Recap :- Bandits → ϵ-greedy & UCB.

MDP :-

- Agent, Environment.
- Sequential Decision Making.
  - trajectory:

$$(s_0, a_0, r_1, s_1, a_2, r_2, s_2, \dots).$$

- Bandit vs MDP :-

$$s, \underbrace{(a_1, a_2, \dots a_n)}_{\text{best arm.}}, r.$$

$$P(s|s,a_i) = 1$$

- MDP :- Sequential decision making

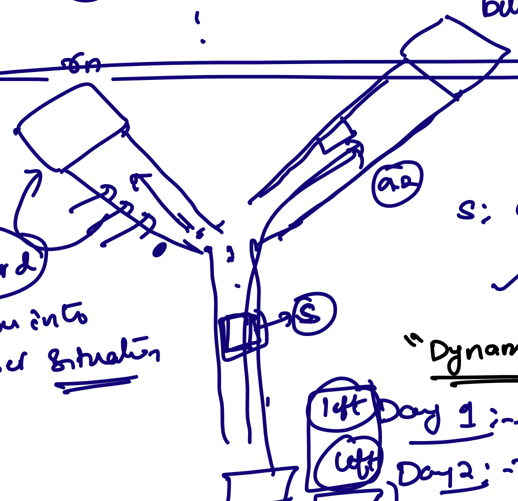
  (b) Find what is the best action/decision to make.

"long-run goal"

$$s: \begin{array}{l} a_1 \to r_1 \to s' \\ a_2 \to r_2 \to s'' \\ a_3 \to r_3 \\ \vdots \end{array}$$

greedily select action that has highest reward.

better than other road.



s, a → high reward

↳ leads you into a worser situation

s: a_1 → 100 ~> -1, -1, -1. .

a_2 → 1 ~> 2, 1, 1, 2 .

"Dynamic".

left Day 1 :- 100 + 1 + -1 + -1 + -1 + -1

left Day 2 :- 100 + 1 + 1 + -1 - 1 - 1.

right Day 3 :- 1 + 1 + + + . . .

obJ:- select actions that has higher long-run rewards.

Model-Known Setting:-

$r : S \times A \times S'$

$P(S' \mid s, a)$:

long-run : G_t :- $[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots]$

Random? of G_t?

assume rewards are bounded $|R_t| \le M$.

$\le M\left(\frac{1}{1-\gamma}\right)$

Machine Replacement:-

(*). $\boxed{r : S \times A \to \mathbb{R}}$

Policy: $\pi : S \to A$; $\pi : S \times A \to \mathbb{R}$ s.t. $\sum_{a} \pi(s,a) = 1$.

deterministic

stochastic

$\pi(s, a_1) = 0.6$

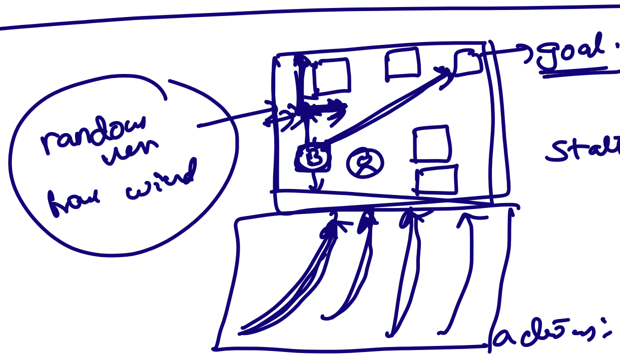
$\pi(s, a_2) = 0.4$.

G_0 :- $s_0, a_0, s_1, a_1, \ldots$

$P(\cdot \mid s_0, a_0)$ ← P.T.M.

$\mathbb{E}_{\substack{s_1, s_2, \ldots \\ a}} \left[r(s_0, a_0) + \gamma r(s_1, a_1) + \gamma^2 r(s_3, a_3) + \cdots \right]$

fix actions.



State: [current distance from nearest obstacle ; distance from goal]

actions: L, R, B, F

1^{st} trajectory, s_0, left, s_1, left,

γ

$$(*).\quad \boxed{V^{\pi}(s)} = E\Big[\underbrace{\sum_{t=0}^{\infty}}_{\text{total}} \underbrace{\gamma^t}_{\text{discount}} \underbrace{r(s_t,a_t)}_{\text{reward}} \,\Big|\, \underbrace{s_0=s}_{\text{starting from state } s},\ \underbrace{a_t=\pi(s_t)}_{\text{following the policy } \pi}\Big].$$

$\boxed{s_1, s_2, s_3 \ldots}$

Assume there is only one policy.

$\boxed{\text{single action}}$ in every state.

$\boxed{A = \{a\}.}$

$\boxed{r(s) = r(s,a).}$

$$V(s) = \underset{(s_1,s_2,\ldots)}{E}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t)\,\Big|\, s_0=s\Big]$$

$\gamma^0 r(s).$

$\boxed{r(s)}$

$$= \sum_{(s_1,s_2,\ldots)} Pr\{s_1=s_1, s_2=s_2, \mid s_0=s\} \sum_{t=0}^{\infty}\gamma^t r(s_t).$$

$$= r(s) + \sum_{(s_1,s_2,\ldots)} Pr\{s_1=s_1, s_2=s_2, \ldots \mid s_0=s\}.(\gamma r(s_1)+\gamma^2 r(s_2)+\gamma^3 r(s_3)+\ldots)$$

$$= r(s)+\gamma\sum_{(s_1,s_2,\ldots)} Pr\{s_1=s_1, s_2=s_2, \ldots \mid s_0=s\}\,[r(s_1)+\gamma r(s_2)+\ldots]$$

$$Pr\{\underbrace{s_1=s_1}_{A}, \underbrace{s_2=s_2, \ldots}_{B} \mid \underbrace{s_0=s}_{C}\} \overset{?}{=}$$

Markov property

$$Pr\{s_2=s_2, s_3=s_3, \ldots \mid s_1=s_1, \cancel{s_0=s}\}.$$

$$Pr\{s_1=s_1 \mid s_0=s\}.$$

$$Pr\{A,B \mid C\} = Pr\{A \mid C\} * Pr\{B \mid A,C\}.$$

$$= r(s)+\gamma\sum_{(s_1,s_2,\ldots)} \boxed{Pr\{s_1=s_1 \mid s_0=s\} * Pr\{s_2=s_2, s_3=s_3 \mid s_1=s_1\}}\,[r(s_1)+\gamma r(s_2)+\cdots]$$

$$= r(s) + \gamma \sum_{s_1} \Pr\{s_1 = s_1 \mid s_0 = s\} \cdot \sum_{(s_2, s_3, \ldots)} \Pr\{s_2 = s_2, s_3 = s_3, \ldots \mid s_1 = s_1\} \cdot [r(s_1) + \gamma r(s_2) + \gamma^2 r(s_3) + \ldots]$$

all the states that can occur at time '1'.

$$= \boxed{r(s) + \gamma \sum_{s'} \Pr\{s = s' \mid s_0 = s\} \cdot V(s')}$$

π : deterministic.

$$V^{\pi}(s) = r(s, \pi(s)) + \gamma \sum_{s'} \Pr(s' \mid s, \pi(s))\, V^{\pi}(s').$$

↳ Bellman equation

$$\boxed{Q^{\pi}(s, a)} = r(s, a) + \gamma \sum_{s'} \Pr(s' \mid s, a) \cdot V^{\pi}(s').$$

↳ Q-Bellman eqn.

π : stochastic :

$$V^{\pi}(s) = \boxed{\sum_{a \in A} \pi(a \mid s)} \left[r(s, a) + \gamma \sum_{s'} \Pr(s' \mid s, a) \cdot V^{\pi}(s') \right]$$

$$** \quad Q^{\pi}(s, a) = r(s, a) + \gamma \sum_{s'} \Pr(s' \mid s, a) \cdot V^{\pi}(s'). \longrightarrow (1)$$

$$V^{\pi}(s) = \boxed{\sum_{a \in A} \pi(a \mid s) \cdot Q^{\pi}(s, a).}$$

$\pi(s) \to a_1 \to 1/2$, $a_2 \to 1/2$

$$\boxed{\tfrac{1}{2} \cdot Q^{\pi}(s, a_1) + \tfrac{1}{2} Q^{\pi}(s, a_2).}$$

$$\boxed{V^{\pi}(s) = \sum_{a \in A} \pi(a \mid s) \cdot Q^{\pi}(s, a).} \longrightarrow (2)$$

w.r.t given policy π.

$$\boxed{Q^{\pi}(s,a) = r(s,a) + \gamma \sum_{s',a'} \Pr\{s' \mid s,a\} \cdot \pi(a' \mid s') \cdot Q^{\pi}(s',a')} \rightarrow (3)$$

→ Bellman eqn for optimal policy

→ Policy Iteration / V.I.

Solve practice