DP

策略改进定理

π, π', if $\forall s \in S$, $q_\pi(s, \pi') \geq V_\pi(s) \Rightarrow V_{\pi'}(s) \geq V_\pi(s)$

Proof:

$$q_\pi(s,a) = \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma V_\pi(s')\right]$$

$$= E_\pi\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, A_t = a\right]$$

original $q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = a]$

or

$$q_\pi(s,a) = \sum_{s' \in S} p(s,a,s')\left[R(s,a) + \gamma \sum_{a' \in A} \pi(s',a') q_\pi(s',a')\right]$$

$V_\pi(s')$

tips:

$V_\pi(s')$ ↓ one parameter

$q_\pi(s,a)$ → about s and a.

$G_t = R_{t+1} + \gamma G_{t+1}$ Bellman equation

$$\Rightarrow E_\pi\left[R_{t+1} + \gamma V_\pi(S_{t+1})\right] \mid S_t = s \quad A_t = a]$$

substitute π'

eg:

$$\Rightarrow E\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, A_t = \pi'(s)\right]$$

$$\Rightarrow E_{\pi'}\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s\right]$$

R_{t+1} has been controlled by π'

$$V_\pi(s) = E_\pi[G_t \mid S_t = s] \quad \text{MC}$$

$$= E_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s]$$

$$= E_\pi[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s] \quad \text{TD}$$

$$= \sum_a \pi(a|s) \sum_{s',r} p(s', r \mid s, a)(r + \gamma V_\pi(s')) \quad \text{DP}$$

V_{t+1}

V_k

DP

$\{V_k\} \to V_\pi$ 自举

$$V_k = \begin{pmatrix} V_k(s_1) \\ V_k(s_2) \\ \vdots \\ V_k(s_n) \end{pmatrix}$$

必须要有P

$$V_1 = \begin{pmatrix} V_1(s_1) \\ \vdots \\ V_1(s_n) \end{pmatrix}$$

MC 采样

$$V_\pi(s) = E_\pi[G_t \mid S_t = s]$$

G_t

$S_0, A_t, R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2} \cdots$

必须走完才能得到 G_t

eg:

$$V_1(S_1) = V_1(S_1) + \alpha(G_t - V_1(S_1))$$

处于 S_1 的状态下，更新 $V(S_1)$，(假设在 V_1 第一轮.

G_t 真实　$V_1(S_1)$ 当前估计

TD

mocule

克服要等延迟的问题.

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

S_t	A_t	R_{t+1}	S_{t+1}	A_{t+1}	R_{t+2}	...	R_T	S_T

TD-0

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

先走一步，得到 R_{t+1}，不走了，直接去表里找 $V(S_{t+1})$

$$V(S_3) \leftarrow V(S_3) + \alpha(R_4 + \gamma V(S_4) - V(S_3))$$

R_4：真实　　从表里拿.

比如得到 $V(S_4)$

$V(S_1)$
$V(S_2)$
$V(S_3)$
$V(S_4)$

注意：这里的 $V(S_4)$ 并不是一定的

因为状态(下一个)是根据 π 抽样得到的

可能是 S_1、S_2、S_3、S_4 之中任何一个.

控制. 单纯的 $V(S)$ 是不能够修改 π 的. (greedy)

因此，先求 $Q(S, a)$

$$Q(S_t, A_t)$$

SARSA

进入这个状态，采样得到 a_6

S_3	a_3	2	S_5	a_6	⋮
S_t	A_t	R_{t+1}	S_{t+1}	A_{t+1}	R_{t+2}

$$= Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$$

(SARSA)

(同轨要新)

A_{t+1}：这是根据 π 随机抽样得到的

离轨轨九：

当我们进入 S_5 的时候，不由策略函数来采样。

我有 Q-table，我分别 $Q(S_5, a_1)$，我自己行！

$Q(S_5, a_2)$

找最大值的 action.

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \beta \max_a Q(S_{t+1}, a) - Q(S_t, A_t)]$$

Q-learning

同轨 SARSA （听教练的）

离轨 Q-learning （不听）

期望 SARSA

教练 (b) or (a)

S_{t+1}	A_{t+1}	P_r
S_5	a_1	0.7
S_5	a_6	0.3

Q-table

S	A	Q(s, a)
S_5	a_1	100
S_5	a_6	99

用期望来做为 ← （根据教练，a 来到 Q-table 找对应的 Q）

$100 \times 0.7 + 99 \times 0.3$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \beta (E_b [Q(S_{t+1}, a) | S_{t+1}] - Q(S_t, A_t))]$$

还是使用了 a，还是同轨的

~~$\pi(a|s_{t+1})$~~)

(I) 默认情况下 $\pi = b$，同轨

(II) 若 π 是贪心策略，也即 $\pi(s_{t+1}) = \arg\max_a Q(s_{t+1}, a) \triangleq a^*$

则期望为离轨策略

$$E_\pi [Q(s_{t+1}, a) | s_{t+1}] = \max Q(s_{t+1}, a)$$

because

$$\Rightarrow \pi(a|s_{t+1}) = \begin{cases} 1 & a = \arg\max_a Q(s_{t+1}, a) \\ 0 & \text{else} \end{cases}$$

Q-learning 是 期望 SARSA 的特例.