Reinforcement Learning

Returns

The agent learns a policy

- Policy at step t, $\boxed{\pi_t}$ which is a mapping from state to action probabilities, $\pi(s,a)$ = prob. that $a_t = a$ when $s_t = s$.

- "Reinforcement ~~algo~~ learning" is not a single algorithm, it is a learning methods that represents how an agent changes its policy as a result of experience. That means the agent's goal is to get as much reward as it can over the long run. i.e the goal of an agent is to maximize its returns.

- Suppose the sequence of reward after step t is $r_{t+1}, r_{t+2}, r_{t+3} \ldots$ we want to maximize the return i.e total reward $\boxed{G_t}$ for every step t

Next term with related to return is important →

"Episodic Task":

Interaction breaks naturally into "episode" as for example "MazeGame".

$G_t = \gamma_{t+1} + \gamma_{t+2} + \gamma_{t+3} \cdots + \gamma_T$ → final step (terminal state)

i.e an end to the episode.

However, there is a possibility of "Return for Continuing Task"

Cont. Task: no natural episodes (for interactions)

∴ Discounted return with the introduction of discounted ~~factor~~ rate γ

∴ The new, $G_t = \gamma_{t+1} + \gamma\gamma_{t+2} + \gamma^2\gamma_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k \gamma_{t+k+1}$

where, γ, $0 \le \gamma \le 1$; γ = discounted factor.

γ is a controlling factor

$\gamma \to 0$ || shortsighted

$\gamma \to 1$ || farsighted

Main intention is to maximize the expected return, $E\{G_t\}$ for each step t.

[$\gamma = 0$ is a immediate reward problem? NO as next state that we are going to see ~~is~~ depends on my current actions.]

Value Functions

∴ How to maximize the expected "Returns" ? at every timestep t..

- To achieve this maximization process we are introducing "Value Function":
  - Expected future rewards
  - Start state (or state-action pair)
  - & following policy π

State-value function for policy π:

$$V_\pi(s) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k \cdot R_{t+k+1} \,\middle|\, S_t = s\right]$$

(G_t : $\sum_{k=0}^{\infty} \gamma^k \cdot R_{t+k+1}$)

t (transition)

S_t —R_{t+1}→ S_{t+1} —R_{t+2}→ S_{t+2} ...

$a_t \sim \pi(\cdot|s)$; $A_{t+1} \sim \pi(\cdot|S_t)$

{ sampled from π

$(S_{t+1}, R_{t+1}) \sim p(\cdot, \cdot | S_t, A_t)$ (sampled from both)

E_π : Expected value = Average future outcome considering probabilities of all possibilities (several trajectories) (average outcome in a related process)

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

Action-value function for policy π:

$$q_\pi(s, a) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

State

Action-value function for policy π:

$$q_\pi(s,a) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

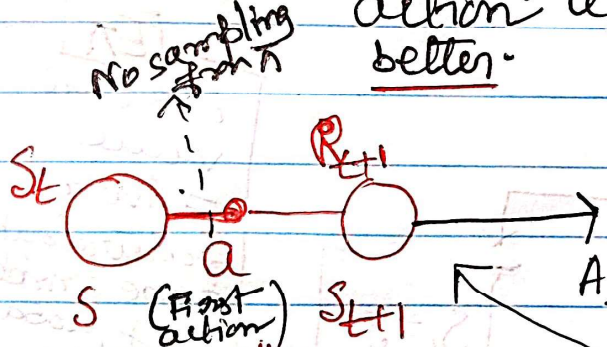
(a: action)

★ State value f^n

- prev. one state-value
- This above is related, is called action value.
- State value: returns expected value from state S if following policy π there after.
- Purpose: which states are better under a fixed policy. Gives the long term value of being in a state.

★★ Action value f^n /

- What if from single state action a_1 or action a_2 is better.



→ [No sampling from π]

$\therefore (S_{t+1}, R_{t+1}) \sim p(\cdot,\cdot|s,a)$ [not S_t a_t it is s,a]

Sampled from

A_{t+1} has been sampled from π after only started from S & action a has been taken

In here, $q_\pi(s,a) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$ ← γ^k

$\therefore$ o better among $q_\pi(s,a_1)$ or $q_\pi(s,a_2)$ } action that policy π shall be followed

Comparisons among actions

* change actions

∴ from a particular state a_1 or a_2: which one is better

$\therefore E_\pi[G_t | S_t=s, A_t=a] \rightarrow$ Two value f^ns
- State value f^n
- Action value f^n.
- Most of time q_π is better than v_π f^n

$= q_\pi(s,a)$

Although these two f^ns are different, they however have a relationship

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a)$$
$$= E_\pi[q_\pi(s,a)]$$

selecting action based on policy π.

But q_π interms of v_π?

∴ from a particular state a_1 or a_2: which one is better

$\therefore E_\pi[G_t | S_t = s, A_t = a]$ → Two value fn
- State value fn
- Action value fn.
- Most of time q_π is better than v_π fn

$= q_\pi(s,a)$

Although these two fn's are different, they however have a relationship

$$v_\pi(s) = \sum_a \pi(a|s)\, q_\pi(s,a)$$

$$= E_\pi[q_\pi(s,a)]$$

selecting action based on policy π.

But q_π integrals of v_π?

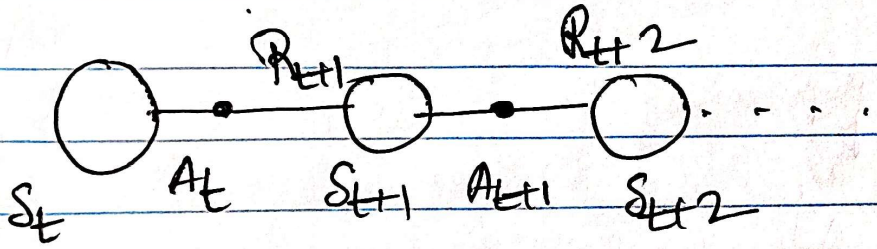
Value fn

$$V_\pi = E_\pi[G_t | S_t = s] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right], \forall s \in S$$

G_t ← Expected value of return $\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right]$ (G_t = G_k)

For a Policy π : Bellman Equation

Let's see, $G_k = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = R_{t+1} + \gamma[R_{t+2} + \gamma R_{t+3} + \cdots]$ ☆



Return at t ⟶ start adding up from R_{t+1}
Return at $t+1$ ⟶ start summing " " R_{t+2}
Return at $t+2$ ⟶ " " " " R_{t+3}
∴ G_t can be written recursively as it express the total future reward as the current reward (R_{t+1}) + discounted future rewards (γG_{t+1}). It means that the return at time t depends on the return at the next time step, linking all future rewards together.

☆ ∴ $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \cdots$
$= R_{t+1} + \gamma [R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} \cdots]$

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

Now let's see the value of v_π

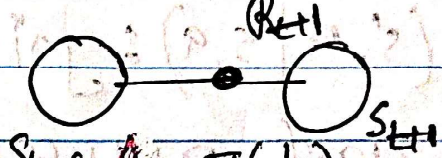
$$V_\pi(s) = E_\pi[G_t | S_t = s]$$

E_π = Expectation

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

$$= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r | s, a)\left[r + \gamma E_\pi[G_{t+1} | S_{t+1} = s']\right]$$

$$= \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a)\left[r + \gamma V_\pi(s')\right], \quad \forall s \in S$$

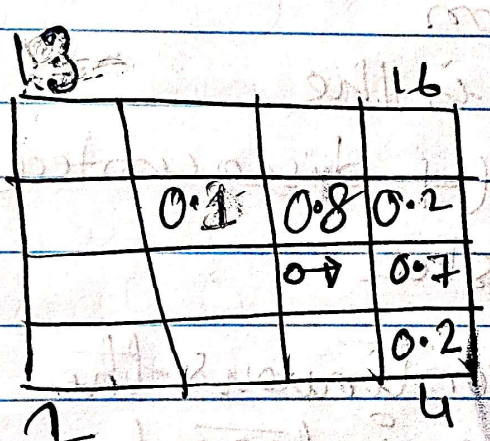


$$(S_{t+1}, R_{t+1}) \sim Pr(\cdot, \cdot | s, A_t)$$

↑ probability distribution

↓ start state 's' is fixed

Grid World



policy → $\pi(\uparrow|s) = 0.5$ [up]

$\pi(\rightarrow|s) = 0.5$ [right]

Two possible actions

let's say up $Pr(\cdot, \cdot | s, \uparrow)$ [we have to sample state]

↑ prob. distribution

let, say current state is 7 i.e. $S = 7$

$$V_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r | s, a) [r + \gamma V_\pi(s')]$$

Explanation:

- $\pi(a|s)$: probability of choosing action a in state s
- $p(s', r | s, a)$: joint prob of next state s' and reward r given (s, a)
- The inner expression $[r + \gamma V_\pi(s')]$ is the immediate reward + discounted value of the next state
- If rewards are continuous the $\sum$ becomes an integral

Here the expectation is computed

1. start at s (fixed)
2. Sample $a \sim \pi(\cdot|s)$
3. Sample $(s', r) \sim p(\cdot, \cdot | s, a)$
4. The contribution is $r + \gamma V_\pi(s')$
5. Average (sum) over all possible a, s', r

Only first step of the probabilities are explicitly expanded in the Bellman equation & the remainder of the trajectory

Linear system & solvability

- For a finite MDP with N states, the Bellman expectation eq^{n}s form an $N \times N$ linear system whose unknowns are the N value-function entries $\{V_\pi(s_i)\}_{i=1}^{N}$

- N states & N unknown values, $V_\pi(s_1), V_\pi(s_2) \cdots V_\pi(s_N)$

- The Bellman Expection equation gives you N equations one for each state

- There is a unique sol^n for V_π (assumption $0 \leq \gamma < 1$)

- Convert infinite-sum-return problem to tractable linear algebra problem.