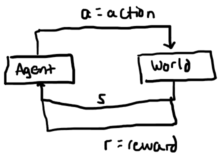
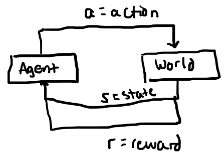
Week 5

Saturday, March 5, 2022 10:34 AM



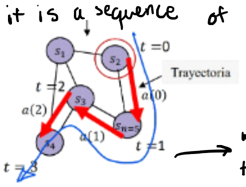
a loop for the desired time of simulation

The input space can be either continuous or discrete
The box is the input space



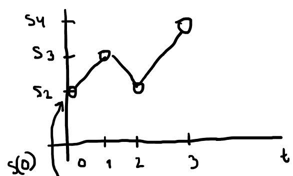
Assuming we defined the World, we analyze the agent

→ Trajectory: they are usually time functions
Since our world is discrete, a trajectory it is a sequence of transitions



→ world is discrete
time is also discrete

We usually describe a trajectory with a function of time: but f(t) is discrete (states) and time is also discrete

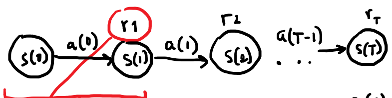


→ But usually people write the trajectory as τ

$\tau = s(0), s(1), s(2), \dots s(T)$ → initial state

$= s_2, s_3, \dots$ } a particular trajectory

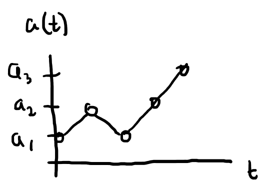
→ Another way to write a trajectory is a graph:



a transition

a reward is given at the end of transitions

a(t): actions are also functions of time:



r(t): also a function of time

belong to the trajectory

Accumulated reward: how much reward an agent accumulated in a trajectory → until the end of the trajectory

$$s(0) \xrightarrow{a(0)} s(1) \xrightarrow{a(1)} s(2) \; \cdots \xrightarrow{a(T-1)} s(T)$$

$r_1 + r_2 + \dots + r_T$ = accumulated reward

$$f_{AR}(T) = \sum_{t=0}^{T-1} \gamma^t f_R(s(t), a(t), s(t+1))$$

f_R ↳ reward function → $f_R(s(t), a(t), s(\text{next state}))$ ↓ gives $r_{(t+1)}$ = reward of current time +1

with $r_{(t+1)} = f_R(s(t), a(t), s(t+1))$

$\gamma \in [0,1] \rightarrow$ usually $\gamma \neq 1$

$$= r_1 + \gamma r_2 + \gamma^2 r_3 + \dots + \gamma^{(T-1)} r_T$$

NOTE: if $t \to \infty$, to make the sum converge:

this factor is to secure convergence, but also to soften the rewards in the future more than the closer rewards

$$\rightarrow \underbrace{\sum_{n=0}^{N} a x^n}_{n=t} = a\left[\frac{1-x^{(N+1)}}{1-x}\right], x \neq 1$$

↳ if $n \to \infty$, the $\lim_{n\to\infty} \Sigma \approx 0$ } converges to zero

The accumulated reward is thus $f_{AR}(\tau)$ ↳ trayectory

To write reward like this gives us a property and allows us to take γ factor out of $= r_1 + \gamma r_2 + \gamma^2 r_3 + \dots + \gamma^{(T-1)} r_T$ gives us

$$= r_1 + \gamma\left[r_2 + \gamma r_3 + \dots + \gamma^{(T-2)} r_T\right]$$ → which gives us another acc reward function, but that starts in $s(1)$

$$= r_1 + \gamma\left[\sum_{t=1}^{T-1} \gamma^{(t-1)} f_R(s(t), a(t), s(t+1))\right]$$

an important property: $\bar{\tau}$ starts in $t=1$

$$= r_1 + \gamma\left[f_{RA}(\bar{\tau})\right] \text{ with } \bar{\tau} = s(1), s(2), \dots s(T)$$

↳ starts in $t=1$ or r_2

We take this functional form so that when $T \to \infty$, the sum converges as mentioned before.

If we have a world that is deterministic in all ways, everytime we start at $s(0)$ we have the same reward

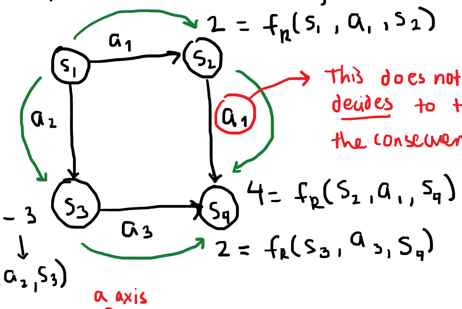
Otherwise, (stochastic) we calculate the mean of the rewards of all possible trayectories. But if the number of trayectories tends to infinite, we need to approximate this mean of rewards.

Example:

A totally deterministic world,

$S = \{s_1, s_2, s_3, s_4\}$

$A = \{a_1, a_2, a_3\}$



This does not mean an agent in S_2 decides to take a_1. It just means the consecuence of in case of a_1.

↓ we need to define an action function for the agent $f_\pi(s)$

$f_R(s, a, s_f) =$ (a axis →, s axis ↓)

$s_f = s_1$

	a_1	a_2	a_3
s_1			
s_2			
s_3			
s_4			

s_1 is never a final state, empty

$s_f = s_2$

	a_1	a_2	a_3
s_1	2		
s_2			
s_3			
s_4			

$s_f = s_3$

	a_1	a_2	a_3
s_1		-3	
s_2			
s_3			
s_4			

$s_f = s_4$

	a_1	a_2	a_3
s_1			
s_2	4		
s_3			2
s_4			

} remaining cells are left empty, but they can have anything

$f_\pi(s) =$

s	a
s_1	a_1
s_2	a_1

} The politic (deterministic)

$$f_\pi(s) = \begin{array}{c|c|} & a \\ s_1 & a_1 \\ s_2 & a_1 \\ s_3 & a_2 \\ s_4 & a_3 \end{array}$$

The politic (deterministic)

$\tau = s_1, s_2, s_4$

$$s_1 \xrightarrow{a_1} \overset{2}{s_2} \xrightarrow{a_1} \overset{4}{s_4}$$

↳ it can finish any where

γ (gamma) $= 1$ (i.e.)

$f_{AR} = 2 + \gamma(4)$
$= 2 + 1(4)$
$= 2 + 4$
$\boxed{= 6}$ → Thus, starting in s_1 the $f_{AR} = 6$ everytime, mean $= 6$

i.e. Another politic:

$$f_\pi^2(s) = \begin{array}{c|c|} & a \\ s_1 & a_2 \\ s_2 & a_1 \\ s_3 & a_3 \\ s_4 & a_3 \end{array}$$

→ it is done, but there's no defined transition ∴ it stays in s_4

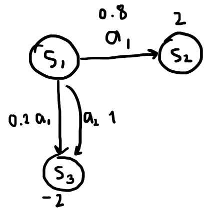
$$s_1 \xrightarrow{a_2} \overset{-3}{s_3} \xrightarrow{a_3} \overset{2}{s_4}$$

$f_{AR} = (-3) + \gamma(2)$ $\gamma = 1$
$= (-3) + 1(2)$
$= -3 + 2$
$\boxed{= -1}$

trajectory: $\tau = s_1, s_3, s_4$

which politic is better? The one that gives us more f_{AR}

→ In the case of an stochastic world,



Let's assume that the reward function only depends on s_f instead of s, a, s_f:

$$f_R(s, a, s_f) = f_R(s, a, s_f) = \begin{array}{c|c|} s_1 & 0 \\ s_2 & 2 \\ s_3 & -2 \end{array}$$

s_f

Politic: the action function, $f_\pi(s)$:

$$f_\pi(s) = \begin{array}{c|c|} s_1 & a_1 \\ s_2 & a_2 \\ s_3 & a_3 \end{array}$$

deterministic, the only non-determ is the transitions in the world

$$s_1 \xrightarrow[a_1]{0.8} s_2 \qquad f_{AR} = 2$$

$$s_1 \xrightarrow[a_1]{0.2} s_3 \qquad f_{AR} = -2$$

Since it is stochastic
↓
The mean?

$\overline{f_{AR}} = 2(0.8) + (-2)(0.2)$
$= 1.2$

to decide which politic is better, now we compare the means $\overline{f_{AR}}$

→ Another politic here,

$$f_\pi(s) = \begin{array}{c|c|} s_1 & a_2 \\ s_2 & a_1 \\ s_3 & a_2 \end{array}$$

$$s_1 \xrightarrow{a_2} \overset{-2}{s_3}$$

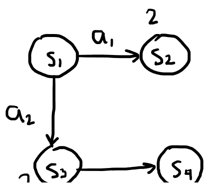
$f_{AR} = -2$

$\overline{f_{AR}} = -2$

$f_\pi(s) \Rightarrow$ the amount of politics is 2^3 in this case, actions^(states)

Which politic is better? The first, since $f_\pi^1(s) = -2$, 80% of the times is -2, and in the second, 100% of the times we lose -2.

Another world

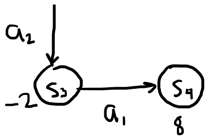


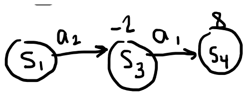
Trajectories

$$s_1 \xrightarrow{a_1} \overset{2}{s_2} \qquad f_{RA} = 2$$

$$s_1 \xrightarrow{a_2} \overset{-2}{s_3} \xrightarrow{a_1} \overset{8}{s_4} \qquad f_{RA} = (-2) + \gamma(8)$$

$\gamma = 1$ → punishment for





$f_{RA} = (-2) + \gamma(8)$

ex. $\gamma = 1 \rightarrow$ punishment for distance

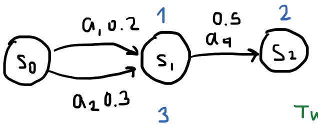
$= (-2) + (1)8$

$= -2 + 8 = 6$

Question: How to calculate f_{AR} in stochastic worlds? You can have two cases:

$\gamma = 0.7$

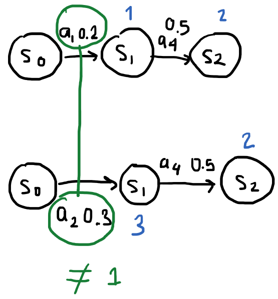
a)



Two trajectories from S_0 to S_2:

Two possible ways and it doesn't state on specific probability

= each 0.5



$f_{AR} = 1 + \gamma(2)$
$= 1 + (0.7)2$
$= 1 + 1.4 = 2.4$

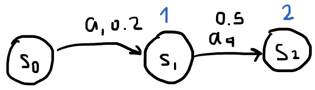
$f_{AR} = 3 + \gamma(2)$
$= 3 + (0.7)2$
$= 3 + 1.4 = 4.4$

Therefore the f_{AR} is $\underbrace{(0.5)(2.4) + (0.5)(4.4)}_{1}$

$= \boxed{\frac{2.4 + 4.4}{2}}$ = the average of the two

$\gamma = 0.7$

b)



Only one possible trajectory

$f_{AR} = 1 + \gamma(2)$
$= 1 (0.7)(2)$
$= 1 + 1.4 = \boxed{2.4}$