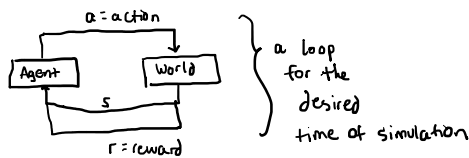


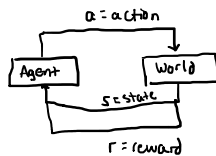
Week 5

Saturday, March 5, 2022 10:34 AM



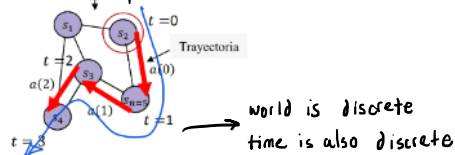
The input space can be either continuous or discrete

The box is the input space

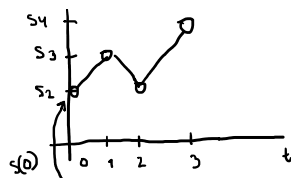


Assuming we defined the World, we analyze the Agent

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

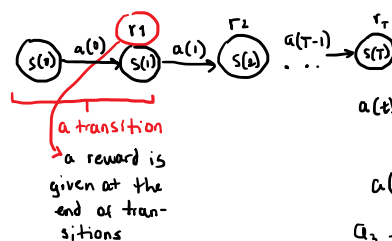


We usually describe a trajectory with a function of time: but $f(t)$ is discrete (states) and time is also discrete

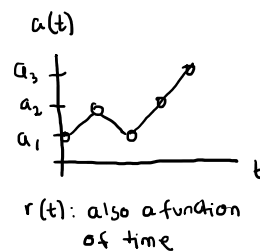


→ But usually people write the trajectory as T
 $T = s(0), s(1), s(2), \dots, s(T)$
initial state
= s_2, s_3, \dots } a particular trajectory

→ Another way to write a trajectory is a graph:



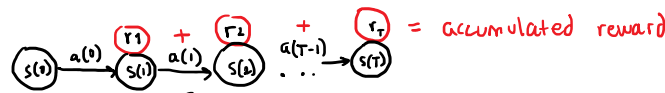
$a(t)$: actions are also functions of time:



belong to the trajectory

$r(t)$: also a function of time

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



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

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

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

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

\downarrow gives r_{t+1} = reward at current time + 1

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

$$\sum_{n=0}^{\infty} \alpha x^n = \alpha \left[\frac{1 - x^{(N+1)}}{1 - x} \right], x \neq 1$$

if $n \rightarrow \infty$, the $\lim_{n \rightarrow \infty} \sum \approx 0$ } converges to zero

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

The accumulated reward is thus $f_{AR}(T)$ $\xrightarrow{\text{trajectory}}$

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

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

\rightarrow which gives us another acc reward function, but that starts in $s(1)$

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

\leftarrow starts in $t=1$ or r_2

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

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

We take this functional form so that when $T \rightarrow \infty$, 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

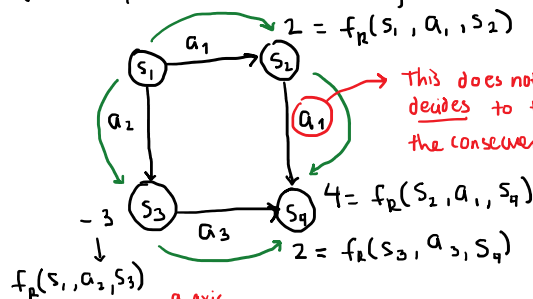
Otherwise, (stochastic) we calculate the mean of the rewards of all possible trajectories. But if the number of trajectories tends to infinite, we need to approximate this mean of rewards.

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 consequence of in case of a_1 .

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

	a axis		
	a_1	a_2	a_3
s_1	2	-3	
s_2			
s_3			
s_4			2

$S_f = S_1$ $S_f = S_2$ $S_f = S_3$ $S_f = S_4$

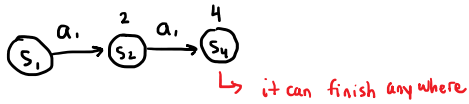
S_1 is never a final state, empty

remaining cells are left empty, but they can have anything

$$f_\pi(s) = \begin{matrix} a \\ s_1 \begin{bmatrix} a_1 \end{bmatrix} \\ s_2 \begin{bmatrix} a_1 \end{bmatrix} \end{matrix} \left. \vphantom{\begin{matrix} a \\ s_1 \begin{bmatrix} a_1 \end{bmatrix} \\ s_2 \begin{bmatrix} a_1 \end{bmatrix} \end{matrix}} \right\} \text{the policy (deterministic)}$$

$$f_{\pi}(s) = \begin{matrix} & a \\ s_1 & a_1 \\ s_2 & a_1 \\ s_3 & a_2 \\ s_4 & a_3 \end{matrix} \quad \left. \vphantom{\begin{matrix} & a \\ s_1 & a_1 \\ s_2 & a_1 \\ s_3 & a_2 \\ s_4 & a_3 \end{matrix}} \right\} \begin{array}{l} \text{The politic} \\ \text{(deterministic)} \end{array}$$

$\mathcal{T} = s_1, s_2, s_4$



$$\gamma(\text{gamma}) = 1 \quad (\text{i.e.})$$

$$f_{AR} = 2 + \gamma(4)$$

$$= 2 + 1(4)$$

$$= 2 + 4$$

$$\boxed{= 6} \rightarrow \text{Thus, starting in } s_1, \text{ the } f_{AR} = 6 \text{ everytime, mean} = 6$$

i.e. Another politic:

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

$$s_1 \xrightarrow{a_2} s_3 \xrightarrow{a_3} s_4$$

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

$$= (-3) + 1(2)$$

$$= -3 + 2$$

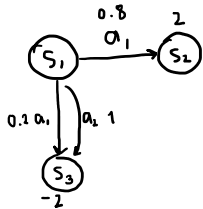
$$\boxed{= -1}$$

trajectory:

$$\mathcal{T} = s_1, s_3, s_4$$

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

which politic is better? The one that gives us more f_{AR}
→ In the case of an stochastic world,



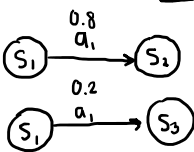
Let's assume that the reward function only depends on s_t instead of s_t, a_t, s_{t+1} :

$$f_R(s, a, s_t) = f_R(s, a, s_t) = \begin{matrix} s_1 & 0 \\ s_2 & 2 \\ s_3 & -2 \end{matrix}$$

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

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

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

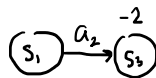


$$\left. \begin{array}{l} f_{AR} = 2 \\ f_{AR} = -2 \end{array} \right\} \begin{array}{l} \text{Since it is stochastic} \\ \downarrow \\ \text{The mean?} \\ \overline{f_{AR}} = 2(0.8) + (-2)(0.2) \\ = 1.2 \end{array}$$

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

→ Another politic here,

$$f_{\pi}(s) = \begin{matrix} & a \\ s_1 & a_2 \\ s_2 & a_1 \\ s_3 & a_2 \end{matrix}$$



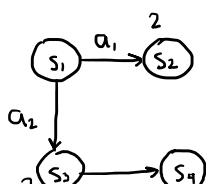
$$f_{AR} = -2$$

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

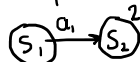
$f_{\pi}(s) \Rightarrow$ the amount of politics is 2^3 in this case, actions[^](states)

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.

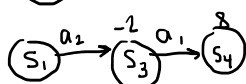
Another world



Trajectories

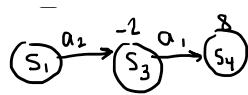
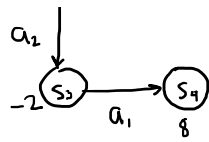


$$f_{RA} = 2$$



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

... $\gamma = 1 \Rightarrow$ punishment for



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

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

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

$$= -2 + 8 = 6$$