第四讲：无模型方法一——蒙特卡洛

主讲人 陈达贵

清华大学自动化系

在读硕士

深蓝学院
shenlanxueyuan.com

强化学习理论与实践

2018-12-28

目录

1 无模型方法简介

2 在策略和离策略

3 蒙特卡洛方法

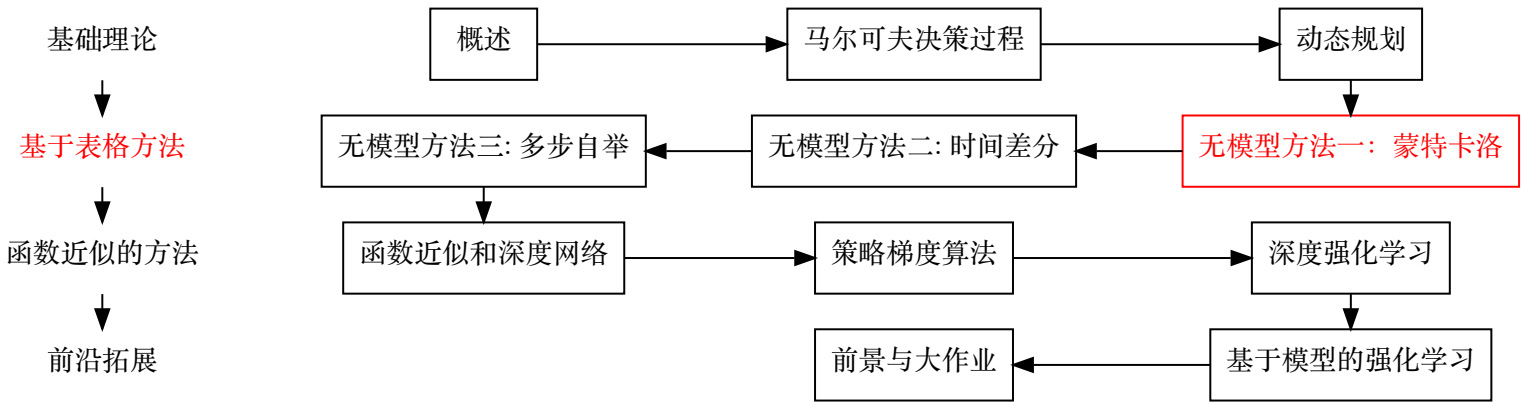
4 蒙特卡洛评价

5 增量式蒙特卡洛算法

6 蒙特卡洛优化

7 蒙特卡洛算法引申

章节目录


基础理论
基于表格方法
函数近似的方法
前沿拓展
概述
马尔可夫决策过程
动态规划
无模型方法一：蒙特卡洛
无模型方法二: 时间差分
无模型方法三: 多步自举
函数近似和深度网络
策略梯度算法
深度强化学习
基于模型的强化学习
前景与大作业

本章目录

1 无模型方法简介

2 在策略和离策略

3 蒙特卡洛方法

4 蒙特卡洛评价

5 增量式蒙特卡洛算法

6 蒙特卡洛优化

7 蒙特卡洛算法引申

无模型 (model-free) 方法

- 属于学习方法的一种
- MDPs 中未知 $\mathcal{P}, \mathcal{R}$——无模型
- 需要智能体和环境进行交互
- 一般采用样本备份
- 需要结合充分的探索

从经验中学习

- 由于未知环境模型，所以无法预知自己的后继状态和奖励值
- 通过与环境进行交互然后观察环境返回的值
- 本质上相当于从概率分布 $\mathcal{P}_{ss'}^{a}, \mathcal{R}_{s}^{a}$ 中进行采样
- 上述是对随机变量 S' 和 R 的采样，需要实现完整的轨迹还需要确定 A
- 采样策略得到 A
  - 可控制
- 采样足够充分时，可以使用样本分布良好地刻画总体分布

和动态规划的区别

- 无模型学习
  - 未知环境模型
  - 需要与环境进行交互，有交互成本
  - 样本备份
  - 异步备份
  - 需要充分的探索
  - 两个策略 (行为策略和目标策略)
  - ...
- 动态规划
  - 已知环境模型
  - 不需要直接交互，直接利用环境模型推导
  - 全宽备份
  - 同步和异步
  - 无探索
  - 一个策略
  - ...

目录

1 无模型方法简介

2 在策略和离策略

3 蒙特卡洛方法

4 蒙特卡洛评价

5 增量式蒙特卡洛算法

6 蒙特卡洛优化

7 蒙特卡洛算法引申

行为策略和目标策略

- 行为策略是智能体与环境交互的策略
- 目标策略是我们要学习的策略，即 v_π, q_π 时的下标

在策略 (on-policy) 学习

- 行为策略和目标策略是同一个策略
- 直接使用样本统计属性去估计总体
- 更简单，且收敛性更好
- 数据利用性更差 (只有智能体当前交互的样本能够被利用)
- 限定了学习过程中的策略是随机性策略

离策略 (off-policy) 学习

- 行为策略和目标策略不是同一个策略
- 一般行为策略 μ 选用随机性策略，目标策略 π 选用确定性策略
- 需要结合重要性采样才能使用样本估计总体
- 方差更大，收敛性更差
- 数据利用性更好 (可以使用其他智能体交互的样本)
- 行为策略需要比目标策略更具备探索性。即，在每个状态下，目标策略的可行动作是行为策略可行动作的子集

$$\pi(a|s) > 0 \Longrightarrow \mu(a|s) > 0$$

重要性采样

- 重要性采样是一种估计概率分布期望值的技术，它使用了来自其他概率分布的样本。
- 主要用于无法直接采样原分布的情况
- 估计期望值时，需要加权概率分布的比值（称为重要性采样率）
- 例子：
  - 估计全班身高，总体男女比例 1:2
  - 由于某些限制，只能按男女比例 2:1 去采样
  - 如果不考虑采样的分布形式，直接平均得到的值就有问题
  - 因此需要加权，加权比例是 $\frac{1}{2} : \frac{2}{1} = 1 : 4$ 去加权

重要性采样

$$\begin{aligned}\mathbb{E}_{X\sim P}[f(X)] &= \sum P(X)f(X) \\ &= \sum Q(X)\frac{P(X)}{Q(X)}f(X) \\ &= \mathbb{E}_{X\sim Q}\left[\frac{P(X)}{Q(X)}f(X)\right]\end{aligned}$$

离策略学习中的重要性采样

考虑 t 时刻之后的动作状态轨迹$\rho_t = A_t, S_{t+1}, A_{t+1}, \cdots, S_T$，可以得到该轨迹出现的概率为

$$\mathbb{P}(\rho_t) = \prod_{k=t}^{T-1} \pi(A_k|S_k)\mathbb{P}(S_{k+1}|S_k, A_k)$$

因此可以得到相应的重要性采样率为

$$\eta_t^T = \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k)\mathbb{P}(S_{k+1}|S_k, A_k)}{\prod_{k=t}^{T-1} \mu(A_k|S_k)\mathbb{P}(S_{k+1}|S_k, A_k)} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}$$

即便是未知环境模型，也能得到重要性采样率

目录

1 无模型方法简介

2 在策略和离策略

3 蒙特卡洛方法

4 蒙特卡洛评价

5 增量式蒙特卡洛算法

6 蒙特卡洛优化

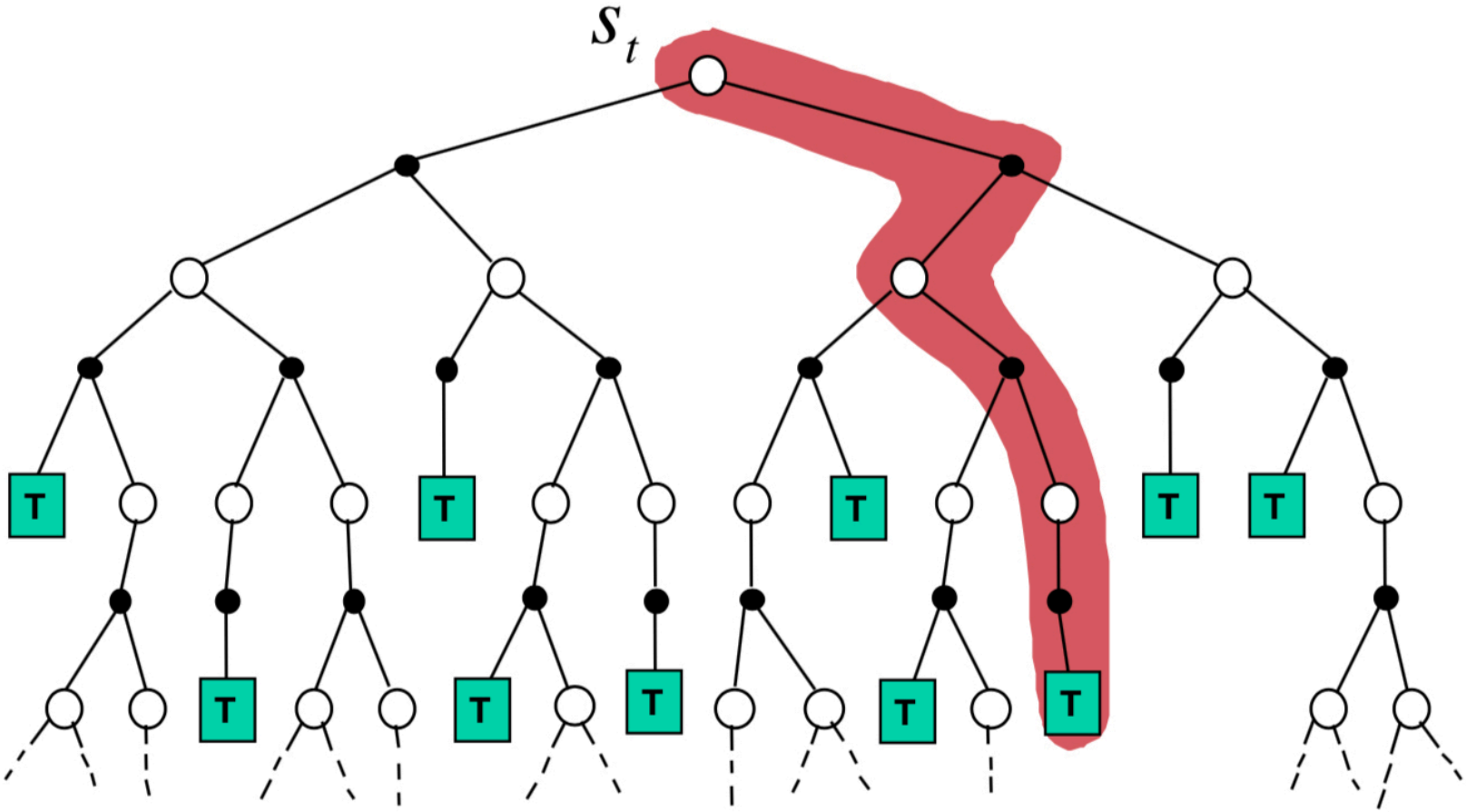
7 蒙特卡洛算法引申

蒙特卡洛 (Monte-Carlo, MC) 方法

- MC 方法可以被用于任意涉及随机变量的估计
- 这里 MC 方法特指利用统计平均估计期望值的方法
- 强化学习中存在很多估计期望值的计算 v_π, v_*
- 使用 MC 方法只需要利用经验数据，不需要 $\mathcal{P}, \mathcal{R}$
- MC 方法从完整的片段中学习
- MC 方法仅仅用于片段性任务 (必须有终止条件)

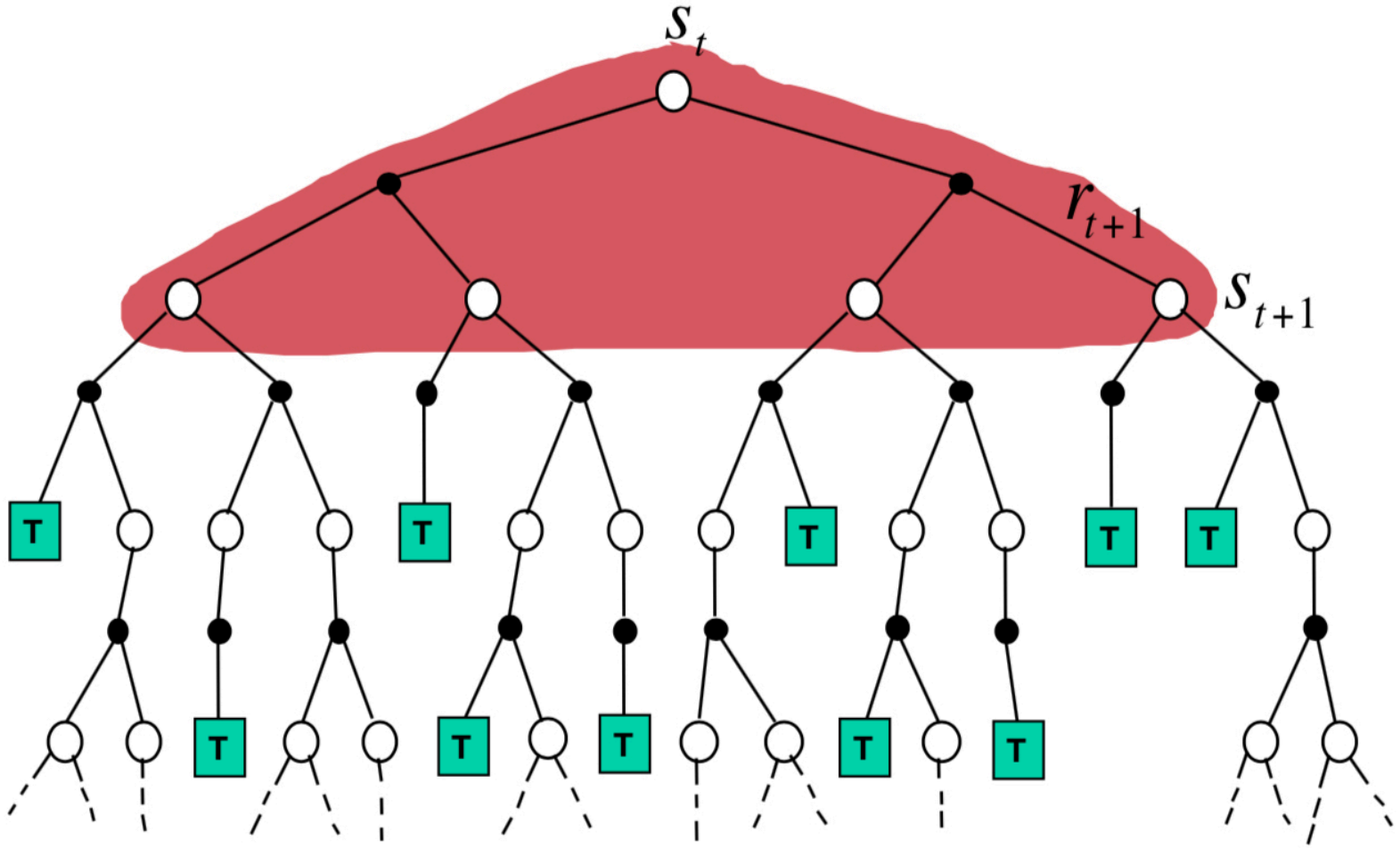
蒙特卡洛方法

最简单的思路，通过不断的采样，然后统计平均回报值来估计值函数，方差较大



动态规划方法

使用所有后继状态进行全宽备份



目录

1 无模型方法简介

2 在策略和离策略

3 蒙特卡洛方法

4 蒙特卡洛评价

5 增量式蒙特卡洛算法

6 蒙特卡洛优化

7 蒙特卡洛算法引申

一些表述说明

- 轨迹：把状态-动作的序列称为一个智能体的轨迹(trajectory)[1]

$$\rho = S_1, A_1, S_2, A_2, \cdots, S_T$$

  - 状态动作序列构成了马尔可夫链图上的一条轨迹
- 从 $\pi, \mathcal{P}^a_{ss'}$ 采样一条轨迹：我们把智能体从初始状态开始和环境进行交互的整个过程中得到的轨迹叫做采样一条轨迹。其中需要考虑两个分布 $\pi, \mathcal{P}^a_{ss'}$
- 从策略中采样一条轨迹 ρ。因为 $\mathcal{P}^a_{ss'}$ 是稳定的，所以轨迹的分布随着策略的变化而变化。我们简述成从一个策略 π 中采样轨迹

$$\rho \sim \pi$$

[1] 有时也会加上奖励值

蒙特卡洛策略评价

- 目标：给定策略 π，求 v_π
- 过去的方法使用了贝尔曼期望方程

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \left(\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} v_\pi(s') \right)$$

  - 直接解
  - 迭代式动态规划
- MC 利用了值函数的定义

$$v_\pi(s) = \mathbb{E}_\pi \left[G_t | S_t = s \right]$$

- MC 策略评价使用回报值的经验平均来估计实际期望值

首次拜访 (First-visit)MC 策略评价

- 为了评价状态 s
- 使用给定的策略 π 采样大量的轨迹
  - 在每一条轨迹中，对于状态 s首次出现的时间 t
  - 增加状态数量 $N(s) \leftarrow N(s) + 1$
  - 增加总回报值 $G_{sum}(s) \leftarrow G_{sum}(s) + G_t$
- 计算平均值得到值函数的估计 $V(s) = \frac{G_{sum}(s)}{N(s)}$
- 每条轨迹都是独立同分布的
- 根据大数定律，随着 $N(s) \to \infty$，$V(s) \to v_\pi(s)$
- 在 MC 方法下，$V(s)$ 是 $v_\pi(s)$ 的无偏估计

每次拜访 (Every-visit)MC 策略评价

- 为了评价状态 s
- 使用给定的策略 π 采样大量的轨迹
  - 在每一条轨迹中，对于状态 s每次出现的时间 t
  - 增加状态数量 $N(s) \leftarrow N(s) + 1$
  - 增加总回报值 $G_{sum}(s) \leftarrow G_{sum}(s) + G_t$
- 计算平均值得到值函数的估计 $V(s) = \frac{G_{sum}(s)}{N(s)}$
- 同样地，随着 $N(s) \to \infty$，$V(s) \to v_\pi(s)$
- 收敛性的证明不如首次拜访 MC 策略评价直观。[2]
- 更容易拓展到函数逼近和资格迹（后述）

[2] Reinforcement learning with replacing eligibility traces.

对 Q 函数的 MC 方法

- 在没有模型的时候，一般我们选择估计 Q 函数
- 因为我们可以通过 Q 函数直接得到贪婪的策略，最优的 Q 函数可以得到最优的策略
- MC 方法和 V 函数类似，但是 Q 函数的拜访从状态 s 变成了在状态 s 下做动作 a
- 一个重要的区别是，在给定策略 π 的情况下，大量的 $\langle s, a \rangle$ 都没有被遍历到
  - 尤其是当策略 π 是确定性策略时，每个 s 只对应一个 a
- 一种避免该困境的方法是假设探索开始，即随机选择初始状态和初始动作

离策略 MC 策略评价

- 在采样轨迹时使用的策略是 μ
- 而我们计算的值函数是 π
- 使用重要性采样率去加权回报值

$$G_t^{\pi/\mu} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} G_t$$

- 将所有在策略的 MC 算法中的 G_t 替换成 $G_t^{\pi/\mu}$ 就得到离策略 MC 算法
- 使用重要性采样会显著增加方差，可能到无限大。(增加了 X^2)

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \bar{X}^2$$

MC 的特点小结

- 偏差为 0，是无偏估计
- 方差较大，需要大量数据去消除
- 收敛性较好
- 容易理解和使用
- 没有利用马尔可夫性，有时可以用在非马尔可夫环境

目录

1 无模型方法简介

2 在策略和离策略

3 蒙特卡洛方法

4 蒙特卡洛评价

5 增量式蒙特卡洛算法

6 蒙特卡洛优化

7 蒙特卡洛算法引申

为什么需要增量式算法？

- 之前的蒙特卡洛算法需要采样大量轨迹之后再统一计算平均数
- 能不能在每一条轨迹之后都得到值函数的估计值呢？
- 平均值能够以增量形式进行计算

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

增量式 MC 更新

- 采样轨迹 $S_1, A_1, S_2, A_2, \cdots, S_T$
- 对于每一个状态 S_t, 统计回报值 G_t,

$$N(S_t) \leftarrow N(S_t) + 1$$

$$V(S_t) \leftarrow V(S_t) + \frac{1}{N(S_t)} (G_t - V(S_t))$$

- 此时 $1/N(S_t)$ 可以认为是更新的步长

常量步长

很多时候，我们会采样常量步长 $\alpha \in (0, 1]$。

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

$$\begin{aligned}
V_{k+1} &= V_k + \alpha(g_k - V_k)) \\
&= \alpha g_k + (1-\alpha) V_k \\
&= \alpha g_k + (1-\alpha) \left[\alpha g_{k-1} + (1-\alpha) V_{k-1}\right] \\
&= \alpha g_k + (1-\alpha)\alpha g_{k-1} + (1-\alpha)^2 V_{k-1} \\
&= \alpha g_k + (1-\alpha)\alpha g_{k-1} + (1-\alpha)^2 \alpha g_{k-2} + \\
&\quad \ldots + (1-\alpha)^{k-1} \alpha g_1 + (1-\alpha)^k V_1 \\
&= (1-\alpha)^k V_1 + \sum_{i=1}^{k} \alpha (1-\alpha)^{k-i} g_i
\end{aligned}$$

由于 $(1-\alpha)^k + \sum_{i=1}^{k} \alpha(1-\alpha)^{n-i} = 1$，**所以也可以认为常量步长是回报值的指数加权**

常量步长

- 仍然是回报值的加权平均
- 会逐渐遗忘过去的轨迹
- 对初始值敏感度更小
- 更简单，不用使用 $N(S_t)$
- 适用于不稳定环境

目录

1 无模型方法简介

2 在策略和离策略

3 蒙特卡洛方法

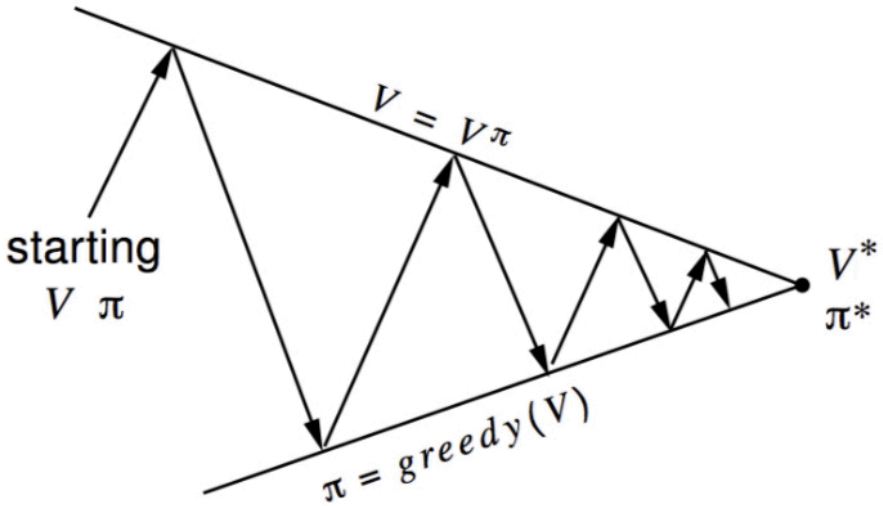
4 蒙特卡洛评价

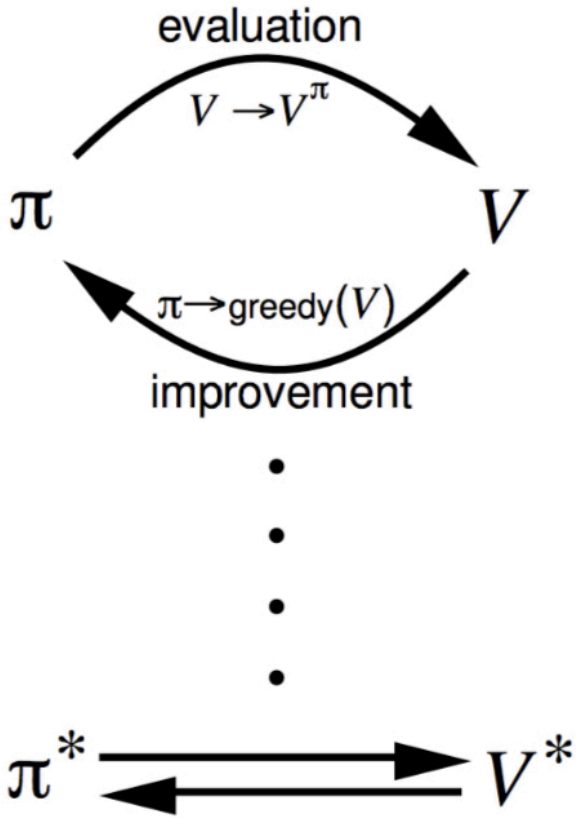
5 增量式蒙特卡洛算法

6 蒙特卡洛优化

7 蒙特卡洛算法引申

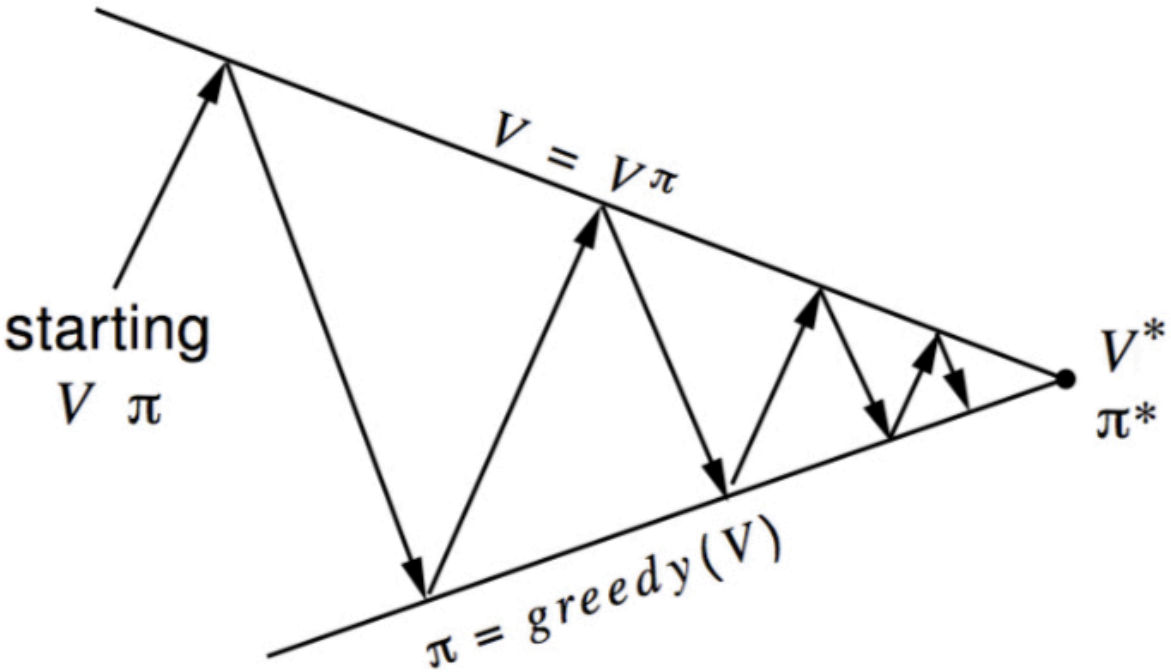
广义策略迭代





- **策略评价**：估计 v_π
  比如迭代式策略评价
- **策略提升**：生成 $\pi' \geq \pi$
  比如贪婪策略提升

MC 中的广义策略迭代



- **策略评价**：估计 v_π
  蒙特卡洛策略评价，$V = v_\pi$?
- **策略提升**：生成 $\pi' \geq \pi$
  贪婪策略提升?

问题一：使用哪个值函数？

- 在 V 函数上做贪婪策略提升要求环境模型

$$\pi'(s) = \arg\max_{a \in \mathcal{A}} \mathcal{R}(s, a) + \mathcal{P}^a_{ss'} V(s')$$

- 在 Q 函数上做贪婪策略提升是无模型的

$$\pi'(s) = \arg\max_{a \in \mathcal{A}} Q(s, a)$$

问题二：贪婪策略提升？

- MC 虽然利用了过去的经验数据，但是
  - 某些状态并未遍历到
  - 遍历不够充分，置信度不高
- 例子:
  - 比如每天早上有两个选择，学习和玩耍
  - 学习如果没成果，获得奖励 0，有成果获得奖励 1000
  - 玩耍获得奖励 1
  - 首次尝试学习获得奖励 0，$Q(s, \text{学习}) = 0$
  - 首次尝试玩耍获得奖励 1，$Q(s, \text{玩耍}) = 1$
  - 根据贪婪策略会选择玩耍
  - 由于不会选择学习，所以 $Q(s, \text{学习})$ 不会更新
- 你确定选到了最优策略?

ε-贪婪探索

- 解决这个问题，需要保证智能体一直在探索新的策略
- 最简单的做法，保证所有的 m 个动作都有一定的概率被采样
  - 用 $1-\varepsilon$ 的概率选择贪婪的动作
  - 用 ε 的概率随机从 m 个动作中选择

$$\pi(a|s) = \begin{cases} \varepsilon/m + 1 - \varepsilon & \text{if } a = \arg\max_{a \in \mathcal{A}} Q(s,a) \\ \varepsilon/m & \text{otherwise} \end{cases}$$

- 能同时解决对 Q 函数的蒙特卡洛策略评价中的探索开始假设

ε-贪婪策略提升

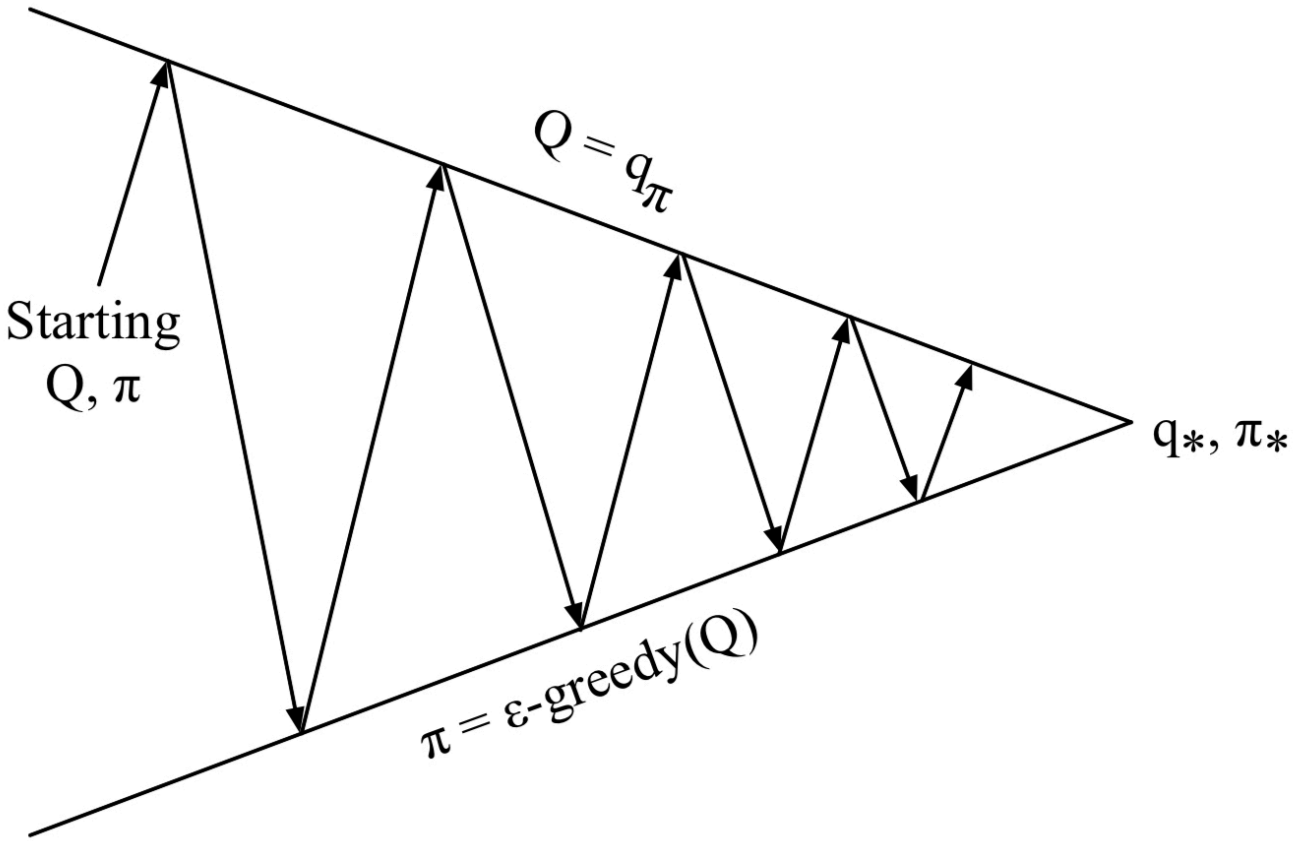
ε-贪婪策略提升定理

对于任意 ε-贪婪策略 π，使用相应的 q_π 得到的 ε-贪婪策略 π' 是在 π 上的一次策略提升，即 $v_{\pi'}(s) \geq v_\pi(s)$

$$
\begin{aligned}
q_\pi(s, \pi'(s)) &= \sum_{a \in \mathcal{A}} \pi'(a|s) q_\pi(s, a) \\
&= \frac{\varepsilon}{m} \sum_{a \in \mathcal{A}} q_\pi(s, a) + (1 - \varepsilon) \max_{a \in \mathcal{A}} q_\pi(s, a) \\
&\geq \frac{\varepsilon}{m} \sum_{a \in \mathcal{A}} q_\pi(s, a) + (1 - \varepsilon) \sum_{a \in \mathcal{A}} \frac{\pi(a|s) - \varepsilon/m}{1 - \varepsilon} q_\pi(s, a) \\
&= \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s, a) = v_\pi(s)
\end{aligned}
$$

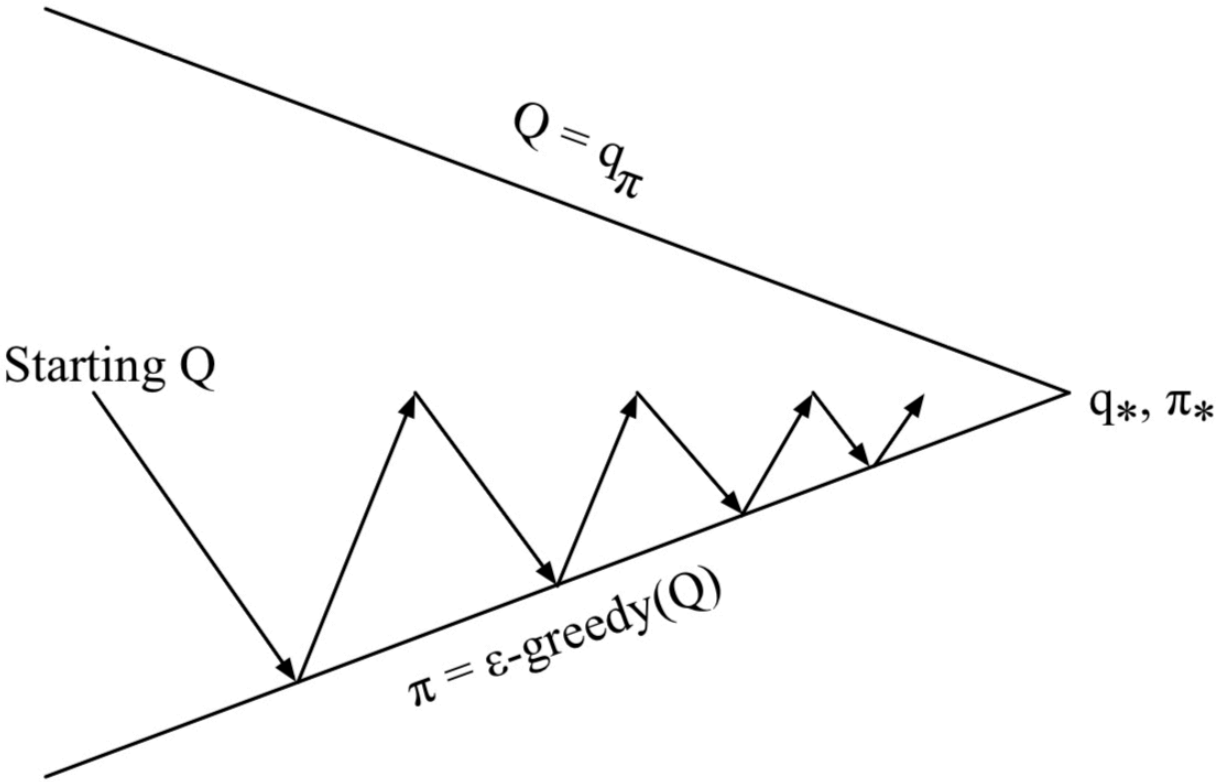
根据策略提升定理，可以得到 $v_{\pi'}(s) \geq v_\pi(s)$

MC 的策略迭代



- **策略评价**：蒙特卡洛策略评价，$Q = q_\pi$
- **策略提升**：ε-贪婪策略提升

增量式策略评价



每条轨迹:

- **策略评价**: 蒙特卡洛策略评价，$Q \approx q_\pi$
- **策略提升**: ε-贪婪策略提升

无限探索下的极限贪婪

定义

无限探索下的极限贪婪 (Greedy in the Limit with Infinite Exploration (GLIE))

- 无限探索：所有的状态动作对能够被探索无穷次

$$\lim_{k\to\infty} N_k(s,a) = \infty$$

- 极限贪婪：在极限的情况下，策略会收敛到一个贪婪的策略

$$\lim_{k\to\infty} \pi_k(a|s) = \mathbf{1}(a = \arg\max_{a'\in\mathcal{A}} Q_k(s,a'))$$

- 设置 ε 逐渐衰减到 0，比如 $\varepsilon_k = \frac{1}{k}$，$\varepsilon$-贪婪策略是 GLIE 的

定理

GLIE 蒙特卡洛优化能收敛到最优的 Q 函数

GLIE 蒙特卡洛优化

算法 1 GLIE 蒙特卡洛优化算法

1: **repeat** $k = 1, 2, 3, \ldots$

2: 使用策略 π 采样第 k 条轨迹，$S_1, A_1, S_2, A_2, \cdots, S_T$

3: 对于轨迹中的每一个 S_t 和 A_t

$$N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t)} (G_t - Q(S_t, A_t))$$

4: 执行 ε-策略提升，并衰减 ε 值

$$\varepsilon \leftarrow 1/k$$

$$\pi \leftarrow \varepsilon - \text{greedy}(Q)$$

5: **until** 收敛

目录

1 无模型方法简介

2 在策略和离策略

3 蒙特卡洛方法

4 蒙特卡洛评价

5 增量式蒙特卡洛算法

6 蒙特卡洛优化

7 蒙特卡洛算法引申

增量式离策略每次拜访蒙特卡洛评价

算法 2 增量式离策略每次拜访蒙特卡洛评价算法

repeat $k = 1, 2, 3, \ldots$
  使用策略μ采样第 k 条轨迹，$S_1, A_1, S_2, A_2, \cdots, S_T$
  $G \leftarrow 0, W \leftarrow 1$
  for $t = T-1, T-2, \cdots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t)]$
    $W \leftarrow W\frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}$
    如果 $W = 0$ 退出 For 循环
  end for
until 收敛

增量式离策略每次拜访蒙特卡洛优化

算法 3 增量式离策略每次拜访蒙特卡洛优化算法

repeat $k = 1, 2, 3, \ldots$
  使用策略 μ 采样第 k 条轨迹，$S_1, A_1, S_2, A_2, \cdots, S_T$
  $G \leftarrow 0, W \leftarrow 1$
  for $t = T-1, T-2, \cdots 0$ **do**
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)}[G - Q(S_t, A_t)]$
    $W \leftarrow W\frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}$
    $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$
    如果 $A_t \neq \pi(S_t)$ 则退出 For 循环
    $W \leftarrow W\frac{1}{\mu(A_t|S_t)}$
  end for
until 收敛