16/3/19

Policy Iteration Algorithm

- Policy Improvement

Find μ_{k+1} s.t $T_{\mu_{k+1}} J_{\mu_k} = T J_{\mu_k}$

- Policy Evaluation (given Policy μ_k)

(Iterative PE can take a long time for convergence)

Value Iteration Algorithm

$$V_{k+1}(s) = \max E\left[R_{t+1} + \gamma v_k(s_{t+1}) \,\middle|\, s_t = s, A_t = a \right]$$

$$= \max_a \sum_{s', r} p(s', r \mid s, a)\left[r + \gamma v_k(s') \right] \quad \forall s \in S.$$

VI can be thought of as Spl. case of PI when the IPE step is ran exactly once before the improvement step is conducted.

Value Iteration for estimating $\pi = \pi^*$

Parameter : small threshold $\theta > 0$ determining the accuracy of estimation.

Loop :

$\Delta \leftarrow 0$

Loop for each $s \in S$

$v \leftarrow V(s)$

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

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

until $\Delta < \theta$

output : $V(s)$ $\forall s$ and a deterministic policy $\pi \approx \pi^*$

s.t $\pi(s) = \operatorname{argmax}_a \sum_{s',r} p(s', r | s, a)[r + \gamma V(s')]$

Modified Policy Iteration

Fix some integers $m_0, m_1, m_2 \cdots \geq 1$

Start with some initial estimate of policy μ_0

At k^{th} step : suppose current policy is μ_k.

Policy Evaluation

$V_{K+1} = T_{\mu_K}^{n\to\infty} V_K$

Policy Improvement (Gives μ_{K+1})

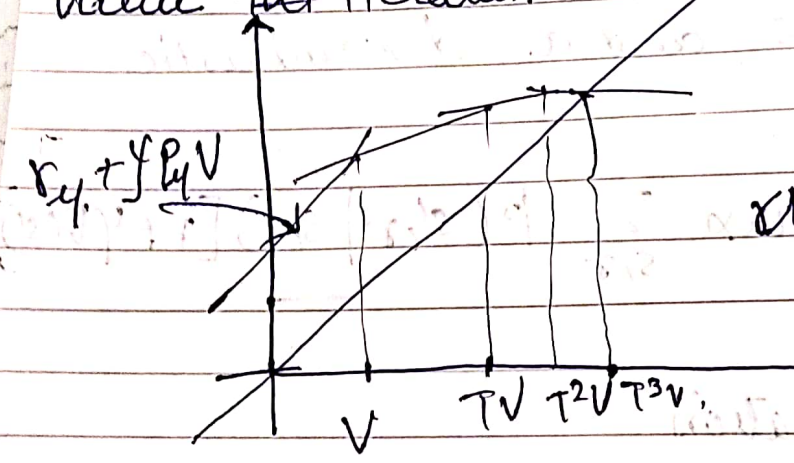
$T_{\mu_{K+1}} V_{K+1} = TV_{K+1}$

$$\boxed{V_{n+1} = TV_n}$$

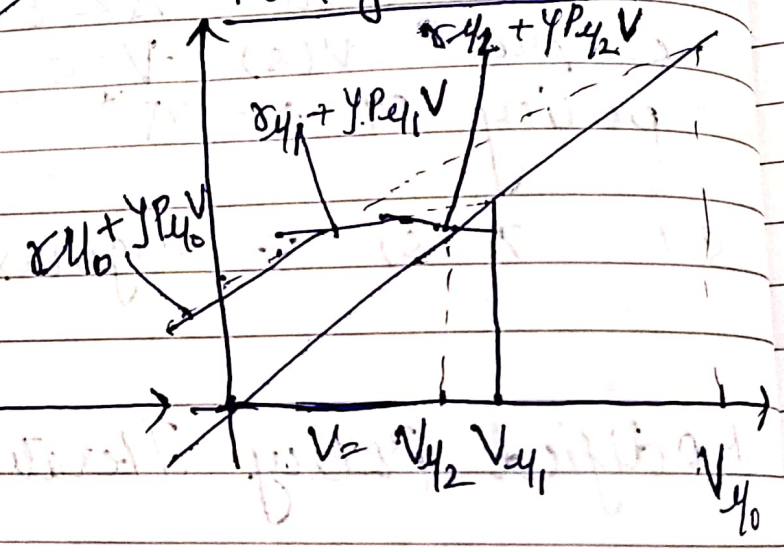
μ_t converges at $n \to \infty$

$V \to V^*$

Value Iteration



Policy Iteration



Geometric Approach PI:

Suppose $J \in \mathbb{R}^n$ is some ~~cost~~ reward function

A policy μ is greedy w.r.t J if

$$T_\mu J = TJ$$

Thus

$$\sum_j p_{ij}(\mu(i))\,(r(i,\mu(i),j) + \alpha J(j)) = \max_{\mu \in A(i)} \sum_{j \in S} p_{ij}(\mu)\,(r(i,u,j) + \alpha J(j))$$

Note :- There can be several greedy policies for a given J and likewise several J for which a policy is greedy.

let $R_\mu = \{J \mid \mu \text{ is greedy w.r.t } J\}$

Note $J \in R_\mu$ if and only if $T_\mu J = TJ$.

In other words,

$$\sum_{j \in S} p_{ij}(\mu(i))\,(r(i,\mu(i),j) + \alpha J(i)) \geq \sum_{j \in S} p_{ij}(\mu)\,(r(i,\mu,j) + \alpha J(i)) \quad \forall \mu \in A(i)$$

Q Is R_μ is a closed set?

Consider a sequence $\{J_n\}$ of functions s.t $J_n \in R_\mu \; \forall \; n \geq 1$. let $J_n \to J$ as $n \to \infty$.

$$J_n(x) \to J(x) \quad \forall \; x$$

since $J_n \in R_y \ \forall n$, we have

$$\sum_{j \in S} p_{ij}(y(i)) \left(r(i, y(i), j) + \alpha J_n(j) \right)$$

$$\geq \sum_{j \in S} p_{ij}(y) \left(r(i, y(i), j) + \alpha J_n(j) \right)$$

$\forall y \in A(i)$
$\forall n \geq 1$

limits on both sides as $n \to \infty$, we obtain.

$$\sum p_{ij}(y(i)) \left(r(i, u(i), j) + \alpha J(i) \right) \geq \sum_{j \in S} p_{ij}(y) \left(r(i, u, j) + \alpha J(j) \right)$$

Property

Let $f_n \geq g_n \ \forall n$.

Let $f_n \leq g_n$.

Let some point for the limit to exist

Suppose J_1 & J_2 are 2 functions in R_y

Let $\beta \in [0,1]$ & let

$$J = \beta J_1 + (1-\beta) J_2.$$

Q Is $J \in R_y$?

since $J_1 \in R_y$, we have.

$$\sum_{j \in S} p_{ij}(y(i)) \left(r(i, y(i), j) + \alpha J_1(j)\right) \geq \sum_{j \in S} p_{ij}(y) \left(r(i, y, j) + \alpha J_1(i)\right) \quad \text{— (1)}$$

since $J_2 \in R_y$, we have.

$$\sum_{j \in S} p_{ij}(y(i)) \left(r(i, y(i), j) + \alpha J_2(j)\right) \geq \sum_{j \in S} p_{ij}(y) \left(r(i, y, j) + \alpha J_2(j)\right) \quad \text{— (2)}$$

$\beta * (1) + (1-\beta) * (2)$ gives

$$\sum_{j \in S} p_{ij}(y(i)) \left[\beta r(i, y(i), j) + \alpha \beta J_1(j) + (1-\beta) r(i, y(i), j) + \alpha(1-\beta) J_2(j)\right]$$

$$\geq \sum_{j \in S} p_{ij}(y) \left[\beta r(i, y, j) + \alpha \beta J_1(j) + (1-\beta) r(i, y, j) + \alpha(1-\beta) J_2(j)\right]$$

or $\sum_{j \in S} p_{ij}(y(i)) \left[r(i, y(i), j) + \alpha J(j)\right]$

$$\geq \sum_{j \in S} p_{ij}(y) \left(r(i, y, j) + \alpha J(j)\right) \quad \forall\ y \in A(i)$$

$\Rightarrow J \in R_y$. Thus R_y is a convex set.

Thus, R_y is both closed and convex.

R_y is obtained their greedy policy y w.r.t J.

$J = (J(1), J(2), \ldots, J(n))$ $\quad J: R^n \to R$

Q Do R_y form a partition of R^n

$\rightarrow R_y = \{J \mid y \text{ is greedy w.r.t } J\}$

$\div$ we showed R_y is closed and convex

$- \quad J \in R_y \iff T_y J = TJ$

$-$ Suppose $J_y \in R_y$

$\Rightarrow T_y J_y = T J_y$

$J_y = T J_y \qquad J_y = T_y J_y$

or $J_y = J^*$

Example

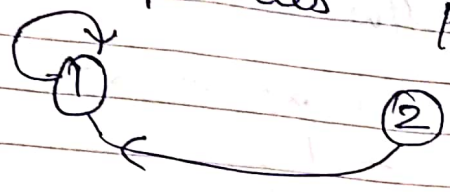
Consider a discounted reward problem with 2 states 1 & 2. Actions available in each state — (1)

1st action — Stay there at a reward of −1

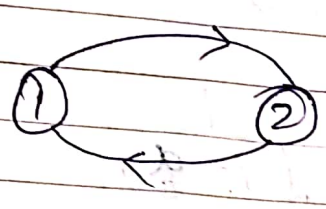
2nd action — Move to other state at a reward of 0.

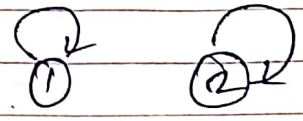
2^2 ← 2 states.

There are 4 policies possible

μ_l 

μ_r 

μ_g 

μ_w 

Consider an arbitary reward vector $J \in \mathbb{R}^4$
In state 1, a greedy policy can choose to stay if and only if

$$-1 + \alpha J(1) \geq 0 + \alpha J(2)$$

$$\text{or} \quad \alpha J(1) \geq \alpha J(2) + 1$$

$$\text{or} \quad J(1) \geq J(2) + 1/\alpha \longrightarrow (1)$$

Similarly in state 2 a greedy policy chooses to stay if and only if

$$-1 + \alpha J(2) \geq 0 + \alpha J(1)$$

$$\text{or} \quad J(2) \geq J(1) + 1/\alpha \longrightarrow (2)$$

We now identify sets R_{μ_L}, R_{μ_R}, R_{μ_g} and R_{μ_w} associated with four policies.

$V(s)$ for each V

Recurrent policy μ_w:

$$J^{\mu_w}(1) = -1 - \alpha - \alpha^2 - \alpha^3 \ldots$$

$$= \frac{-1}{1-\alpha}$$

$$J^{\mu_w}(2) = \frac{-1}{(1-\alpha)}$$

For μ_w to be greedy w.r.t J_1 we require both (1) & (2) to hold. So we require

$$J(1) \geq J(1) + \frac{2}{\alpha}$$

similarly

$$J(2) \geq J(2) + \frac{2}{\alpha}$$

This is impossible

$$\Rightarrow R_{e\mu_w} = \phi$$

Consider now the left policy μ_L

$$J^{\mu_L}(1) = \frac{-1}{(1-\alpha)}$$

$$J^{\mu_L}(2) = \frac{-\alpha}{(1-\alpha)}$$

μ_L is greedy w.r.t J if and only if

$$J(1) \geq J(2) + 1/\alpha.$$

$$0 + \alpha J(1) \geq -1 + \alpha J(2)$$

Thus
$$\begin{cases} J(1) \geq J(2) + 1/\alpha \\ J(1) \geq J(2) - 1/\alpha \end{cases}$$

$$\Rightarrow \boxed{J(1) \geq J(2) + 1/\alpha}$$

if true the μ is the optimal policy.

Consider J^{μ_L}

Is $J^{\mu_L} \in R_{\mu_L}$ if Yes then it is optimal.

i.e Is $\frac{-1}{1-\alpha} \geq \frac{-\alpha}{1-\alpha} + \frac{1}{\alpha}$.

$$\frac{-\alpha^2 + 1 - \alpha}{\alpha(1-\alpha)} = \frac{(1-\alpha)(-\alpha^2+1-\alpha)}{\alpha(1-\alpha)}$$

$$= \frac{(1-\alpha)(1+\alpha)}{1-\alpha}$$

$$\frac{-1}{1-\alpha} \neq 1+\alpha \quad \text{Not true.}$$

No optimal policy in this case.

$$\frac{-1}{1-\alpha} \geq \frac{-\alpha^2 + 1 - \alpha}{\alpha(1-\alpha)}$$

$$-\alpha(1-\alpha) \geq (1-\alpha)(-\alpha^2 + 1 - \alpha)$$

Or $-\alpha + \alpha^2 \geq -\alpha^2 + \alpha^3 + (1-\alpha) - \alpha + \alpha^2$

$$(1-\alpha)(1+\alpha) \leq \alpha(1-\alpha)(1+\alpha)$$

If $\alpha = 1$ then only it is true.

Consider μ_g

$J^{\mu_g}(1) = 0 \quad ; \quad J^{\mu_g}(2) = 0.$

for μ_g to be greedy w.r.t J

$J(1) \geq J(2) - 1/\alpha$

& $J(2) \geq J(1) - 1/\alpha$

$0 \geq -1/\alpha$
$0 \geq -1/2$
} Both true.

$\therefore$ μ_g is the optimal policy.

$J^{\mu_g} \in R_{\mu_g}$

$J^{\mu_g} = J^*$

$\mu_g = \mu^*$

Chapter 5 : Monte Carlo Methods

- We do not know system model.
- We don't know transition probabilities $P_{ij}(a)$
- We don't know reward function $r(i,a)$
- We have a generative or simmulation model of the system.

Input / State (i) → next state (j), reward (r)
action (a)

Goal : Learn value function from the sample rewards.

Monte Carlo Predictions

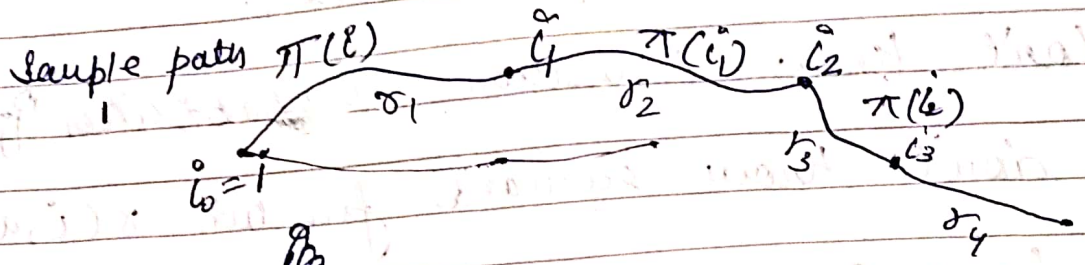
This is Estimating value of a given policy π

$$V^{\pi}(i) = E\left[\sum_{k=0}^{\infty} \gamma^k r(i_n, a_n, i_{n+1}) \,\middle|\, i_0 = i, a_n \sim \pi\right]$$

$$\boxed{0 < \gamma < 1}$$

The generative model is giving the new state j according to a distribution a.

$i_1 \sim P(\cdot | i, \pi(i))$



Rewards are Random.

$\hat{V}_1^{\pi}(i) = r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots + \gamma^N r_{N+1}$

for sample path 2 : Sample cost :-

$\hat{V}_2^{\pi}(i) = \cdots$

$\hat{V}_M^{\pi}(i) = \cdots$

$\frac{1}{M} \sum_{K=1}^{M} \hat{V}_K^{\pi}(i)$ — taking Avg. of all sample cost

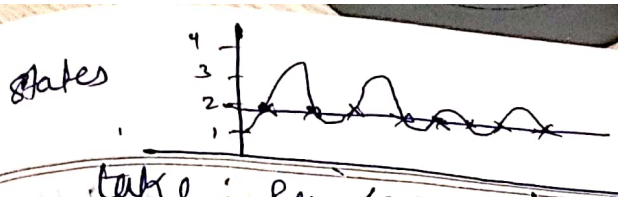
$\approx V^{\pi}(i)$

Two types of Methods in Monte Carlo

① First visit methods.

$i = i_0 \to i_1 \to i_2 \to i_3 \cdots$

i_1, i_2 can be same states



take single estimate from a trajectory.

(2) Every visits methods.

$i = i_1$ (multiple estimates from same trajectory)

First visit of MC Prediction for estimating $V \approx v_\pi$

input : a policy π to be evaluated.

Initialize:

$V(s) \in \mathbb{R}$, arbitrarily $\forall s \in S$

Returns (s) ← an empty list, $\forall s \in S$

Loop forever (for each episode):

Generate an episode following π:

$S_0, A_0, R_1, S_1, A_1, R_2, \ldots S_{T-1}, A_{T-1}, R_T$.

$G \leftarrow 0$

loop for each step of episode, $t = T-1, T-2, \ldots, 0$

$G \leftarrow \gamma G + R_{t+1}$

Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$

Append G to returns (S_t)

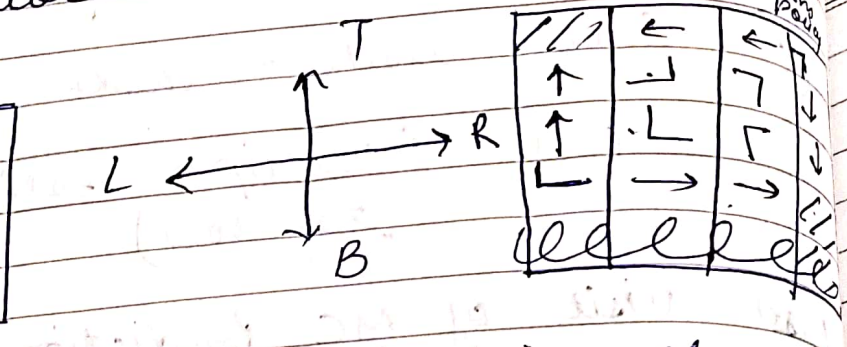
$V(S_t) \leftarrow$ average (Returns (S_t))

Ag #2

Consider Grid world example.

States

	1	2	3
4	5	6	7
8	9	10	11
12	13	14	



Recall the example from previous class where the random policy was evaluated.

Consider same reward structure as before.

Transition Probabilities

$p(0 \mid 1, \leftarrow) = 0.7$

$p(5 \mid 1, \leftarrow) = 0.1$

$p(2 \mid 1, \leftarrow) = 0.1$ " " & from 1

$p(1 \mid 1, \leftarrow) = 0.1$ going to 1 from 1

Transition Probabilities :

$p(1 \mid 5, \uparrow) = 0.4$

$p(4 \mid 5, \uparrow) = 0.4$

$p(9 \mid 5, \downarrow) = 0.1$

$p(6 \mid 5, \rightarrow) = 0.1$

Apply Monte-Carlo first visit method averaged over 20 iterations to estimate.

$$V_\pi(s) \quad \forall s \in \{1, \ldots, 14\}.$$

Pick start state randomly on each trajectory.