# 强化学习第III课

# Reinforcement Learning

七月在线

2017/10/14

# Outline

- 内容回顾
  - Known Environment MDP Prediction / Control
  - Unknown Environment MDP Prediction
- Unknown Environment MDP Control
  - Exploration and Exploitation
    - Multi-Armed Bandit Problem
    - ε-greedy strategy
  - On Policy / Off Policy Learning
    - Monte Carlo Method
    - TD Method: Sarsa (on policy TD), Q-Learning (off policy TD)

# 快速回顾I

- 马尔科夫决策过程 <S, A, P, R, γ>
- 状态值函数V(s), 动作值函数q(s, a), 策略 π (s)
- Bellman Expectation Equation

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

- Bellman  Optimality Equation

$$v_*(s) = \max_a \mathcal{R}_s^a + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}_{ss'}^a v_*(s') \qquad q_*(s,a) = \mathcal{R}_s^a + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}_{ss'}^a \max_{a'} q_*(s',a')$$

- MDP Planning: 已知模型 P & R
  - 策略评估
  - 寻找最优策略  (值迭代  & 策略迭代)

# 快速回顾II

- Unknown Environment MDP <S, A, P?, R?, γ> Prediction
- 策略评估 for unknown MDP
  - 生成轨迹 under π, i.e., $S_1, A_1, R_2, \ldots, S_k \sim \pi$
  - 估计$V_\pi(s)$
    - Monte-Carlo：$V(s) \leftarrow V(s) + \alpha(G_t - V(s))$
    - Temporal-Difference：$V(s_t) \leftarrow V(s_t) + \alpha(R_{t+1} + \gamma V(s_{t+1}) - V(s_t))$

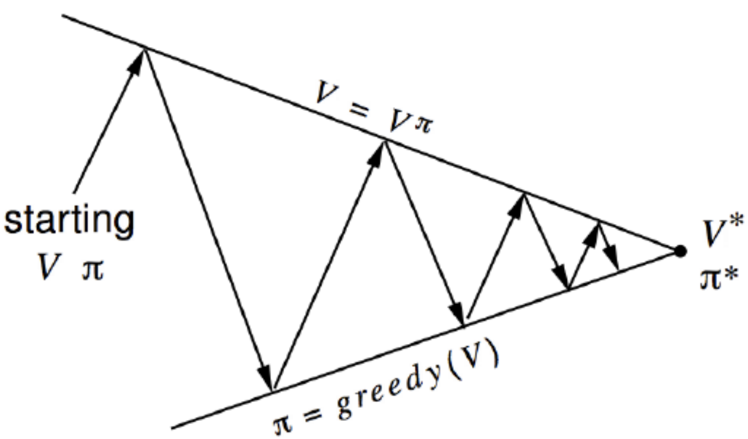
| MC | TD(0) |
|---|---|
| 要等到episode结束才能获得return | 每一步执行完都能获得一个return |
| 只能使用完整的episode | 可以使用不完整的episode |
| 高variance，零bias | 低variance，有bias |
| 没有体现出马尔可夫性质 | 体现出了马尔可夫性质 (use MDP) |
| No Bootstrapping | Bootstrapping |
| 收敛慢， steady | 收敛快， not steady |

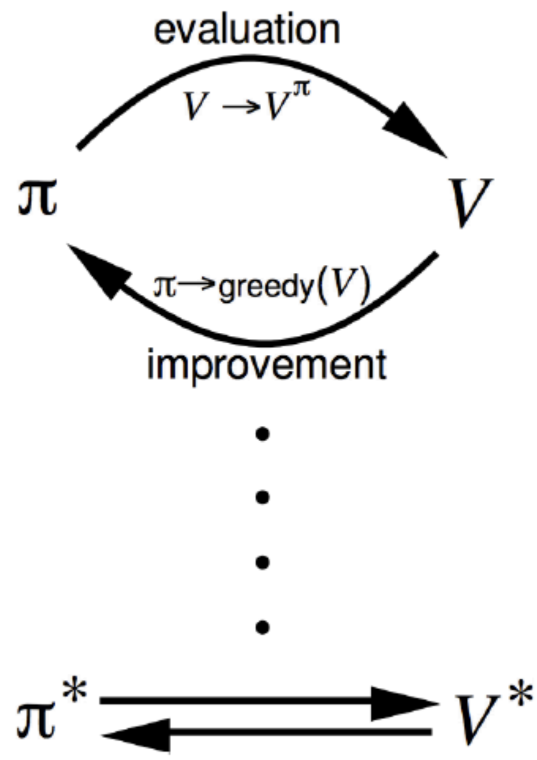
# Unknown Environment MDP Control

- 基本思路：广义策略迭代（策略评估＋策略改进）

  回顾策略迭代for known environment MDP

  - 给定策略π, 评估策略得到$V_\pi(s)$
  - 改进策略：$\pi' = \text{greedy}(V_\pi)$ => $\pi' \geq \pi$



# Unknown Environment MDP Control

- 基本思路：广义策略迭代（策略评估＋策略改进）

- 问题I：
  - 策略评估 For known Environment MDP (solve Bellman Expectation Equation)
  - 策略评估 For unknown Environment MDP (Estimate from sample trajectories)

- 问题II:
  - 策略改进over V(s) require model $\pi'(s) = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \mathcal{R}_s^a + \mathcal{P}_{ss'}^a V(s')$
  - 策略改进over Q(s,a) is model-free $\pi'(s) = \underset{a \in \mathcal{A}}{\operatorname{argmax}} Q(s, a)$

问题解决了！！！？

# Unknown Environment MDP Control

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

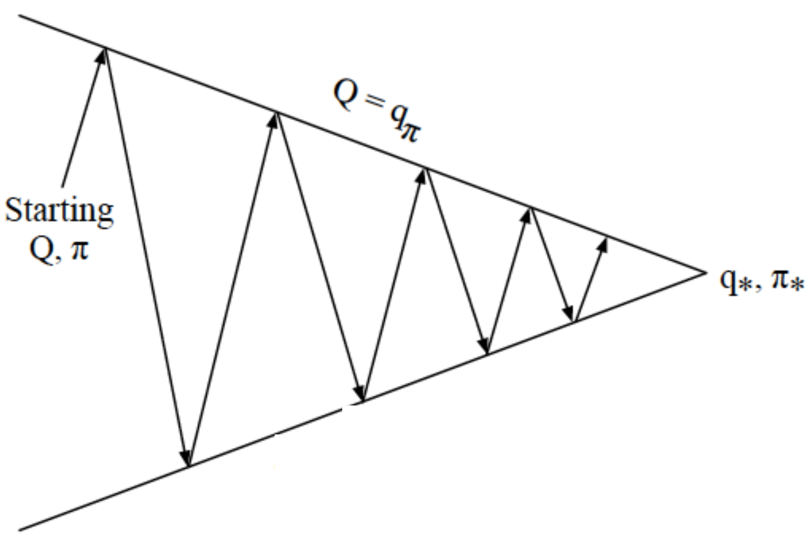
Repeat forever:
(a) Generate an episode using exploring starts and $\pi$
(b) For each pair $s,a$ appearing in the episode:
$R \leftarrow$ return following the first occurrence of $s,a$
Append $R$ to $Returns(s,a)$
$Q(s,a) \leftarrow$ average$(Returns(s,a))$
(c) For each $s$ in the episode:
$\pi(s) \leftarrow \arg\max_a Q(s,a)$

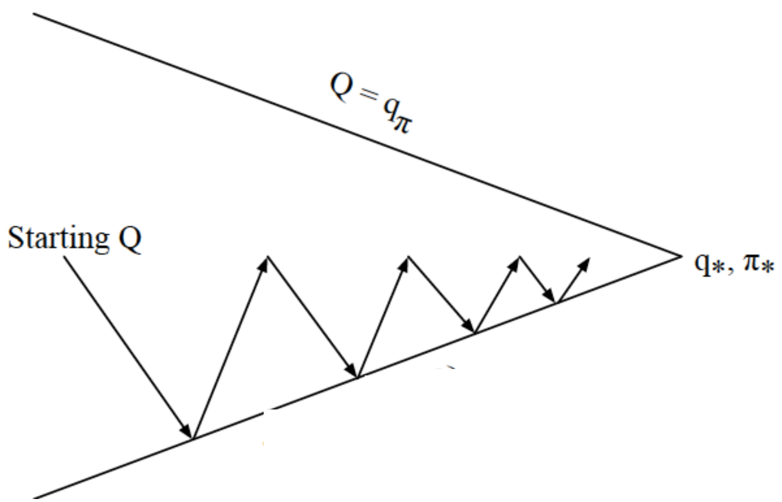
Fixed point is optimal policy $\pi^*$

Proof is open question

# Unknown Environment MDP Control

- 如何保证每个状态行为对(Q,a)都可以被访问到？
- No greedy!!!
- 确保历经每个状态行为对，$\pi(a|s) > 0$ for all a, s
- 每次迭代确保 $\pi' \geq \pi$ （回顾policy ordering）



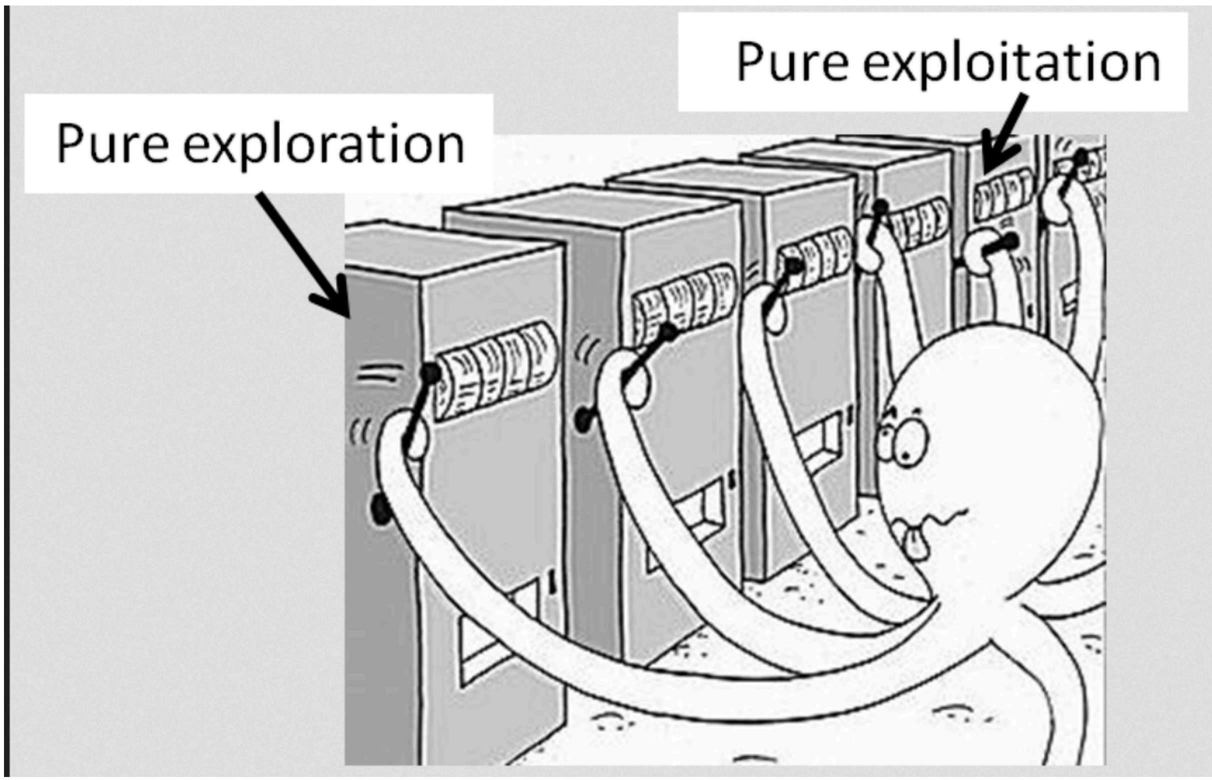



# Exploration and Exploitation

- 实时在线决策
  - Exploitation: 基于之前所有的信息做出最优选择
  - Exploration: 收集更多信息
- 最好的长远策略可能需要牺牲短期利益
- 只有收集到足够多的数据才能作出全局最好决策

# Exploration and Exploitation

多臂自动机(Multi-Armed Bandit)




应用：推荐系统问题

# Exploration and Exploitation

- Naive-Exploration: ε-greedy (Add noise to greedy strategy)

$$\pi\left(a|s\right) \leftarrow \begin{cases} 1-\varepsilon+\frac{\varepsilon}{|A(s)|} \; if \; a = arg\max_a Q\left(s,a\right) \\ \\ \frac{\varepsilon}{|A(s)|} \; if \; a \neq arg\max_a Q\left(s,a\right) \end{cases}$$

- Thompson Sampling

- Upper Confidence Bound(置信区间上界)

  Choose the arm with max value of $\bar{x}_j(t) + \sqrt{\frac{2\ln t}{T_{j,t}}}$

# Exploration and Exploitation

**Theorem**

*For any $\epsilon$-greedy policy $\pi$, the $\epsilon$-greedy policy $\pi'$ with respect to $q_\pi$ is an improvement, $v_{\pi'}(s) \geq v_\pi(s)$*

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

Therefore from policy improvement theorem, $v_{\pi'}(s) \geq v_\pi(s)$

-- From David Silver RL Lecture

# On Policy and Off Policy Learning

- On Policy Learning: 探索策略与评估策略为同一策略
  - “Learn on the job”
  - Learn about policy $\pi$ from experience sampled from $\pi$

- Off Policy Learning:探索策略与评估策略为不同策略
  - “Look over someone's shoulder”
  - Learn about policy $\pi$ from experience sampled from $\mu$
  - Learn from observing humans or other agents
  - Re-use experience generated from old policies $\pi_1, \pi_2, \ldots, \pi_{t-1}$
  - Learn about optimal policy while following exploratory policy
  - Learn about multiple policies while following one policy

# On Policy Monte Carlo

**On-policy first-visit MC control (for $\varepsilon$-soft policies), estimates $\pi \approx \pi_*$**

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s, a) \leftarrow$ arbitrary
    $Returns(s, a) \leftarrow$ empty list
    $\pi(a|s) \leftarrow$ an arbitrary $\varepsilon$-soft policy

Repeat forever:
    (a) Generate an episode using $\pi$
    (b) For each pair $s, a$ appearing in the episode:
        $G \leftarrow$ return following the first occurrence of $s, a$
        Append $G$ to $Returns(s, a)$
        $Q(s, a) \leftarrow \text{average}(Returns(s, a))$
    (c) For each $s$ in the episode:
        $A^* \leftarrow \arg\max_a Q(s, a)$
        For all $a \in \mathcal{A}(s)$:

$$\pi(a|s) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(s)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(s)| & \text{if } a \neq A^* \end{cases}$$

# On Policy TD (sarsa)

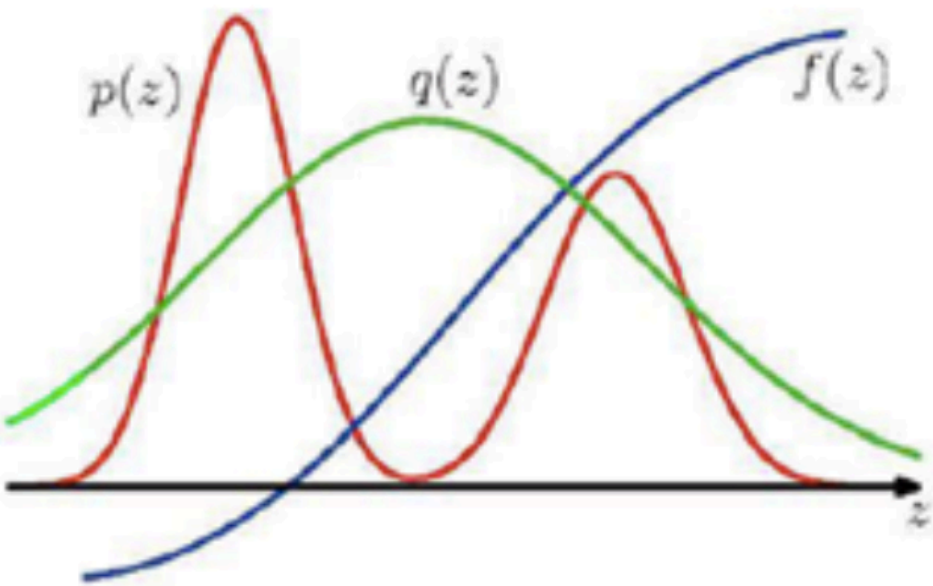
**Sarsa (on-policy TD control) for estimating $Q \approx q_*$**

Initialize $Q(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$, arbitrarily, and $Q(\textit{terminal-state}, \cdot) = 0$
Repeat (for each episode):
  Initialize $S$
  Choose $A$ from $S$ using policy derived from $Q$ (e.g., $\epsilon$-greedy)
  Repeat (for each step of episode):
    Take action $A$, observe $R$, $S'$
    Choose $A'$ from $S'$ using policy derived from $Q$ (e.g., $\epsilon$-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma Q(S',A') - Q(S,A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until $S$ is terminal

# 重要性抽样：Importance Sampling

Why Importance Sampling:

- Not easy to sample over original distribution
- To reduce variance




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

# 重要性抽样：Importance Sampling

Importance Sampling for MDP

Under policy π
$$Pr\left(A_t, S_{t+1}, \cdots, S_T\right) = \prod_{k=t}^{T-1} \pi\left(A_k \middle| S_k\right) p\left(S_{k+1} \middle| S_k, A_k\right)$$

Under Policy μ
$$Pr\left(A_t, S_{t+1}, \cdots, S_T\right) = \prod_{k=t}^{T-1} \mu\left(A_k \middle| S_k\right) p\left(S_{k+1} \middle| S_k, A_k\right)$$

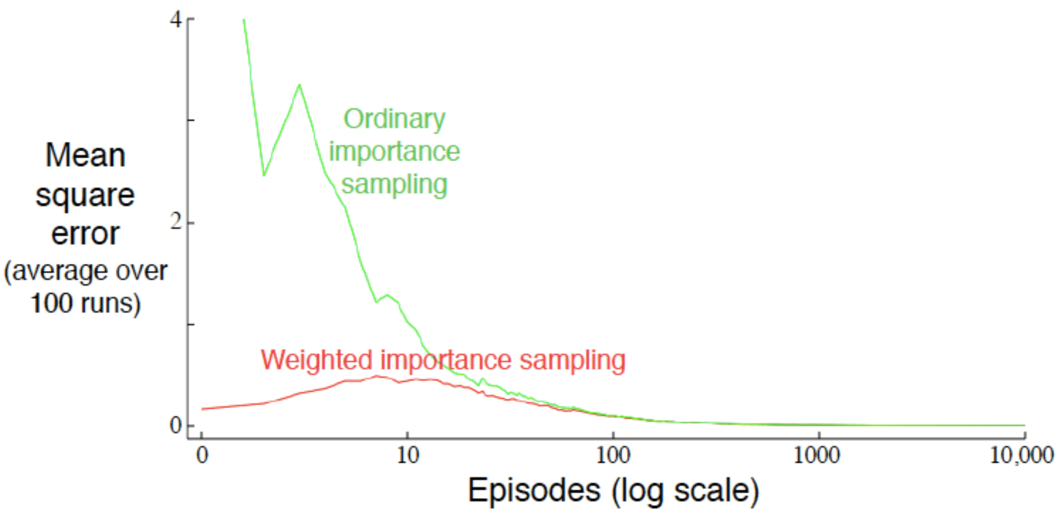
Import Sampling weights
$$\rho_t^T = \frac{\prod_{k=t}^{T-1} \pi\left(A_k \middle| S_k\right) p\left(S_{k+1} \middle| S_k, A_k\right)}{\prod_{k=t}^{T-1} \mu\left(A_k \middle| S_k\right) p\left(S_{k+1} \middle| S_k, A_k\right)} = \prod_{k=t}^{T-1} \frac{\pi\left(A_k \middle| S_k\right)}{\mu\left(A_k \middle| S_k\right)}$$

# 重要性抽样：Importance Sampling

Importance Sampling for MDP

Ordinary importance sampling $V(s) \doteq \frac{\sum_{t\in\mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{|\mathcal{T}(s)|}$

Weighted importance sampling $V(s) \doteq \frac{\sum_{t\in\mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{\sum_{t\in\mathcal{T}(s)} \rho_{t:T(t)-1}}$



# 重要性抽样：Importance Sampling

Incremental MC

Suppose we have a sequence of returns $G_1$, $G_2$, …, $G_{n-1}$ with weight $W_i = \rho_{t:T(t)-1}$

MC Estimate $$V_n \doteq \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}, \qquad n \geq 2,$$

Incremental MC Estimate

$$V_{n+1} \doteq V_n + \frac{W_n}{C_n}\Big[G_n - V_n\Big], \qquad n \geq 1$$

and

$$C_{n+1} \doteq C_n + W_{n+1},$$

# Off Policy Monte Carlo

**Off-policy MC control, for estimating $\pi \approx \pi_*$**

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
    $Q(s,a) \leftarrow$ arbitrary
    $C(s,a) \leftarrow 0$
    $\pi(s) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)

Repeat forever:
    $b \leftarrow$ any soft policy
    Generate an episode using $b$:
        $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$
    $G \leftarrow 0$
    $W \leftarrow 1$
    For $t = T-1, T-2, \ldots$ downto 0:
        $G \leftarrow \gamma G + R_{t+1}$
        $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
        $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
        $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)
        If $A_t \neq \pi(S_t)$ then ExitForLoop
        $W \leftarrow W \; 1$

# Off Policy TD (Q-learning)

One - step Q - learning :

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right]$$

Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
  Initialize $s$
  Repeat (for each step of episode):
    Choose $a$ from $s$ using policy derived from $Q$ (e.g., $\varepsilon$-greedy)
    Take action $a$, observe $r$, $s'$
    $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$
    $s \leftarrow s'$;
  until $s$ is terminal

No Importance Sampling, 更多关于Q-Learning见Lecture IV

# Thanks and Questions!!!