20/8/2018

Monte Carlo Methods:

— Prediction (first problem)

— Control (Second problem)

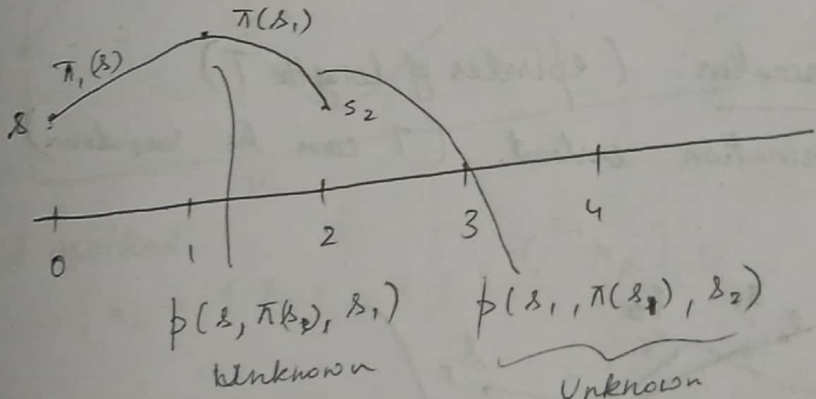
Idea:- Model of system is unknown. One has access to samples / transitions either through simulation / real data.

Prediction Problem: Two approaches

- first visit method
- Every visit method.

~~Rec~~ Recall, value f^n under a given policy π

$$V_\pi(s) = E\left[\sum_{k=0}^{\infty} \gamma^k r_{k+1} \mid S_0 = s\right]$$

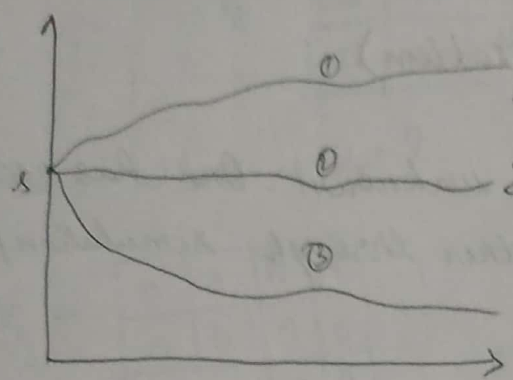


In order to find $V_\pi(s)$, we need to calculate or approx $E[\cdot]$ but we cannot calculate $E[\cdot]$ since we don't know $p(s, \pi(s), \cdot)$

$\pi(\cdot|s)$

[Most elementary procedure],

Run sample paths with intial state s.



$$V_\pi(s) \approx \frac{1}{M} \sum_{n=1}^{M} \sum_{k=0}^{N} \gamma^k r^n_{k+1}$$

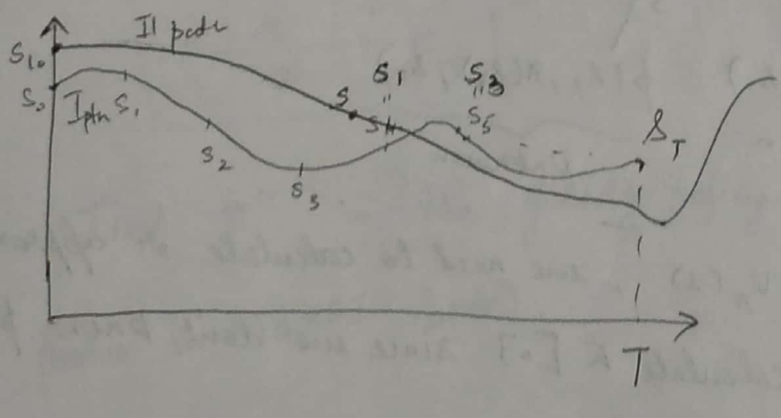
M is large
N is large

$$V_\pi(s) = E\left[\sum_{k=0}^{\infty} \gamma^k r_{k+1} \,|\, s_0\right] \text{ (approx)}$$

this method is inefficient since everytime intial state 's' is changed, the steps of M & N has to ~~carr~~ repeated agai

Problems with termination: (episodes of length T)

T is the termination instant. (T can be random)



First visit method:

- state s_0 is visited first time at time 0
- state s_1 is visited — — — 1
- state s_2 — — — — 2
- state s_3 3
- state s_1 is visited second time at time 4

I path:

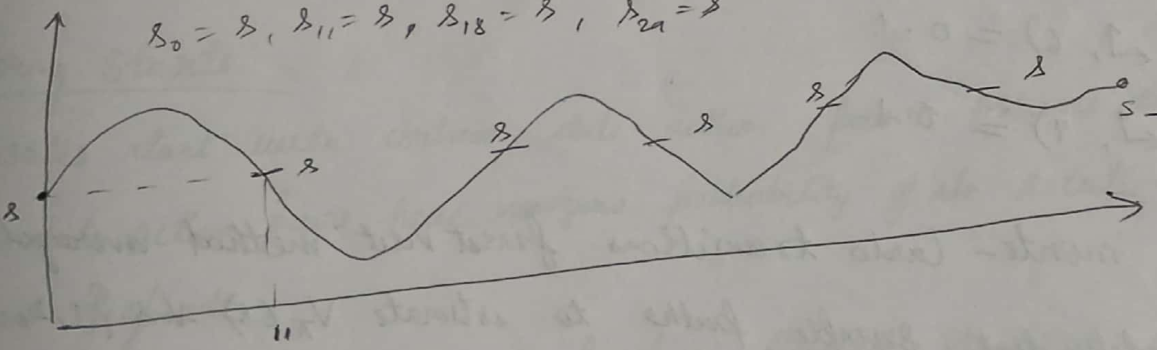
state s $r_1 + \gamma r_2 + \cdots \gamma^{T-1} \cdot r_T$

II path

state s $r_1^{(2)} + \gamma r_2^{(2)} \cdots \gamma^{T-1} r_T^{(2)}$

Every visit method:

$s_0 = s, s_{11} = s, s_{18} = s, s_{29} = s$



I visit method:

$[r_1 + \gamma r_2 + \cdots \gamma^{T-1} r_{k-}]$

Every visit method:

$r_1 + \gamma r_2 + \cdots \gamma r_T$ — X_1

$r_{11} + \gamma r_{13} + \cdots \gamma^{T-1} r_T$ — X_2

$r_{19} + \gamma r_{20} + \cdots \gamma^{T-1} \cdot r_T$ — X_3.

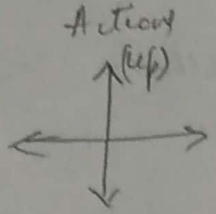
$$\frac{1}{N} \sum_{i=1}^{N} X_i$$

Exercise: Grid World Example

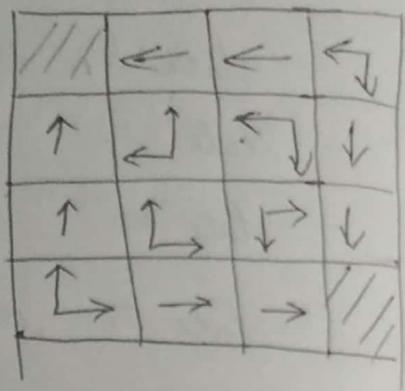
state

0	1	2	3
4	5	6	7
8	9	10	11
12	13	14	0

Actions



optimal policy



Transitions are stochastic, and are obtained as follows:

$p(1, \leftarrow, 0) = 0.7$ $\quad$ $p(1, \leftarrow, 2) = 0.1$ $\quad$ $p(i, a, j)$

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

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

$p(5, \leftarrow\uparrow, 4) = 0.4$ $\quad$ ← similar rule for all other actions.

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

$p(5, \leftarrow\uparrow, 6) = 0.1$

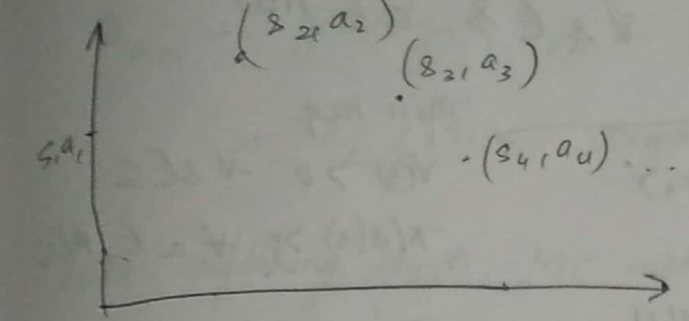
$p(5, \leftarrow\uparrow, 9) = 0.1$

Apply monte-Carlo transitions first visit method averaged over 28 independent sample paths to estimate $V_\pi(s)\ \forall s\ \{1, 2, \ldots 14\}$

Note:

- Monte carlo estimate can be used to estimate q-values (q_*) [one can obtain optimal policy here]

- Makes sense to look at randomized policies instead of ~~at~~ deterministic $q(s,a)$ is being estimated for various (s,a) tuples.



- Actions a_i in state s_i will be picked up from the given policy π
- If π is deterministic, then in state s_i, action picked $= \pi(s_i) = a_i$
- This is a disadvantage since we only learn about how good action $a_i \equiv \pi(s_i)$ is in state s_i but not about other actions feasible in state s_i
- Thus, we should consider randomized policies $\pi(a|s)$, $a \in A(s)$

  s.t. $\pi(a|s) = P(A_i = a | S_i = s) > 0 \; \forall a \in A(s)$.

Exploring Starts:

Episodes start with certain state-action pairs. Assume that all state action pairs have non-zero probability of ~~the~~ selection at the start of episode.

$\Rightarrow$ All state action pairs will be ~~limi~~ ~~limited~~ visisited a infinite number of times in the limit as number of episodes $\to \infty$

Suppose, $\nu(s) = p(s_0 = s)$, $s \in S$ be the initial distribution on states.

Then, $\mu(s, a) = p(s_0 = s, A_0 = a) = P(s_0 = s) P(A_0 = a | s_0 = s)$

$= \nu(s) \, \pi(a|s)$

We assume that,

$$\mu(s,a) = \underbrace{\nu(s)\,\pi(a|s) > 0}_{\text{min req}} \quad \forall s \in S,\ a \in A(s)$$

min req
$\nu(s) > 0 \ \ \forall s \in S$
$\pi(a|s) > 0 \ \ \forall a \in A(s)$

$$\sum_{s,a} \mu(s,a) = \sum_{s} \sum_{a} \nu(s)\,\pi(a|s)$$

$$= \sum_{\pi} \nu(s) \sum_{a} \pi(a|s)$$

$$= 1$$

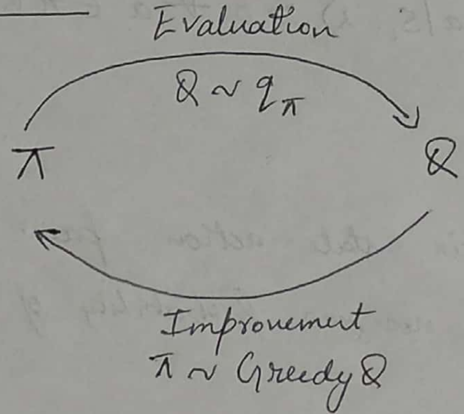
---

① ~~Avg~~ (Inverse R.L problem - construct optimal reward) (N.G)

→ Avg cost MDP - Q learning.

$\boxed{\text{raghub@iisc.ac.in}}$

Monte Carlo Control:



Given an intial policy π_0,

$$\pi_0 \xrightarrow{E} q_0 \xrightarrow{I} \pi_1 \xrightarrow{E} q_{\pi_1} \xrightarrow{I} \cdots \xrightarrow{I} \pi_* \xrightarrow{E} q_{\pi_*}$$

Goal: Find (π_*, q_{π_*})

$$(\pi, q_\pi) \xrightarrow[\text{Improvent}]{\text{policy Eval}} (\pi^*, q_{\pi^*})$$

$$\pi_{k+1} = \text{greedy}(q_{\pi_k})$$

$$= \pi_{k+1}(s) = \underset{a}{\text{argmax}}\ q_{\pi_k}(s,a) \quad \forall s \in S$$

with exploring starts, MC methods will compute q_{π_k} exactly for arbitrary π_k

π_{k+1} is a better policy than π_k because,

$$q_{\pi_k}(s, \pi_{k+1}(s)) = q_{\pi_k}(s, \underset{a}{\text{argmax}}\ q_{\pi_k}(s,a))$$

$$= \max_a q_{\pi_k}(s,a)$$

$$\geq q_{\pi_k}(s, \pi_k(s))$$

$$= v_{\pi_k}(s)$$

If $q_{\pi_k}(s, \pi_{k+1}(s)) = q_{\pi_k}(s, \pi_k(s))$

then both π_k & π_{k+1} are optimal polices

Else, $\exists\ s_0 \in S$ s.t

$$q_{\pi_k}(s_0, \pi_{k+1}(s_0)) > q_{\pi_k}(s_0, \pi_k(s_0))$$

Note that $q_{\pi_k}(s,a)$ are estimated using Monte-Carlo based policy evaluation. We dont need an infinite no of iterats for PE to converge for given policy.

Work around 1:

Stop when $\left| V_{\pi_{k+1}}(s) - V_{\pi_k}(s) \right| < \delta$

Work around 2:

Use a priori defined integer M_1, M_2, M_3 etc steps of iterate P.E which is followed by policy improvement

(Modified Policy Iteration)

This takes less no of steps of P.E before an improvement before an improvement step is conducted.

Monte Carlo Exploring Starts (ES) for Estimating $\pi \approx \pi_*$

Initialize:

$\pi(s) \in A(s) \quad \forall s \in S$

$Q(s,a) \in \mathbb{R}$ (arbitrary) $, \forall \; s \in S, a \in A(s)$

$Returns(s,a) \leftarrow$ Empty list $\forall \; s \in S, \; a \in A(s)$

(similar to Q-Value iteration)

Loop for each episode:

- Choose $S_0 \in S, A_0 \in A(S_0)$ randomly such that all pairs of states & actions have prob > 0
- Generate an episode from S_0, A_0 following π:
  $S_0, A_0, R_1, S_1, A_1, R_2 \ldots \ldots S_{T-1}, A_{T-1}, R_T$
- $G \leftarrow 0$

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

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

- ~~Ap~~ Append G to returns (S_t, A_t)

$Q(S_t, A_t) \leftarrow \text{Average}(\text{Returns}(S_t, A_t))$

$\pi(S_t) \leftarrow \underset{a}{\text{argmax}}\ Q(S_t, a)$

for each episode

$G \leftarrow 0$

For $t = T-1, T-2, \ldots 0$

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

$G_{T-1} = R_T$

$G_{T-2} = \gamma G_{T-1} + R_{T-1} = \gamma R_T + R_{T-1}$

$G = R_1 + \gamma R_2 + \gamma^2 R_3 + \ldots \gamma^{T-1} R_T$

$\bar{A}_n = \frac{1}{n} \sum_{i=1}^{n} G_i$ avg of G_t's over n episodes

$\bar{A}_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} G_i$

$= \frac{1}{n+1} \left(\sum_{i=1}^{n} G_i + G_{n+1} \right) = \frac{n}{n+1} A_n + \frac{1}{n+1} G_{n+1}$

$= \bar{A}_n + \frac{1}{n+1} (G_{n+1} - \bar{A}_n)$

Q) Can Monte-Carlo control procedure (with ES) converge to a suboptimal policy.

No [illegible], MC control with ES gives us optimal policy since otherwise corresponding value f^n will be suboptimal & policy improvement step happening on that value f^n will give a better policy $\Rightarrow$ convergence didn't happen.

(Inverted pendulum - MC with E-S)

A4 College Book

Kavitha

Monte Carlo "without" Exploring Starts:

If we don't use exploring starts, alternatively:

- All actions need to be selected infinitely often in each action.

R.L methods:

On-Policy Method — off policy methods.

On policy method:

Generate an episodes using policy π

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

Goal:

Estimate $v_\pi(S_0)$ [Value of state s_0 under policy π)

$$G_{T_1}^{(1)} = R_1^{(1)} + \gamma R_2^{(1)} + \ldots \gamma^{T-1} R_{T_1}^{(1)}$$

path

If $G_{T_1}^{(1)}$ - sample path, $G_{T_2}^{(2)}, \ldots G_{T_m}^{(m)}$ are estimates of $V_\pi(S_0)$ from M episodes, then

$$V_\pi(s_0) \approx \frac{1}{M} \sum_{i=1}^{M} G_{T_i}^{(i)}$$

T_i is termination for i^{th} episode.

Off policy Methods:

Generate episodes using policy π

$S_0, A_0, R_1 \ldots S_{T-1}, A_{T-1}, R_T$

Goal: Estimate $v_b(s_0)$ value of state s_0 under policy b

Note: → On-policy methods evaluate or improve policy used to make decisions.

→ Off-policy methods evaluate or improve policy different from policy used to generate data.