强化学习

Python编程实验三

MC / TD

Random Walk

实验回顾

实验一

马尔科夫奖励过程（MRP）

马尔科夫决策过程（MDP）

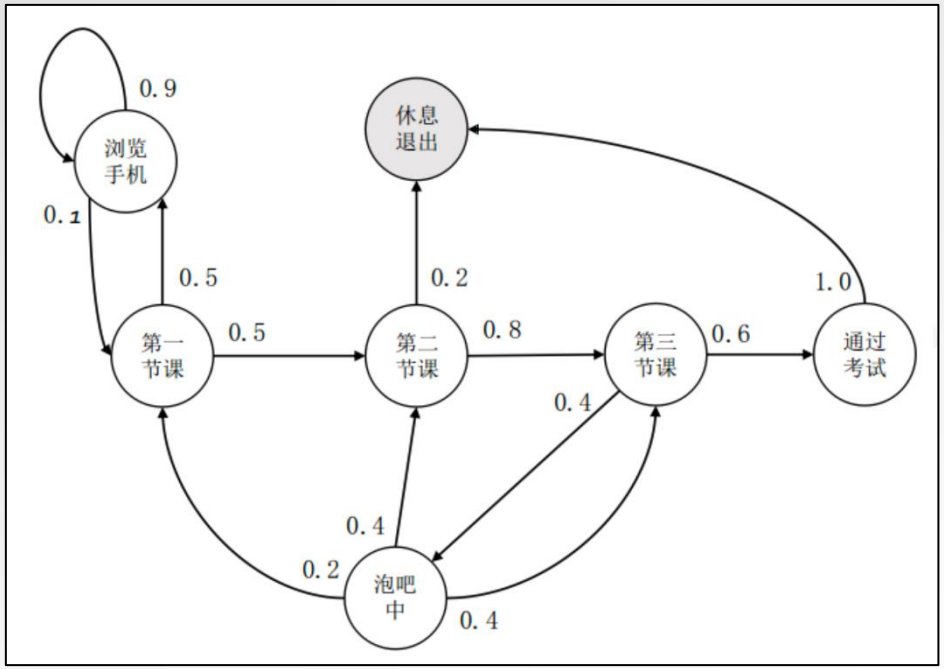
收获（return)、价值（value）

实验二

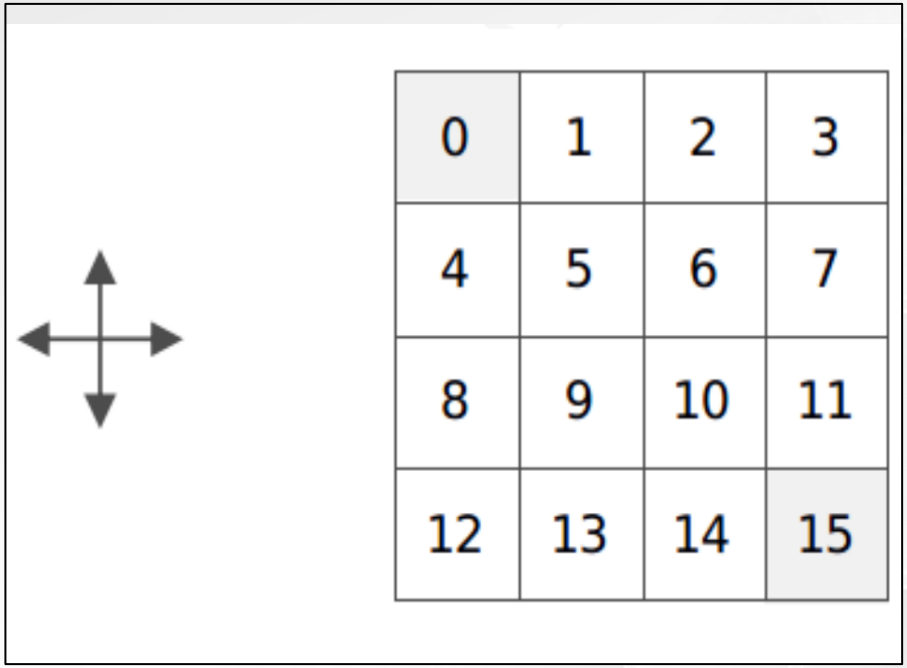
策略评估（Policy Evaluation）

策略迭代（Policy Iteration）

价值迭代（Value Iteration）



学生马尔可夫过程

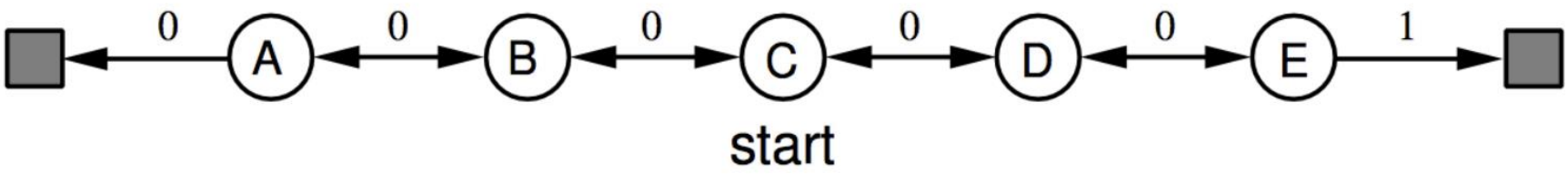


4✕4 小型方格世界

实验示例

Random Walk

在这个例子中，用TD(0) / MC方法应用于马尔科夫奖励过程(MRP)



① 从 C 出发，左右概率相同

② 最右侧 reward 1，其余为 0

③ 假设 $v(C)=\frac{1}{2}$ ，$\gamma=1$

$$v(A)=\frac{1}{2}*0+\frac{1}{2}*v(B)$$
$$v(B)=\frac{1}{2}*v(A)+\frac{1}{2}*v(C)$$
$$v(C)=\frac{1}{2}*v(B)+\frac{1}{2}*v(D)$$

每个状态的真实value: $v(A)=\frac{1}{6}$; $v(B)=\frac{2}{6}$; $v(C)=\frac{3}{6}$; $v(D)=\frac{4}{6}$; $v(E)=\frac{5}{6}$ 。

$$v(s)=R_s+\gamma\sum_{s'\in S}P_{ss'}v(s')$$

实验理论知识

- 蒙特卡罗 （MC）
- 时序差分（TD）

蒙特卡罗 (MC)

蒙特卡罗强化学习 (Monte-Carlo reinforcement learning, MC 学习)：指在不清楚 MDP 状态转移概率的情况下，直接从经历完整的状态序列 (episode) 来估计状态的真实价值，并认为某状态的价值等于在多个状态序列中以该状态算得到的所有收获的平均。

蒙特卡罗强化学习有如下**特点**：不依赖状态转移概率，直接从经历过的完整的状态序列中学习，使用的思想就是用平均收获值代替价值。理论上完整的状态序列越多，结果越准确。

累进更新平均值

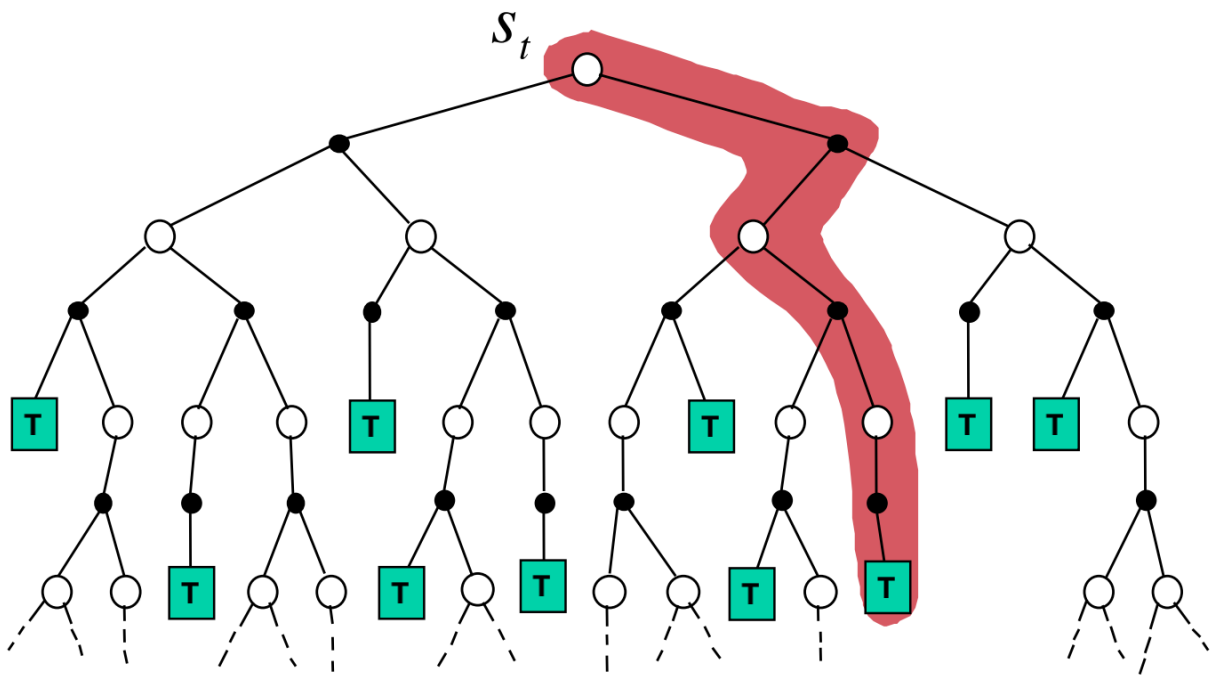
一种非常实用的不需要存储所有历史收获的计算方法：**累进更新平均值**（incremental mean）

$$
\begin{aligned}
\mu_k &= \frac{1}{k}\sum_{j=1}^{k} x_j \\
&= \frac{1}{k}\left(x_k + \sum_{j=1}^{k-1} x_j\right) \\
&= \frac{1}{k}\left(x_k + (k-1)\mu_{k-1}\right) \\
&= \mu_{k-1} + \frac{1}{k}(x_k - \mu_{k-1})
\end{aligned}
$$

累进更新平均值：利用前一次的平均值和当前数据以及数据总个数来计算新的平均值。

MC方法

$$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t - V(S_t)\right)$$



时序差分（TD）

时序差分强化学习（temporal-difference reinforcement learning，TD 学习）：指从采样得到的不完整的状态序列学习，该方法通过合理的**引导**（bootstrapping），先估计某状态在该状态序列完整后可能得到的收获，并在此基础上利用前文所述的累进更新平均值的方法得到该状态的价值，再通过不断的采样来持续更新这个价值。

具体地说，在 TD 学习中，算法在估计某一个状态的收获时，用的是离开该状态的即刻奖励 R_{t+1} 与下一时刻状态 S_{t+1} 的预估状态价值乘以衰减系数 γ 组成：

TD 误差

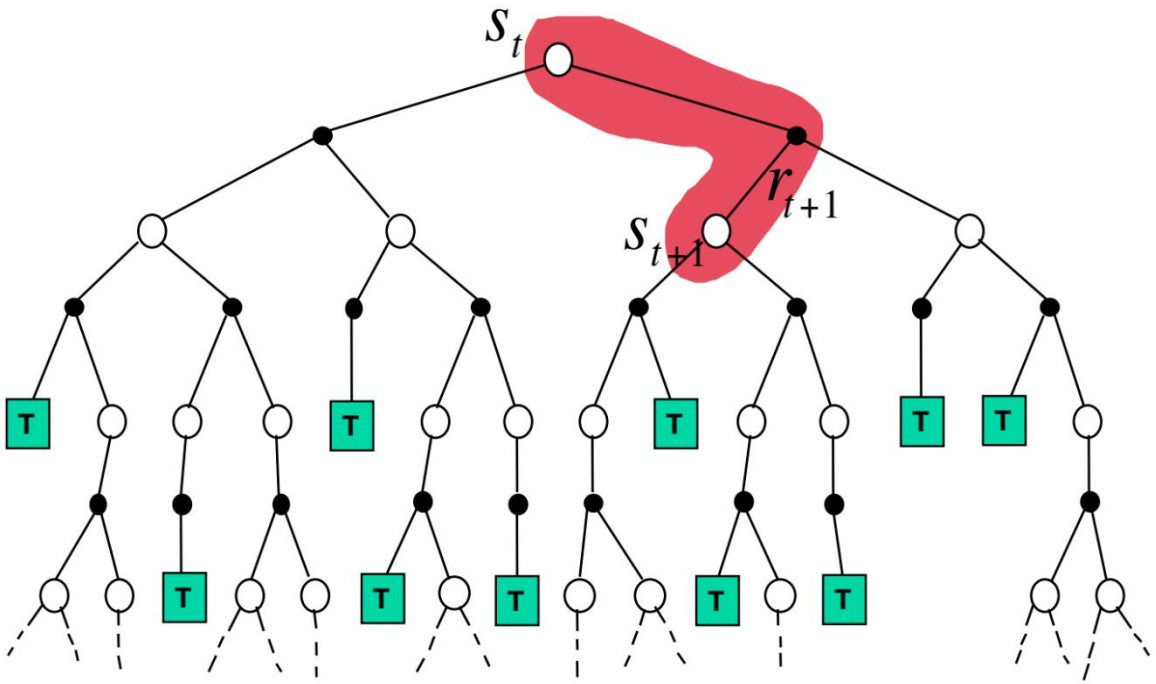
$$V(S_t) \leftarrow V(S_t) + \alpha\left(R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right)$$

TD目标值

引导（bootstrapping）：指的是用 TD 目标值代替收获 G_t 的过程。

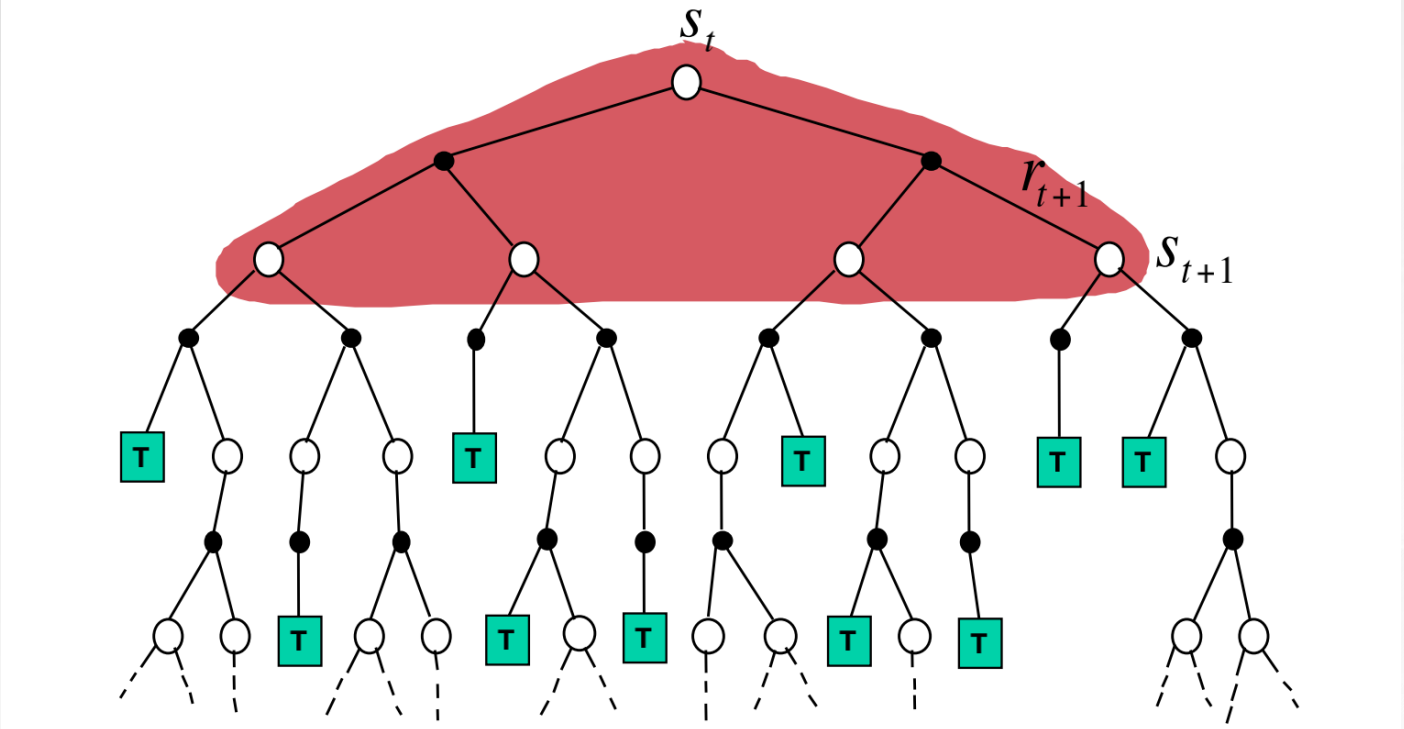
TD 方法

$$V(S_t) \leftarrow V(S_t) + \alpha\left(R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right)$$



DP 方法

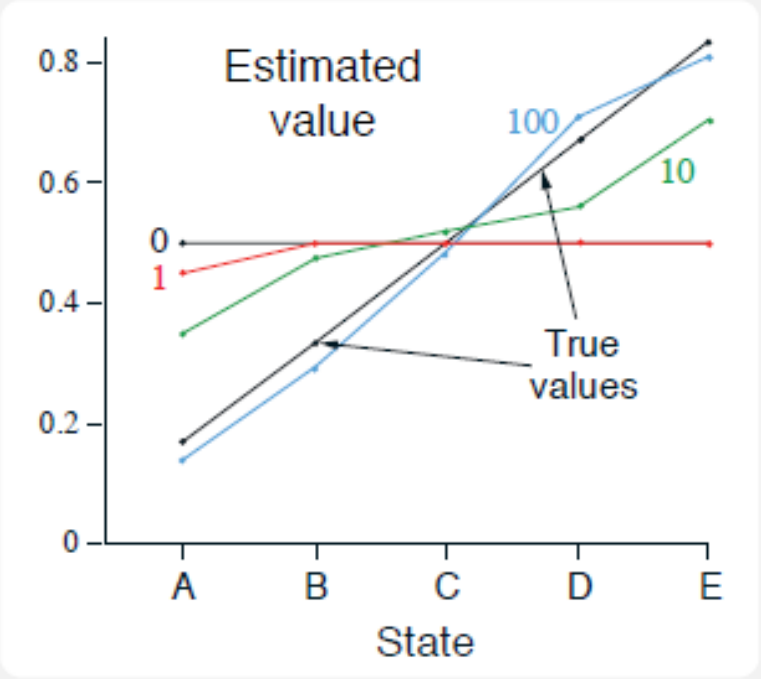
$$V(S_t) \leftarrow \mathbb{E}_\pi \left[R_{t+1} + \gamma V(S_{t+1}) \right]$$

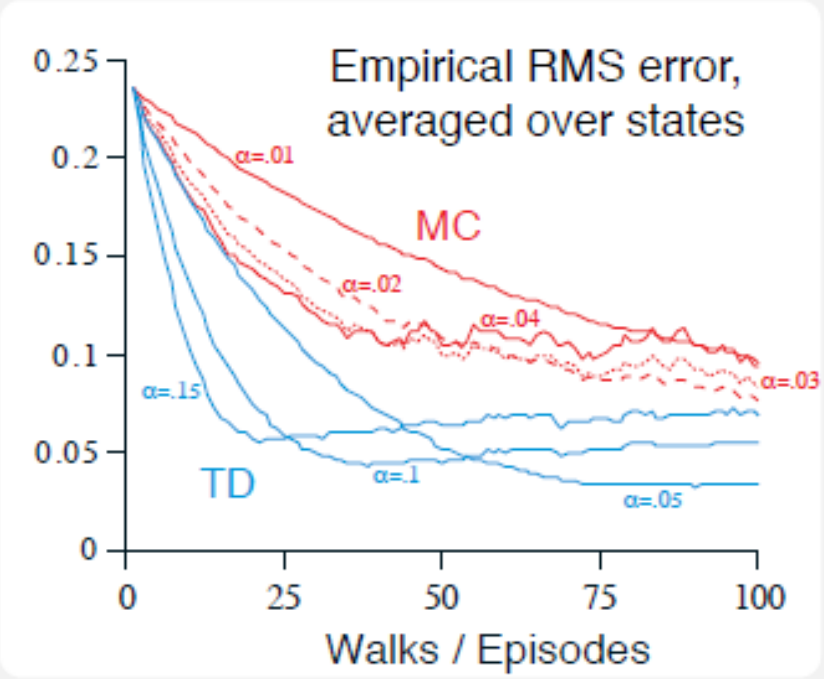


DP vs. MC vs. TD

方法	值函数估计	是否自举	是否采样	偏差	方差
DP	$V_\pi(S) = E_\pi[R_{t+1} + \gamma V(S_{t+1}) \mid S_t = s]$	自举	无须采样	无	无
MC	$V_\pi(S_t) \approx G_t \mid S_t = s$	不自举	采样 完整轨迹	无	高
TD	$V_\pi(S_t) \approx R_{t+1} + \gamma V(S_{t+1}) \mid S_t = s$	自举	采样 不完整轨迹	无(真实TD目标) 有(预估TD目标)	低

实验示例





$\alpha = 0.1$，$episode$数量增加

TD 估计值与真实值变化趋势

α 取不同时，TD/MC的误差收敛随 $episodes$ 变化趋势

Python代码实现

- MC 方法
- TD(0) 方法
- MC / TD(0) 误差比较
- 均方根误差计算