Decision Process

Q-Learning

$s_1 \rightarrow a_1$, $s_1 \rightarrow a_2$

Q-table

r	a1	a2
s1	-2	1

$Q(s1,a1) < Q(s1,a2)$ → ~~policy~~: choose a_2

state update

$s_2 \rightarrow a_1$, $s_2 \rightarrow a_2$

Q-table

r	a1	a2
s2	-4	2

$Q(s2,a1) < Q(s2,a2)$ → ~~policy~~: choose a_2

How Q-table update and improve

→ estimated Q-table →

$s_1 \rightarrow a_1$, $s_1 \rightarrow a_2$ ~~wrong~~ → In state s1, choose a_2 → imaginary s_2 → update Q-table (before making next move)

no real action → imagine: $\rightarrow a_1$, $\rightarrow a_2$

Q-table

r	a1	a2
s1	-2	1

according to previous Q-table, a_2 better than a_1

r	a1	a2
s1	-2	1
s2	-4	2

$R + \gamma * \max Q(s2)$

s2下最大的Q值

衰减值 attenuation coefficient : attenuated rate of future reward

reward of reaching s_2.

⇓

real $Q(s1,a2)$

recall estimated $Q(s1,a2)$

$d = |Q_{real}(s1,a2) - Q_{estimate}(s1,a2)|$

$newQ(s1,a2) = Q(s1,a2) + \alpha * d$

estimated

learning efficiency : decide how much of the difference will be studied this time.

$$Q(s,a) \leftarrow Q(s,a) + \alpha[(r + \gamma \max_{a'} Q(s',a')) - Q(s,a)]$$

↓

attenuated maximum estimated Q(s2)

Basically, we are using the maximum estimated value of next step to ~~approach reality~~ plus the reward as the reality value of this step

$Q(s1) = r2 + \gamma Q(s2) = r2 + \gamma[r3 + \gamma Q(s3)] = r2 + \gamma[r3 + \gamma[r4 + \gamma Q(s4)]]$

$\Rightarrow Q(s1) = r2 + \gamma r3 + \gamma^2 r4 + \gamma^3 r5 + \gamma^4 r6 + \cdots$

∴ If $\gamma = 1$, $Q(s) = r2 + r3 + r4 + \cdots$

which means seeing the future clearly without any attenuation, i.e. it can perfectly predict the future.

If $\gamma = 0$, $Q(s) = r2$.

which means it cannot see any future step but only cares about the most recent reward

If $0 < \gamma < 1$

means it see near future steps more clearly than remote future steps.

As the training moves forward, more real future steps are used to update the Q-table, the machine can better predict the near future

see beyond short-term benefit