Week 9: New V(s) and Q(s,a)

Saturday, April 2, 2022 9:57 AM

Solución: Iteración de Valor	$V(s) \leftarrow f_R(s, f_\pi(s), s_f) + \gamma V(s_f)$	$V(s) \leftarrow \sum_{s_f \in S} P_{M_T}(s_f \mid s, f_\pi(s)) [f_R(s, f_\pi(s), s_f) + \gamma V(s_f)]$

$V(s) = \overline{f_{AR}(\tau)}\Big|_{s(0)=s}$ we said we can write this function in a nondeterministic way, and also in a det form. Both cases lead us to a linear system of equations

↳ we can solve this either by an analytical way or a numerical way, which is the method : Value Iteration

We can apply this num method to either deterministic and non det.

We said V(s) is the average of acc reward of all trajectories starting on s.

↳ all this is done assuming we have a defined politic

but how to define the politic?

↳ the principle: Obtain the maximum average reward

↓

Bellman's idea: Bellman's Optimality Equation

↓

How?

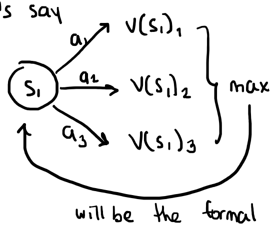
Instead of $V_s = r + \gamma V(s_f)$ we thus

$$\underbrace{V(s)}_{\text{new def of } V} = \underbrace{\max_a}_{\text{max with the action}} \left[\overline{f_R(s, a, s_f) + \gamma V(s_f)} \right]$$

this a iterates and becomes each and every action

} the maximum average of the accumulated reward (previous V(s)) for all availabe actions (s in V(S) for a would be the state that has the availabe actions)

Let's say



will be the formal $V(s_1)$

Thus, by obtaining the $\underbrace{\text{max } V(s)}$, the agent discovers the optimal politic.

↳ for all states

This $V(s) = \max_a [\text{Bellman}]$ creates a different equation system: it becomes a non-linear system of equations, since max() is not continuous. Therefore, to solve this nonlinear system we use a numerical method.

In literature, to avoid this, people defined :

$$V(s) = \max_a \left[\underbrace{\overline{f_R(s, a, s_f) + \gamma V(s_f)}} \right] \quad (1) \Rightarrow \quad s \xrightarrow{a} s_f$$

$$\overbrace{s \xrightarrow{a} s_f}^{Q(s,a)} \overbrace{\xrightarrow{a_f} s_{f2}}^{Q(s_f, a_f)}$$

$$Q(s,a) = \overline{f_R(s, a, s_f) + \gamma V(s_f)}$$

Considering $V(s_f) = \max_a \left[f_R(s_f, a_f, s_{f2}) + \gamma V(s_{f2}) \right]$

↳ state after s_f

So, by defining $V(s) = \max_a [Q(s,a)]$, thus we rewrite V(s) as Q(s,a)

Considering $V(s_f) = \max_a \left[f_R(s_f, a_f, s_{f2}) + \gamma V(s_{f2}) \right]$ ↳ state after s_f

So, by defining $V(s) = \max_a [Q(s,a)]$, thus we rewrite $V(s)$ as $Q(s,a)$

$$Q(s,a) = \left[f_R(s,a,s_f) + \gamma \overline{\max_a [Q(s_f,a_f)]} \right] \quad (2)$$

(1) and (2) are two ways of visualizing the same thing

The relationship between V and Q is

$$\underbrace{V(s)}_{\text{one}} = \max_a [\underbrace{Q(s,a)}_{\text{two variables}}]$$

In theory, $V(s,a)$, but we don't write it, since to V, "a" doesn't matter

only one action, which? the one that gave you the max
grabs the max Q and THAT is $V(s_1)$, or $Q(s_1, a_1)$

	V(s)	say
s1	V(s1)	a_1
s2	V(s2)	a_3
s3	V(s3)	a_2
s4	V(s4)	a_1
s(n=5)	V(s5)	a_1

	Q(s,a)		
	a1	a2	a(m=3)
s1	Q(s1,a1)	Q(s1,a2)	Q(s1,a3)
s2	Q(s2,a1)	Q(s2,a2)	Q(s2,a3)
s3	Q(s3,a1)	Q(s3,a2)	Q(s3,a3)
s4	Q(s4,a1)	Q(s4,a2)	Q(s4,a3)
s(n=5)	Q(s5,a1)	Q(s5,a2)	Q(s5,a3)

is the acc reward when executing a_1 when in S_2.

↳ optimal politic: it tells the agent the sequence of actions that will give the maximum benefit.

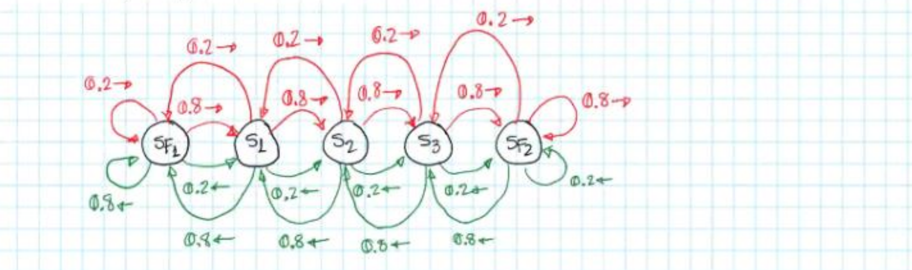
↳ in order to know this politic, we need to calculate the whole matrix for all a's in order to know the optimal a.

Thus, the problem is basically solving the nonlinear systems V(S) or Q(s,a) and the num method is once again the value iteration.

Example

S_{F_1}	S_1	S_2	S_3	S_{F_2}

a. Construya el grafo del mundo



b. Escriba la función de transición $P_{M_T}(s_f|s,a)$

$$P_{M_T}(s_f|s,a) = \begin{array}{c} \\ s_1\\ s_2\\ s_3\\ s_{F_1}\\ s_{F_2}\end{array}\!\!\overset{s_f=s_1}{\begin{array}{cc} \leftarrow & \rightarrow \\ 0 & 0 \\ 0.8 & 0.2 \\ 0 & 0 \\ 0.2 & 0.8 \\ 0 & 0 \end{array}} \quad \begin{array}{c} \\ s_1\\ s_2\\ s_3\\ s_{F_1}\\ s_{F_2}\end{array}\!\!\overset{s_f=s_2}{\begin{array}{cc} \leftarrow & \rightarrow \\ 0.2 & 0.8 \\ 0 & 0 \\ 0.8 & 0.2 \\ 0 & 0 \\ 0 & 0 \end{array}} \quad \begin{array}{c} \\ s_1\\ s_2\\ s_3\\ s_{F_1}\\ s_{F_2}\end{array}\!\!\overset{s_f=s_3}{\begin{array}{cc} \leftarrow & \rightarrow \\ 0 & 0 \\ 0.2 & 0.8 \\ 0 & 0 \\ 0 & 0 \\ 0.8 & 0.2 \end{array}} \quad \begin{array}{c} \\ s_1\\ s_2\\ s_3\\ s_{F_1}\\ s_{F_2}\end{array}\!\!\overset{s_f=s_{F_1}}{\begin{array}{cc} \leftarrow & \rightarrow \\ 0.8 & 0.2 \\ 0 & 0 \\ 0 & 0 \\ 0.8 & 0.2 \\ 0 & 0 \end{array}} \quad \begin{array}{c} \\ s_1\\ s_2\\ s_3\\ s_{F_1}\\ s_{F_2}\end{array}\!\!\overset{s_f=s_{F_2}}{\begin{array}{cc} \leftarrow & \rightarrow \\ 0 & 0 \\ 0 & 0 \\ 0.2 & 0.8 \\ 0 & 0 \\ 0.2 & 0.8 \end{array}}$$

$$V(s) = \max_a \left[f_R(s,a,s_f) + \gamma V(s_f) \right]$$ Bellman's Optimality Eq.

First Bellman's Eq for optimal politic:

$$V(s_1) = \max_a \left[\underbrace{f_R(s_1, \leftarrow, s_f) + \gamma V(s_f), \; f_R(s_1, \rightarrow, s_f) + \gamma V(s_f)}_{\text{iterate each possible action}} \right]$$

stochastic: s_f is either one that is s_f

This new V(S) is called: V(S) under Bellman's Optimality, different from the other V(S)

$$= \max_a \left[\underbrace{\sum_{s_f \in S} P_{MT}(s_f | s_1, \leftarrow)[f_R(s_1, \leftarrow, s_f) + \gamma V(s_f)]}_{\text{action: } \leftarrow \text{ (5 terms in the sum)}}, \; \sum_{s_f \in S} P_{MT}(s_f | s_1, \rightarrow)[f_R(s_1, \rightarrow, s_f) + \gamma V(s_f)] \right]$$

$$= \max_a \left[P_{MT}(s_2 | s_1, \leftarrow)[f_R(s_1, \leftarrow, s_2) + \gamma V(s_2)] + P_{MT}(s_{F_1} | s_1, \leftarrow)[f_R(s_1, \leftarrow, s_{F_1}) + \gamma V(s_{F_1})], \right.$$

$$\left. P_{MT}(s_2 | s_1, \rightarrow)[f_R(s_1, \rightarrow, s_2) + \gamma V(s_2)] + P_{MT}(s_{F_1} | s_1, \rightarrow)[f_R(s_1, \rightarrow, s_{F_1}) + \gamma V(s_{F_1})] \right]$$

$$= \max_a \left[0.2[-0.4 + \gamma V(s_2)] + 0.8[-10 + \gamma V(SF_1)],\ 0.8[-0.4 + \gamma V(s_2)] + 0.2[-10 + \gamma V(SF_2)] \right] \quad (1)$$

↳ we will have 5 equations like this

Apart from the numerical method (value iteration), there is another method: Temporary Difference

We will see that in the real cases we do not know the Transition Model Function nor the reward function. Those can be <u>learned</u>, though. Only then will we discover the politic: optimal politic $f_\pi(s)$, that will give us the highest reward posible.