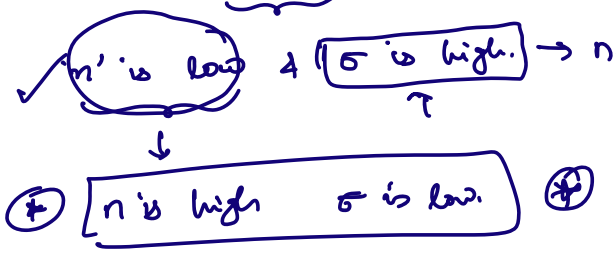
(1). (*). UCB → intelligent exploration.



(2). $\pi P = \pi$ ✓ $P\pi = \pi$

$$[x \quad y] \begin{bmatrix} 1/3 & 2/3 \\ 1/2 & 1/2 \end{bmatrix} = [x \quad y].$$

$x + y = 1$

$$\frac{x}{3} + \frac{y}{2} = x$$

$$\Rightarrow \frac{y}{2} = \frac{2x}{3} \Rightarrow y = \frac{4x}{3}$$

$$x + y = 1 \Rightarrow x + \frac{4x}{3} = 1 \Rightarrow x = 3/7.$$

$$y = \frac{4}{7}.$$

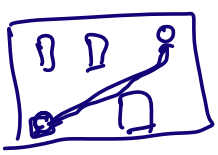
(3A) A : 1000 w.p $1/1000$.

B : (2)

$$E[r_A] = 1 \; ; \; E[r_B] = 2.$$ ✓

(4). (C). Episodic ends after fixed no. of timesteps.

random



(5). $V_\pi(s) = \sum_a \pi(a|s) \cdot Q^\pi(s,a)$. ∀s

$V_\pi(s_1) = \frac{1}{3}(2+4+6) = 4.$

$V_\pi(s_2) = \frac{1}{3}(1+5+3) = 3.$

$\therefore V_\pi(s_1) + V_\pi(s_2) = 7.$

Practice Questions :-

(1). (a). $E[r^i] = (1-p_i)\,p_i + 0\,(1-p_i)$

$= (1-p_i)\,p_i$

(b). $f(p) = (1-p)p$ ✓

$1-2p = 0 \Rightarrow p = 1/2$

$-2 < 0$ ✓

Arm 1 : $\frac{3}{4} \cdot \frac{1}{4}$.

2 : $\frac{1}{3} \cdot 2/3$

3. $1/4$ ✓

4.

5.

(2). Policy Iteration :-

1. Evaluating a given policy (π) : ← stochastic.

$s, a \rightarrow r$ is fixed

$$V^\pi(s) = \sum_a \pi(a|s) \cdot \sum_{s'} p(s'|s,a) \cdot (r(s,a) + \gamma V^\pi(s'))$$

↳ system of equations

(*). Contraction Mapping thm :-

- f is a contraction :
- ⇓
- (f) has a fixed point : unique.

(x) : $f(x) = x$

$\|f(x) - f(y)\| \leq \alpha \|x - y\|$

$0 < \alpha < 1$

$x \longrightarrow y$

$\downarrow \qquad \downarrow$

$f(x) \longmapsto f(y)$

how to find fixed point :-

1. $x_0 \in \mathbb{R}$ ✓
2. Repeatedly apply f:

$x_0 \to f(x_0) \to f(f(x_0)) \to \cdots \to x^*$ s.t $f(x^*) = x^*$

$$f(V) = \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) \cdot (r(s,a) + \gamma V(s'))$$

$V = f(V)$ → ↳ fixed point of this function

(*). Show f is contraction. ✓ $\|f(V_1) - f(V_2)\| \leq \alpha \|V_1 - V_2\|$.

$0 < \alpha < 1$.

Iterative policy Evaluation :-

1. Initialising $V(s) \equiv 0$
2. Loop :- → 3 step

   Loop for each state :-

   $$V'(s) \leftarrow \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)(r(s,a) + \gamma V(s'))$$

3. $V^\pi(s) \leftarrow$ result

0.001

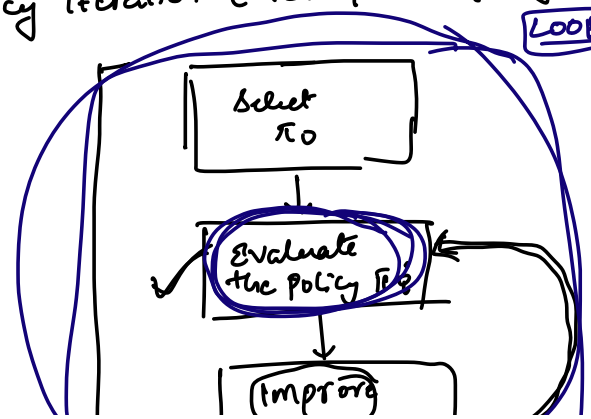
(*) $\{\frac{1}{n}\} \to 0$

$1, \frac{1}{2}, \frac{1}{3}, \ldots$

$\frac{1}{2} \quad \frac{1}{6}, \ldots$

Policy Iteration (for optimal policy) :-

Loop :-



S, A

1, 2, ?
A A A ..

deterministic ? A

$\pi : S \to A$.

π is the optimals

Policy improvement:-

$\pi'(s) \leftarrow \arg\max_a \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$.

(new policy)

$V^{\pi'} \geq V^{\pi}$ ✓ → improvement.

$v^*(s) = \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$.

(2 A). $V_0(s) = 0$, $\forall s$. ∀

π : $[1/4, 1/4, 1/4, 1/4]$.

Step 1 :-

$V_1(1) := \frac{1}{4}[-2 + \gamma \cancel{V_0(1)}^{0}] + \frac{1}{4}[-1 + \gamma \cancel{V_0(4)}^{0}] +$

top, down

$\frac{1}{4}[-1 + \gamma 0] + \frac{1}{4}[-1 + \gamma \cdot \cancel{V_0(2)}^{0}]$.

left, right

$= V_1(7)$.

$V_1(2) := \frac{1}{4}[-2 + \gamma V_0(2)] + \frac{1}{4}[-1 + \gamma V_0(5)] +$

top

$\frac{1}{4}[-1 + \gamma V_0(1)] +$

$\frac{1}{4}[-2 + \gamma V_0(2)]$.

$= V_1(6)$.

$V_1(4) := \frac{1}{4}[-1 + \gamma V_0(1)]$

$+ \frac{1}{4}[-1 + \gamma V_0(7)] + \frac{1}{4}[-1 + \gamma V_0(5)] +$

$\frac{1}{4}[-1 + \gamma V_0(5)]$.

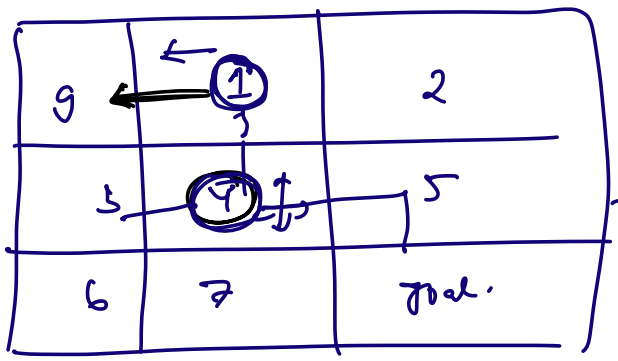
Step 2 :

Step 3 :-

(b). Policy improvement :-

$V_1(1) = V_1(7) = V_1(3) = V_1(5) = -1.25$.

$V_1(4) = -1.$; $V_1(2) = V_1(6) = -1.5$



$\gamma = 1$

$\pi(a|s)$; → action taken

$P(s', r|s, a)$ → transition

π_0 : equi prob.

(*) $\pi_1 :-$ $\underset{a}{\text{argmax}} \left\{ \sum_{s',r} P(r, s'|s, a) \left(r + \gamma V^{\pi_0}(s') \right) \right\}$.

(transition p.)

$\underset{\{U, D, L, R\}}{\text{argmax}} \left\{ \begin{array}{l} -2 + 1.(-1.25). \\ -1 + 1(-1), \\ -1 + 1(0), \\ -1 + 1(-1.5). \end{array} \right\}$ → best.

$\pi_1(4)$: $\underset{\{U, D, L, R\}}{\text{argmax}} \left\{ \begin{array}{l} -1 + 1(-1.25), \\ -1 + 1(-1.25), \\ -2 + 1(-1.25), \\ -1 + 1(-1.25) \end{array} \right\}$.

---

(3). Status :- { high, low }

actions :- { (wait, search), (recharge, search, wait) }.

$$v^*(s) = \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma v^*(s')].$$

$$v^*(\text{high}) = \max \{\text{wait}, \text{search}\} \begin{bmatrix} r_{\text{wait}} + \gamma v^*(\text{high}), \\ \alpha[r_{\text{search}} + \gamma v^*(\text{high})] + \\ (1-\alpha)[r_{\text{search}} + \gamma v^*(\text{low})]. \end{bmatrix}$$

$$v^*(\text{low}) = \max \{\text{recharge}, \text{wait}, \text{search}\} \begin{bmatrix} 0 + \gamma v^*(\text{high}), \\ r_{\text{wait}} + \gamma v^*(\text{low}), \\ \beta[r_{\text{search}} + \gamma v^*(\text{low})] + \\ (1-\beta)[-3 + \gamma v^*(\text{high})]. \end{bmatrix}$$

Bellman eq. for optimal policy.

(b). $\pi = (\text{search}, \text{recharge})$

v^π (Bellman eqn for given policy).

$$v^\pi(\text{high}) = -10 + 3 \cdot v^\pi(\text{low}).$$
$$v^\pi(\text{low}) = 5 + 4 \cdot v^\pi(\text{high})$$