强化学习

马尔可夫过程

- 每个节点的当前状态只和上一个状态和状态转移矩阵有关

马尔可夫奖励过程

定义效用函数 两种定义
$V(s)=\sum_{s'}P(s'|s)(V(s')+R(s'))$
$V(s)=\sum_{s'}P(s'|s)(R(s')+\gamma V(s'))$
只和下一个状态有关
末状态的V为0，反向传播

马尔可夫决策过程

输入为$<S,A,R,P>$
策略π实际上为给定状态下各个动作的概率
修改效用函数的计算方法，实际上要对两个东西求和：一个是动作，一个是采取动作能够转移到的状态

MDP:

$$V^\pi(s)=\sum_a \pi(a|s)\sum_{s'}P(s'|s,a)\big(R(s,a,s')+V^\pi(s')\big)$$

定义Q值函数

实际上就是在当前节点的某个动作a下的效用函数
$Q^\pi(S,a)=\sum_{S'}P(S'|S,a)(V^\pi(S')+R(S,a,S'))$

consequently,

$$V^\pi(s) = \sum_a \pi(a|s) Q(s,a)$$

Q-function => policy

最优解

Bellman optimality equations

s	0	0.3
	1	0.7
c	0	0.6
	1	0.4
r	0	0.1
	1	0.9

$$V^*(s) = \max_a Q^*(s,a)$$

from the relation between V and Q

$$Q^*(s,a) = \sum_{s'} P(s'|s,a)\big(R(s,a,s') + \gamma V^*(s')\big)$$

we have

$$Q^*(s,a) = \sum_{s'} P(s'|s,a)\big(R(s,a,s') + \gamma \max_a Q^*(s',a)\big)$$

$$V^*(s) = \max_a \sum_{s'} P(s'|s,a)\big(R(s,a,s') + \gamma V^*(s')\big)$$

the unique fixed point is the optimal value function

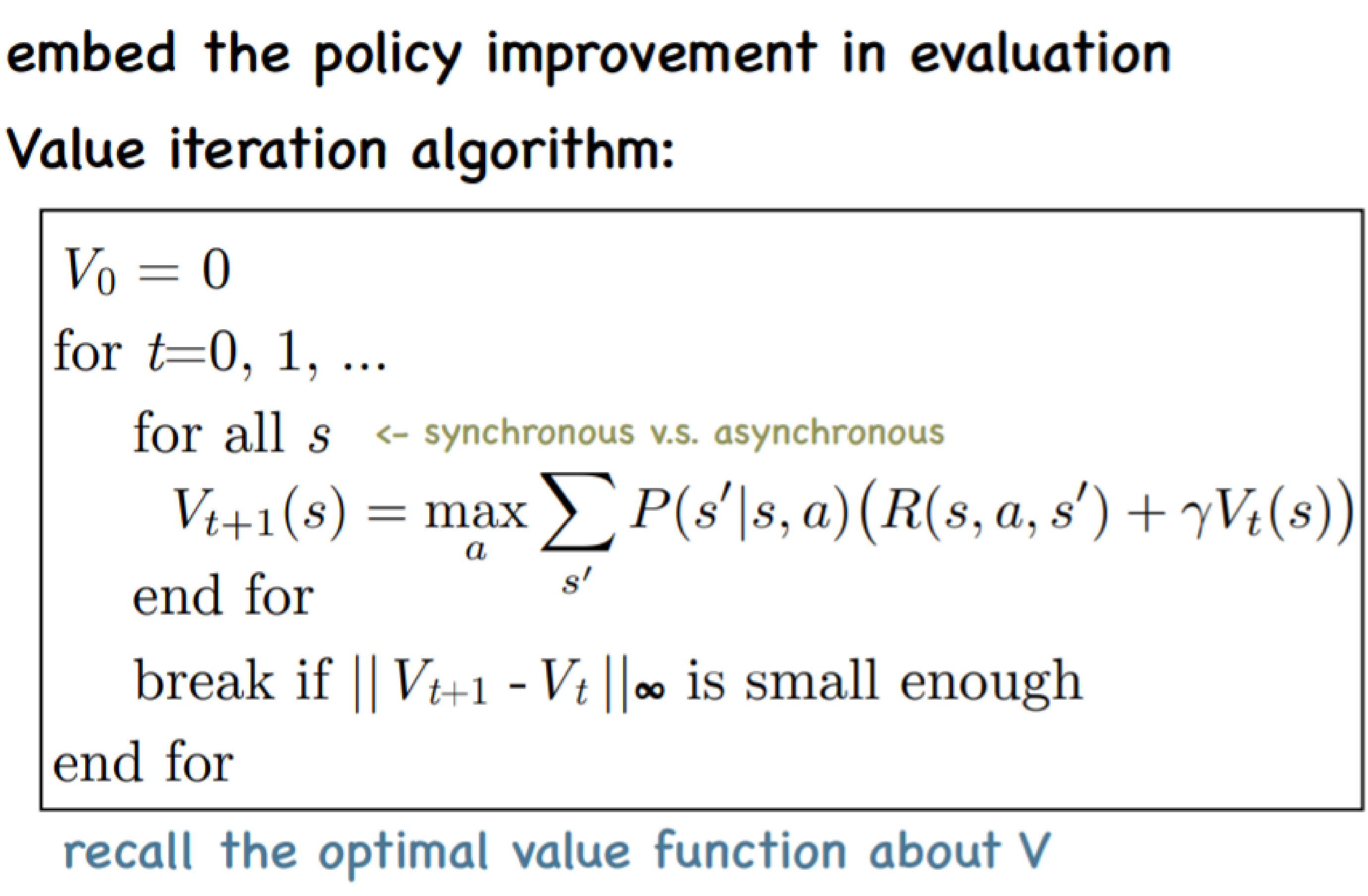
也就是说，这是一个固定点，可以被不断强化逼近

找到马尔可夫决策过程中的最优解

策略评估

策略评估就是V和Q的反向更新公式

策略改进



价值迭代

$$
\begin{aligned}
&V_0 = 0 \\
&\text{for } t\text{=0, 1, ...} \\
&\quad \text{for all } s \quad \text{<- synchronous v.s. asynchronous} \\
&\quad\quad V_{t+1}(s) = \max_a \sum_{s'} P(s'|s,a)\big(R(s,a,s') + \gamma V_t(s)\big) \\
&\quad \text{end for} \\
&\quad \text{break if } ||V_{t+1} - V_t||_\infty \text{ is small enough} \\
&\text{end for}
\end{aligned}
$$

recall the optimal value function about V

策略迭代

Policy iteration algorithm:

loop until converges

  policy evaluation: calculate V

  policy improvement: choose the action greedily

$$\pi_{t+1}(s) = \arg\max_a Q^{\pi_t}(s,a)$$

converges: $V^{\pi_{t+1}}(s) = V^{\pi_t}(s)$

$$Q^{\pi_{t+1}}(s,a) = \sum_{s'} P(s'|s,a)\big(R(s,a,s') + \gamma \max_a Q^{\pi_t}(s',a)\big)$$

recall the optimal value function about Q

在未知R和P的情况下学习

蒙特卡洛算法

Monte Carlo RL - evaluation+improvement

$Q_0 = 0$

for i=0, 1, ..., m

  generate trajectory <s_0, a_0, r_1, s_1, ..., s_T>

  for t=0, 1, ..., T-1

    R = sum of rewards from t to T

    $Q(s_t, a_t) = (\mathrm{c}(s_t, a_t)\, Q(s_t, a_t) + \mathrm{R}) / (\mathrm{c}(s_t, a_t) + 1)$

    $\mathrm{c}(s_t, a_t)$++

  end for

  update policy $\pi(s) = \arg\max_a Q(s, a)$

end for

improvement ?

每次更新的是一条蒙特卡洛采样路径上的点和对应动作

ϵ-greedy policy:

given a policy π

$$\pi_\epsilon(s) = \begin{cases} \pi(s), \text{with prob. } 1 - \epsilon \\ \text{randomly chosen action, with prob. } \epsilon \end{cases}$$

ensure probability of visiting every state > 0

Monte Carlo RL

$Q_0 = 0$

for i=0, 1, ..., m

  generate trajectory $<s_0, a_0, r_1, s_1, ..., s_T>$ by π_ϵ

  for t=0, 1, ..., T-1

    R = sum of rewards from t to T

    $Q(s_t, a_t) = (\mathrm{c}(s_t, a_t)\, Q(s_t, a_t) + \mathrm{R}) / (\mathrm{c}(s_t, a_t) + 1)$

    $\mathrm{c}(s_t, a_t)$++

  end for

  update policy $\pi(s) = \arg\max_a Q(s, a)$

end for

Monte Carlo RL -- off-policy

$Q_0 = 0$

for i=0, 1, ..., m

generate trajectory <s_0, a_0, r_1, s_1, ..., s_T> by π_ϵ

for t=0, 1, ..., T-1

R = sum of rewards from t to $T \times \prod_{i=t+1}^{T-1} \frac{\pi(x_i, a_i)}{p_i}$

$Q(s_t,a_t)$= (c(s_t,a_t) $Q(s_t,a_t)$+R)/(c(s_t,a_t)+1)

c(s_t,a_t)++

end for

update policy $\pi(s) = \arg\max_a Q(s, a)$

end for

$$p_i = \begin{cases} 1 - \epsilon + \epsilon/|A|, a_i = \pi(s_i), \\ \epsilon/|A|, a_i \neq \pi(s_i) \end{cases}$$