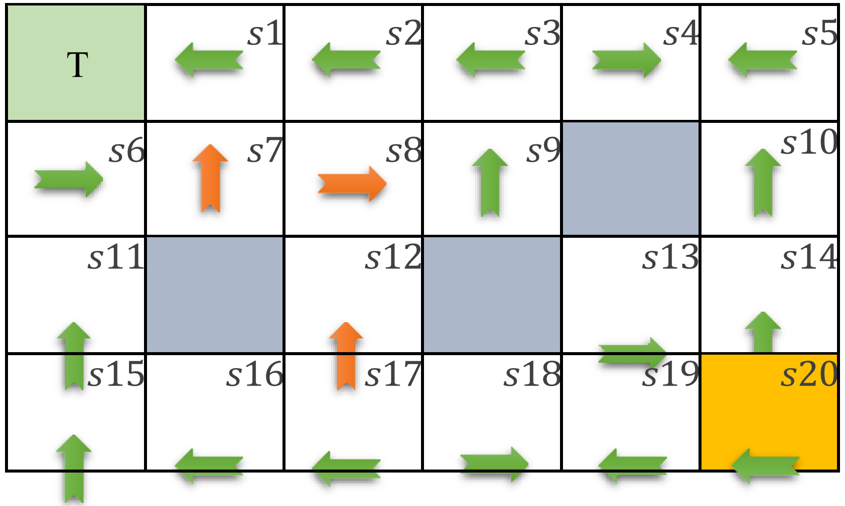
[54.953]

Returns_list(s_8 , R)=[62.171, 62.171]
Returns_list($s12$, U)=[54.953]
Returns_list($\boldsymbol{s7}$, $\boldsymbol{U}$) =[89]

蒙特卡洛算法-探索起点（例）。



任意的政策



4集后的更新政策

MC Epsilon Greedy算法（无探索性启动）。(控制)

理念

- ✓ 在一些问题中，我们无法计算所有的边缘情况。
  - ➢ 例如，在一个应用程序中，重设环境始终是一个问题。回到一个状态，而不是随机的一个状态!
- ✓ 在这种情况下，MC探索式启动就变得不可行了!

解决方案

- ✓ 我们**消除了**从MC探索开始的所有**起点的随机选择**。
- ✓ 有时通过**Epsilon-Greedy**技术**应用随机政策**

管委会的做法

问题

为了更新价值和政策，我们需要**等到剧集结束**。

还有什么更好的主意吗？

- ✓ 解决方案是时空差异（TD） RL

归纳

- ✓ 我们讨论了基于模型的学习和无模型的学习
- ✓ 我们讨论了状态-价值函数和状态-行动函数
- ✓ 我们宣布了首次访问法MC预测的例子
- ✓ 我们宣布了 "每次访问法MC预测"。
- ✓ 我们宣布了蒙特卡洛探索启动（MC-ES）。
- ✓ 我们知道MC Epsilon的贪婪（没有探索性的启动）。