1 Dynamic Programming

2nd Mid Term.
31th March.
11:30 - 1:00

Policy Evaluation

Solve value function for a given policy π

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

(Bellman Equation for Policy)

$$V^{\pi} = r^{\pi} + \gamma P^{\pi} V^{\pi}$$

$$r^{\pi} = (r^{\pi}(s), s \in S) \quad \text{s.t}$$

$$r^{\pi}(s) = \sum_{a \in A(s)} \pi(a|s) \sum_{s',r} p(s',r|s,a)\, r$$

$$P^{\pi} = \left[\left[P^{\pi}(s,s') \right]\right]_{s,s' \in S}$$

$$P^{\pi}(s,s') = \sum_{a \in A(s)} \pi(a|s) \sum_{r} p(s',r|s,a)$$

$$\sum_{s' \in S} P^{\pi}(s,s') = 1$$

$$P^{\pi}(s,s') \geq 0 \quad \forall\, s, s' \in S$$

$$(I - \gamma P^{\pi})\, V^{\pi} = r^{\pi}$$

$$\text{or} \quad V^{\pi} = (I - \gamma P^{\pi})^{-1} r^{\pi}$$

Iterative Policy Evaluation

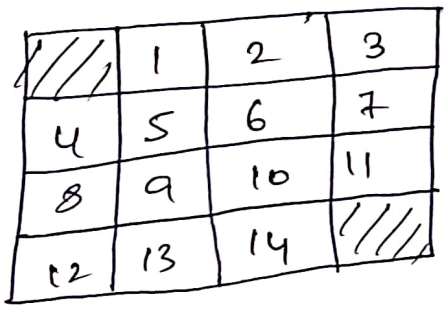
Is value Iterties for a given policy π

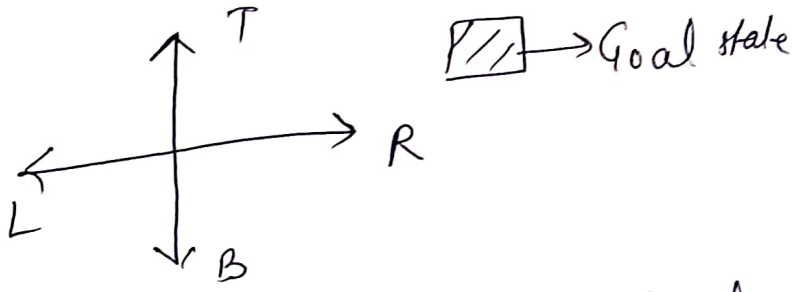
$$V_{n+1}(s) = \sum_{a \in A(s)} \pi(a|s) \sum_{s',r} p(s',r|s,a)(r + V_n(s'))$$

Starting for same V_0

$V_n \to V^\pi$ as $V \to V_n$.

Example Consider a 4×4 grid world.





Transition happens when you are in non shadded region

$R_t = -1$ on all transitions.

Non terminal states = 1, 2 . . . 14.

$P(6, -1 | 5, R) = 1$

↑ Right = action

Probability that u are in state 5 and go to state 6 with $R_t = -1$ and action = Right.

$P(7, -1 | 7, R) = 1$

$P(10, r | 5, R) = 0$ $\forall r \in \mathbb{R}$.

We apply iterative policy Evaluation for the random policy.

picks all directions w.prob 1/4.

$K=0$, $\gamma=1$

0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

v_K for random policy.

////	↕	+	+
+	+	+	+
+	+	+	+
+	+	+	////

greedy policy w.r.t v_K.

↕ becoz u can take any action and u will be at same state 0 w.r.t to v_K.

$K=1$, $\gamma<1 := 0.9$

0	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	-1
-1	-1	-1	0

	←	↕	+
↑	+	+	+
+	+	+	↓
+	+	→	

$\frac{1}{4}$ $\frac{1}{4}$ $\frac{1}{4}$ $\frac{1}{4}$ $\times 0$

$K=2$.

0	-1.75	-2	-2
-1.7	-2	-2	-2
-2	-2	-2	-1.7
-2	-2	-1.7	0

	←	←	+
↑	↑ ←	+	↓
↑	+	↓ →	↓
↕	→	→	

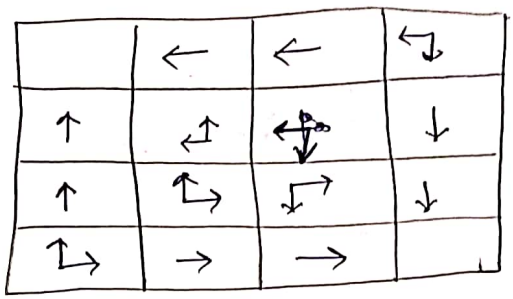
Reward.

$-1 + \frac{1}{4} \times 0 + \frac{3}{4} \times -1 = -1.75$.

↑ prob. of moving left ($\$ = 1$)

$K=3$

0	-2.4	-2.9	-3
-2.4	-2.85	-3	-2.9
-2.9	-3	-2.85	-2.9
-2.9	-2.9	-2.4	0



lowest cost path.

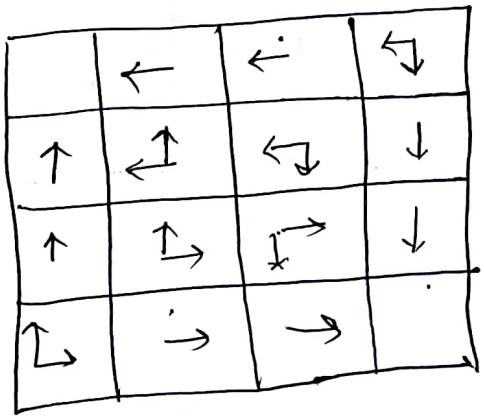
$$-1+\frac{1}{4}\times 0+\frac{1}{4}\times -1.7+\frac{2}{4}\times -2 \Rightarrow S=1$$

$$-1+\frac{1}{4}\times(-1.7)+\frac{3}{4}\times -2 \quad -S=2$$

$$-1+\frac{1}{4}\times 2\times(-1.7)+\frac{1}{4}(2\times -2) \quad S=5.$$

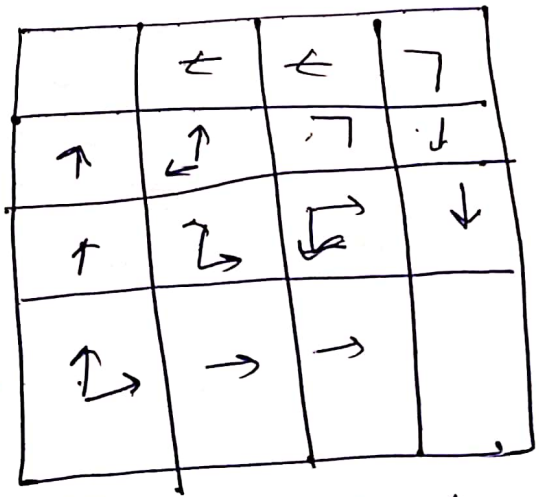
$K=10$

0	-6.1	-8.4	-9
-6.1	-7.7	-8.4	-8.4
-8.4	-8.4	-7.7	-6.1
-9	-8.4	-6.1	0



$K=\infty$

0	-14	-20	-22
-14	-18	-20	-20
-20	-20	-18	-14
-22	-20	-14	0



converged.

The greedy policy converge at $k=3$ only, but the optimal policy v_k converge at $k=\infty$.

Exercise

(1) If π is the equiprobable random policy, what is $q_\pi(11, \text{down})$? in $k=3$.

(11 ← state.)

q_π is optimal policy.

Soln $q_\pi(11, \text{down}) = -1$

(2) $q_\pi(7, \text{down})$? in $k=\infty$.

Soln $q_\pi(7, \text{down}) = -1 - 14 = -15.$

$v_\pi(11) = -14.$, $q_\pi(11, \text{down}) = -1$

As $\boxed{q_\pi(11, \text{down}) > v_\pi(11)}$

$q_\pi(7, \text{down}) = -15$

$v_\pi(7) = -20$

$\boxed{q_\pi(7, \text{down}) > v_\pi(7)}$

equiprobable Random Policy. can go anywhere

Hence q_π Policy is better than v_π Policy.

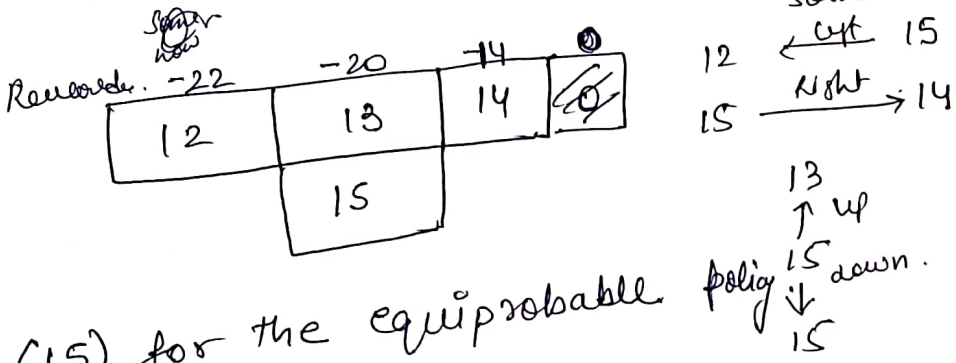
Likewise :-

$q_\pi(6, \text{left}) = -19$

$v_\pi(6) = -20$

$q_\pi(B, \text{right}) = -21$

Exercise We now add a new state 15 to the grid world just below state 13.



Find $V_\pi(15)$ for the equiprobable policy

every transition Reward $= -1$

Consider the gridworld for $\gamma = 0.9$.

$$V_\pi(15) = -1 - \frac{1}{4}[22 + 20 + 14] + \frac{1}{4} V_\pi(15)$$

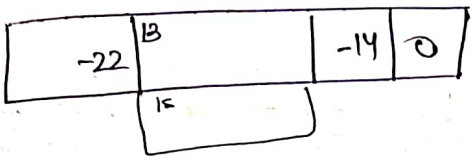
(Left, up, Right, down.)

$$\frac{3}{4} V_\pi(15) = -15$$

$$V_\pi(15) = -\frac{60}{3} = -20$$

Suppose now that dyanics of 13 is also changed such that action down from 13 takes agent to state 15. find $V_\pi(15)$?.

If $V_\pi(13)$ changes than its neighbouring values also changes and ultimately all values of the grid world changes. $\therefore$ $13 \to 15$ changes values but $15 \to 13$ do not.



Assuming neighbouring values of state 13 do not changes. Calculating $V_\pi(13)$

$$V_\pi(13) = -1 - \frac{1}{4}(-22 + 20 + 14) + \frac{1}{4} V_\pi(15)$$

$$= -1 - 14 + \frac{1}{4} V_\pi(15)$$

$$= -15 + \frac{1}{4} V_\pi(15)$$

$$V_\pi(15) = -1 - \frac{1}{4}[22 + 14] + \frac{1}{4} V_\pi(13) + \frac{1}{4} V_\pi(15)$$

$$= -10 + \frac{1}{4} V_\pi(13) + \frac{1}{4} V_\pi(15)$$

$$= -10 + \frac{1}{4}\left(-15 + \frac{1}{4} V_\pi(15)\right) + \frac{1}{4} V_\pi(15)$$

$$V_\pi(15)$$

$$= -10 - \frac{15}{4} + \frac{5}{16} \cdot V_\pi(15)$$

$$V_\pi(15) = -20$$

$$V_\pi(13) = -20$$

But. In general it is not true. ~~That~~ that $V_\pi(B) = V_\pi(15)$

Mid Term Solution

(1) (a) $P(T = +ve) = P(T = +ve \mid S)P(S) + P(+ve/\overline{NS}).P(\overline{NS})$

$= 0.10425.$

(b) $P(S/+ve) = \dfrac{P(+ve \mid S)P(S)}{}$

Q2 $V_n(i) = \frac{1}{n} \sum_{m=n}^{n-1} \left[r(i, a x_m^i) + \gamma V_m(x_m^i) \right]$

$V_{n+1}(i) = \frac{1}{n+1} \left[r(i, a, x_n^i) + \gamma V_n(x_n^i) + \sum_{m=0}^{n-1} r(i, a x_m') + \gamma V_m(x_m^i) \right]$

$= \frac{1}{n+1} \left[r(i, a, x_n^i) + \gamma V_n(x_n') \right] +$

$\frac{1}{n} \frac{n}{n+1} [\quad] + \frac{n}{n+1} \left[V_n(i) \right]$

Ans. $X_{n+1} = x_n + a(n) \left[h(x_n) + M_{n+1} \right].$

$\rightarrow = V_n(i) + \frac{1}{n+1} \left[r(i, a, x_n') + \gamma V_n(x_n') - V_n \right.$

$\left. \frac{\gamma}{\gamma_n} \right]$

$+ \gamma V_n(x_n^i) - V_n(i)]$

$+ E[r(i, a, x_n^i) + \gamma V_n(x_n^i) -$

$V_n(i) / \gamma_n$

③ (a) $V^*(i) = \max_{a \in A_i} Q(i,a)$

(b) $q^*(i,a) = \sum_j [r(i,a,j) + \gamma' V^*(j)]$

(c) $\pi^*(i) = \arg\max_{a \in A_i} q^*(i,a)$

(d) $\pi^*(i) = \arg\max_{a \in A_i} E[r(i,a,j) + \gamma V^*(j)]$

Q4

$V^*(\text{high}) = \max_{\{\text{wait, search}\}} \begin{bmatrix} 3 + \gamma V^*(\text{high}), \\ 0.4[5 + \gamma V^*(\text{low})] \end{bmatrix}$

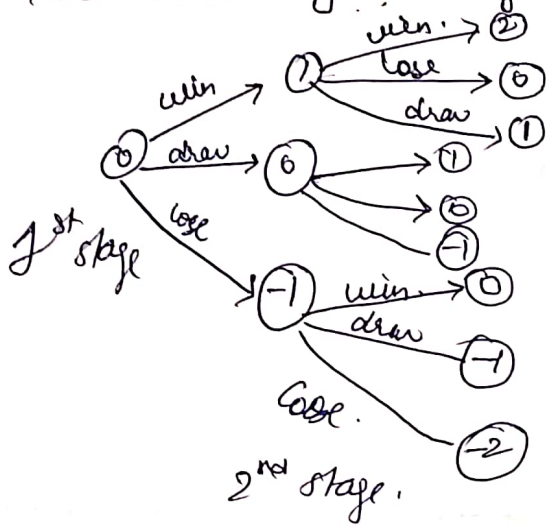
$V^*(\text{low}) = \max_{\{\text{search, recharge, wait}\}} \begin{bmatrix} \{0.2[5 + \gamma V^*(\text{low}), \\ \text{search} \quad 0.5[-3 + \gamma V^*(\text{high}), \\ 0 + \gamma V^*(\text{high}), \\ 3 + \gamma V^*(\text{low})\}, \end{bmatrix}$

0.8 ∴

Q5 State space = {difference in points between you & opponent.

Action state = {Timid, Bold}.

Terminal single stage = +1, 0, -1



Rewards are given at the end of 2 games.

$$J^*(1) = \max_{\{Bold,\ Timid\}} \begin{cases} 0.3[0+1] + 0.7[0+0], & \text{Bold} \\ 0.6[0+1] + 0.4[0+0] & \text{draw, loose.} \end{cases}$$

Optimal action is Timid.

$$J^*(0) = \max_{\{Bold,\ Timid\}} \begin{cases} 0.3[0+1] + 0.7[0-1], \\ 0.6[0+0) + 0.4[0-1] \end{cases}$$

Bold.

$$J^*(-1) = \max_{\{B,\ T\}} \begin{cases} 0.3[0+0] + 0.7[0-1], & 0.7 \\ 0.6[0-1] + 0.4[0-1] \end{cases}$$

-10.

$$J_K^*(0) = \max_{\{B,T\}} \Big[0.3\,[0 + J_2^*(1)] + 0.7\,[0 + J_2^*(-1)],$$
$$0.6\,[0 + J_2^*(0)] + 0.4\,[0 + J_2^*(-1)]$$
$$0.6\,[-0.(1) + 0.4(-0.7)$$
$$= -2 + -2.8.$$

Given any policy π, if we find an action a in state s, such that $q_\pi(s,a) \geq v_\pi(s)$, then it is advantageous to pick action a in state s.

Suppose we have a new policy π' such that

$$\pi'(i) = \pi(i) \quad \forall i \neq s$$
$$\pi'(s) = a$$
while $\pi(s) \neq a$

Q) Is $v_{\pi'}(s) \geq v_\pi(s)$? $\forall\ s \in S$.

$$v_\pi(s) \leq q_\pi(s,a)$$
$$= E\Big[R_{t+1} + \gamma\, v_\pi(S_{t+1}) \,\Big|\, S_t = s, A_t = \pi'(s)\Big]$$
$$= E\big[R_{t+1} + \gamma\, v_\pi(S_{t+1}) \,\big|\, S_t = s\big]$$
$$= E_{\pi'}\big[R_{t+1} + \gamma\, v_\pi(S_{t+1}) \,\big|\, S_t = s\big]$$
$$\leq E_{\pi'}\big[R_{t+1} + \gamma\, q_\pi(S_{t+1}, \pi'(S_{t+1})) \,\big|\, S_t = s\big]$$
$$= E_{\pi'}\Big[R_{t+1} + \gamma\, E_{\pi'}\big[R_{t+2} + \gamma\, v_\pi(S_{t+2}) \,\big|\, S_{t+1}, A_{t+1} = \pi'(S_{t+1})\big] \,\Big|\, S_t = s\Big]$$

$$= E_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 V_\pi(S_{t+2}) \,\middle|\, S_t = s\right]$$

$$\leq E_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 V_\pi(S_{t+3}) \,\middle|\, S_t = s\right]$$

$$\vdots$$

$$\leq E_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \,\middle|\, S_t = s\right]$$

$$= V'_\pi(s)$$

$$\pi'(s) = \underset{a}{\operatorname{argmax}}\, q_\pi(s,a)$$

$$\pi'(s) = \underset{a}{\operatorname{argmax}}\, E\left[R_{t+1} + \gamma V_\pi(S_{t+1}) \,\middle|\, \begin{matrix} S_t = s, \\ A_t = a \end{matrix}\right]$$

$$= \underset{a}{\operatorname{argmax}} \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma V_\pi(s')\right]$$

[Policy Improvement]

suppose π' [the greedy policy] is as good as (but not better than) the old policy π, Then

$$V_\pi(s) = V_{\pi'}(s) \quad \forall s \in S$$

Then

$$V_{\pi'}(s) = \max_{a \in A(s)} E\left[R_{t+1} + \gamma V_{\pi'}(S_{t+1}) \,\middle|\, \begin{matrix} S_t = s, \\ A_t = a \end{matrix}\right]$$

$$= \max_{a \in A(s)} \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma V_\pi(s')\right]$$

$$= \boxed{V_{\pi'} = V^* = v_\pi}$$

Consider a state s where there are $a_1, a_2, \ldots$ as the feasible actions.

Let a_i, a_j, a_k be the maximizing actions.

Consider a stochastic policy.

$$\pi'(s) = \begin{cases} a_i & \text{w.p. } p_1 \\ a_j & p_2 \\ a_k & p_3 \end{cases}$$

$$\text{s.t} \quad p_1 + p_2 + p_3 = 1 \qquad [\text{& } p_1, p_2, p_3 \geq 0]$$

Then $\pi'(s)$ gives the optimal actions in state s regardless of p_1, p_2, p_3.

Policy Iteration

Involves two nested loops

outer loop — policy Improvement

Inner loop — Policy evaluation. (can be done via say iterative PE)

$$\pi_0 \xrightarrow{PE} V_{\pi_0} \xrightarrow{PI} \pi_1 \xrightarrow{PE} V_{\pi_1} \xrightarrow{PI} \pi_2 \cdots \pi^*$$

$$\pi^* \xrightarrow{PE} V_{\pi^*} = V^*$$

V^* is the optimal value Policy.

π^* is the optimal policy.

Consider a state s where there are $a_1, a_2 \ldots a_n$ as the feasible actions.

Let a_i, a_j, a_k be the maximizing actions.

Consider a stochastic policy.

$$\pi'(s) = \begin{cases} a_i & \text{w.p. } p_1 \\ a_j & p_2 \\ a_k & p_3 \end{cases}$$

s.t $p_1 + p_2 + p_3 = 1$ [& $p_1, p_2, p_3 \geq 0$]

Then $\pi'(s)$ gives the optimal actions in state s regardless of p_1, p_2, p_3.

Policy Iteration

Involves two nested loops

outer loop policy Improvement

Inner loop Policy Evaluation. (can be done via say iterative PE)

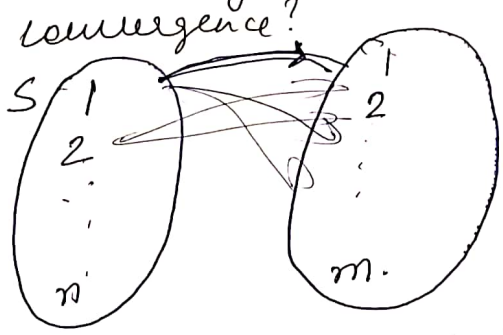
$$\pi_0 \xrightarrow{PE} V_{\pi_0} \xrightarrow{PI} \pi_1 \xrightarrow{PE} V_{\pi_1} \xrightarrow{PI} \pi_2 \cdots \pi^*$$

$$\pi^* \xrightarrow{PE} V_{\pi^*}^* = V^*$$

V^* is the optimal value Policy.

π^* is the optimal policy.

Q how many steps it will take for convergence?



no. of functions btw S & A $= \underbrace{m \times m \times \cdots m}_{n \text{ times}}$

$= m^n$

Hence No. of deterministic policies is also finite.

if S & A is finite

Policy Iteration algorithm converges in a finite no. of state [measured in terms of policy Improvement]

policy Iteration (using iterative PE) for estimating

$\pi \approx \pi^*$

1. Initialize :- $V(s) \in \mathbb{R}$ and $\pi(s) \in A(s)$ arbitarily for $s \in S$.

2. Policy Evaluation :-

Loop

$\Delta \leftarrow 0$

Loop for each $s \in S$

$v \leftarrow V(s)$

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

$$\Delta \leftarrow \max(\Delta, |v - V(s)|)$$

until $\Delta < \theta$ [sme small positive No.]

3. policy Improvement

For each $s \in S$,

old actions $\leftarrow \pi(s)$

$$\pi(s) \leftarrow \underset{a}{\text{argmax}} \sum_{s,r} p(s',r|s,a)[r+\gamma V'(s)]$$

If old actions $\neq \pi(s)$

Go back to step2.

If $V_\pi(s) = V_{old}(s)$ $\forall s$, then $V_\pi = V^*$ & π & old are both optimal.

policy Iteration

(1) Policy evaluation

Given Policy update π_k, evaluate $V_{\pi k}$ as

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

(2) Policy Improvement.

$$\pi_{k+1}(s) \underset{a \in A(s)}{\text{argmax}} \sum_{s',r} p(s',r|s,a)[r+\gamma V_{\pi k}(s')]$$

Or

$$T V_{\pi k} = T_{\pi_{k+1}} V_{\pi k}$$

$$\max_{a \in A(s)} \sum_{s', r} p(s', r | s, a) \left[r + \gamma V_{\pi_k}(s') \right]$$

$$= \sum_{s', r} p(s', r | s, \pi_{k+1}(s)) \left[r + \gamma V_{\pi_k}(s') \right]$$