25/Aug/2019

Tabular Solution Method

Small state/action spaces.

Value function $V^*(s)$ f^n of a state $\forall s \in S$

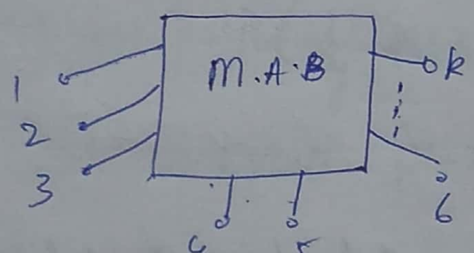
$Q^*(s,a)$ f^n of a state. $\forall s \in S, a \in A$

Multistage problem states:

$$V^*(s) = \max_{a \in A} Q^*(s,a)$$

We estimate $V^*(s)$ & $Q^*(s,a)$. These estimates are stored in arrays.

CHAPTER - 2 Multi-Arm bandits.



Action: pulling the arm.

Note: MAB has single state, but multiple action due to different rewards.

Let number of arms be 'k'

Actions available $\in \{1, 2 \ldots k\}$

Once arm is pulled, we get a reward. Given an action, i.e., follows that arm is pulled, the reward obtained follows a certain distribution. The distribution depends on the arm that is pulled.

Objective:-

Find the arm the gives highest total expected reward.

Let A_t = action selected at time t

r_t = reward at time t.

Let $q_*(a) = E[R_t \mid A_t = a]$

Let $Q_t(a)$ = estimated value of action a at time "t".
(Avg rewards)

$$Q_t(a) = \frac{\sum_{i=1}^{t} R_i \, I\{A_i = a\}}{\sum_{i=1}^{t} I\{A_i = a\}}$$

← total reward obtained by playing action 'a' till 't'

← num of times out of t, a was played.

Where $I_{\{A_i=a\}} = \begin{cases} 1 & \text{if } A_i = a \\ 0 & \text{otherwise} \end{cases}$

$Q_t(a)$ = Avg reward by playing 'a' till time t

A4 College Book

Kavitha

Scanned by CamScanner

We expect that

$$Q_t(a) \longrightarrow q_*(a) \quad \text{as time} \longrightarrow \infty$$

Example 1:

Suppose we have 5 arm ~~bandit~~ bandit m/c with action

$$A \in \{a, b, c, d, e\}$$

Actions	R_1	R_2	R_3	R_4	R_5
a	5	6	3	2	4
b	8	4	9	2	6
c	10	12	3	1	1
d	2	3	4	8	15
e	4	3	4	2	6

Q_1	Q_2	Q_3	Q_4	Q_5
5	5.5	~5	.4	4
8	6	7	>5	<5
$\boxed{10}$	$\boxed{11}$	$\boxed{\sim 8}$	$\boxed{6.5}$	5.4
2	2.5	3	>4	$\boxed{6.4}$
4	3.5	<3	>3	<4

$$Q_t = \frac{\sum R_i \, I_{A_i = a}}{\sum I_{A_i = a}}$$

for a

$$Q_2(a) = \frac{R_1 \, I_{A_1 = a} + R_2 \, I_{A_2 = a}}{1+1}$$

$$= \frac{5 \times 1 + 6 \times 1}{2}$$

$$= 5.5$$

Offline Scheme: The above table represent offline scheme. That is

5 independent copies of simulation are running (starting with same intial seed)

I simulation — arm 'a' is pulled always.

II simulation — 'b' is pulled

III simulation — 'c' is pulled.

In an online scheme, we need to decide that each time which arm to pull.

Online scheme:

Actions	R_1	R_2	R_3	R_4	R_5
a					
b		8			
c	10			12	
d			2		
e					

$$Q_t(a) = \frac{\sum_{i=1}^{t} R_i \, I_{\{A_i = a\}}}{\sum_{i=1}^{t} I_{\{A_i = a\}}}$$

Exploitation based scheme [Greedy approach]

$$a_t^* = \arg\max_a Q_t(a)$$

Exploration based scheme: [ϵ-Greedy approach] ← one eg, there are many.

$$a_t^* = \begin{cases} \arg\max_a Q_t(a) & \text{with prob } (1-\epsilon) \\ \text{Random action } \cancel{\text{than arg}} \text{ other than argmax} & \text{with prob } \epsilon \end{cases}$$

A4 College Book

Kavitha

Scanned by CamScanner

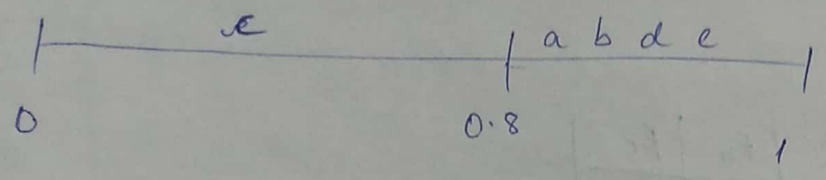
suppose $\epsilon = 0.2$

suppose $\arg\max_a Q_t(a) = c$

<u>ϵ-greedy scheme</u>

Pick action c with prob 0.8

Pick action a, b, d, e with prop 0.2



Eg 2: Ten arm, thus $k = 10$. Consider 2000 randomly generated MAB. Total no of time steps in each each ~~question~~ problem is 1000. i.e, $t = 1, 2, \ldots 1000$.

Select: $q_*(a)$ according to $N(0,1)$ $\quad N = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad x \in \mathbb{R}$

gaussian.

Here each sequence is played by 1000 time instances.

(Box Mueller method for $N(0,1)$ generator).

When action $A_t = a$ is picked at time t,

reward R_t is picked according to $N(q_*(a), 1)$

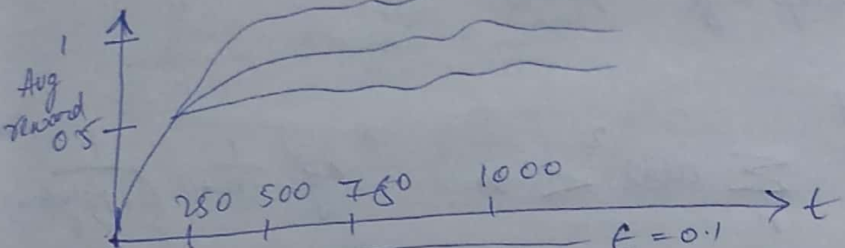
Since w.k.t, $E[Y] = E[X] + b$

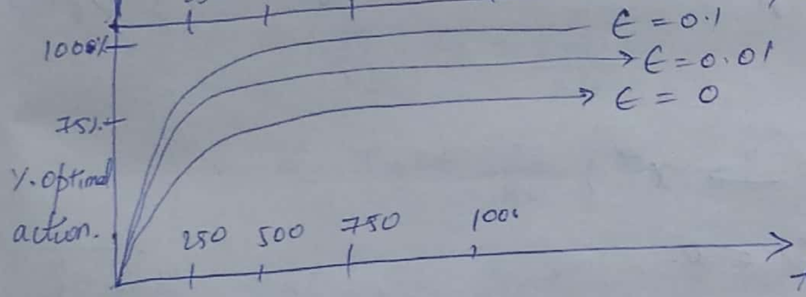
$$N(q_*(a), 1) = N(0, 1) + q_*(a)$$

Scanned by CamScanner

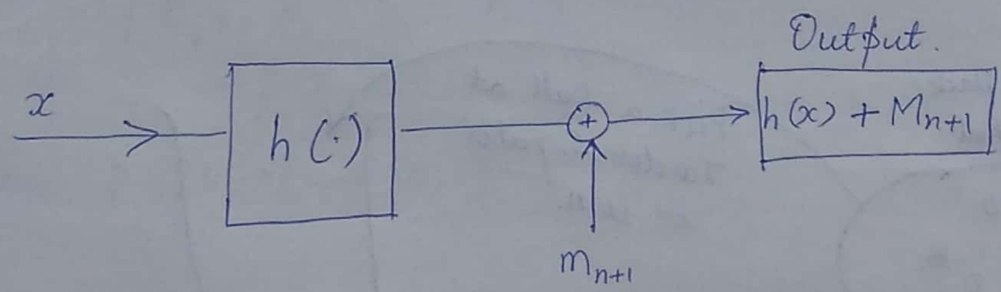
Plot $Q(a) = \dfrac{\sum R_i \, \mathbb{I}_{\{A_i = a\}}}{\sum \mathbb{I}_{\{A_i = a\}}}$

as a function of t. (With avg over 2000 independent runs)





Stochastic Approximation Algorithms



If we input x, we observe $h(x) + M_{n+1}$ (Where M_{n+1} is zero mean noise)

Goal: Find x^* such that $h(x^*) = 0$, without trying to estimate function h.

A4 College Book

Kavitha

Scanned by CamScanner

Algorithm: Start with some x_0 as an estimate ~~of~~ x for which $h(x) = 0$.

(Robbins & Monro 1951)

$$x_{n+1} = x_n + a_n \left(h(x_n) + M_{n+1} \right)$$

Here, $a(n)$, $n \geq 0$ is a step size sequence.

$a(n)$, $n \geq 0$ must satisfy.

$$\sum_n a(n) = \infty \qquad \sum_n a(n)^2 < \infty \qquad \text{eg } \frac{1}{(n+1)}, \ \frac{1}{(1+n)\log(1+n)}$$

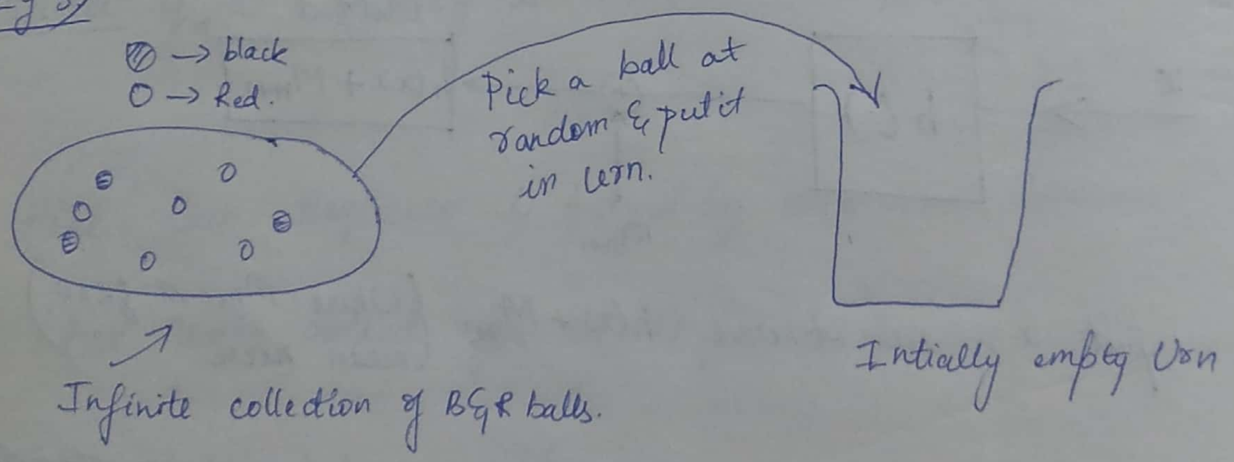
Result. As $n \to \infty$, $x_n \to x^*$,

where $h(x^*) = 0$.

$$E\left[\|x_n - x^*\|^2 \right] \longrightarrow 0 \text{ as } n \to \infty \qquad \text{(Convergence in mean sq sense)}$$

Modern day result $\quad x_n \longrightarrow x^*$ with prob 1

Eg 3)



Suppose: y_n = No of Red balls in Urn by time n.

$$y_{n+1} = y_n + \mathcal{E}_{n+1} \qquad \text{—— (1)}$$

Where $= \begin{cases} 1 & \text{if red ball is picked at } n+1 \\ 0 & \text{otherwise.} \end{cases}$

Scanned by CamScanner

Define: $x_n = \frac{y_n}{n}$

i.e. x_n is the fraction of red balls in urn at n.

from ①

$$\frac{y_{n+1}}{n+1} = \frac{y_n}{n+1} + \frac{\xi_{n+1}}{n+1}$$

$$= \frac{n}{n+1} \cdot \left(\frac{y_n}{n}\right) + \frac{\xi_{n+1}}{n+1}$$

$$x_{n+1} = \frac{n}{n+1} \cdot x_n + \frac{\xi_{n+1}}{n+1}$$

$$= x_n + \frac{1}{n+1}\left(\xi_{n+1} - x_n\right) \quad \text{—— ②}$$

Suppose that av fraction of redballs in urn at time n given entire history of red & black balls that were picked. is the past depends only on x_n

$$P(\xi_{n+1} = 1 \mid \xi_0, \xi_1, \xi_3 \ldots \xi_n) = p(x_n)$$

Where $p : [0,1] \to [0,1]$

reWrite ②

$$x_{n+1} = x_n + \frac{1}{n+1}\left(p(x_n) - x_n\right) + \frac{1}{n+1}\left(\xi_{n+1} - p(x_n)\right) \quad \text{—③}$$

let ~~M_{n+1}~~ $M_{n+1} \triangleq \left(\xi_{n+1} - p(x_n)\right) \Rightarrow E[M_{n+1}] = 0$

This is similar to Robbins-Manro algo

$$x_{n+1} = x_n + a_n\left(h(x_n) + M_{n+1}\right)$$

Where $h(x_n) = p(x_n) - x_n$

$a_n = \frac{1}{n+1}$ $\quad M_{n+1} = \xi_{n+1} - p(x_n)$

A4 College Book Kavitha

Scanned by CamScanner

We can consider ③ to be imitating an ODE,

$$\dot{x}(t) = p(x(t)) - x(t)$$

eg: Suppose the ODE

$$\dot{x}(t) = h(x(t)), \quad x(0) = x_0$$

$$x(t) = x(0) + \int_0^t h(x(z))\,dz$$

Suppose t is small, let $t \approx \alpha$

$$x(\alpha) = x(0) + \int_0^\alpha h(x(z))\,dz$$

$$\approx x(0) + \alpha h(x(\alpha))$$ assuming h doesn't vary much.

$$x((n+1)\alpha) \simeq x(n\alpha) + \alpha h(x(n\alpha)) \qquad n \geq 0$$

Eulers discretization.

Suppose ODE is

$$\dot{x}(t) = p(x(t)) - x(t)$$

Then Eulers discretization,

$$x((n+1)\alpha) \simeq x(n\alpha) + \alpha\,(p(x(n\alpha)) - x(n\alpha))$$

suppose $\hat{x}_n = x(n\alpha)$

Then $\hat{x}_{n+1} = \hat{x}_n + \alpha\,(p(\hat{x}_n) - \hat{x}_n)$

Scanned by CamScanner

comparing with (3)

$$x_{n+1} = x_n + \frac{1}{n+1}(p(x_n) - x_n) + \underbrace{\frac{1}{n+1}(\xi_{n+1} - p(x_n))}_{\text{suppose noise is dropped.}}$$

It turns out that,

$$\sum_n \frac{1}{n+1}(\xi_{n+1} - p(x_n)) < \infty$$

This converges!

$$Z_n = \sum_{m=0}^{n-1} \left(\frac{1}{m+1}\right)\left(\xi_{m+1} - p(x_m)\right) ; n \geq 1$$ is a martingale sequence.

If $E\left[\sum_n \left[\|Z_{n+1} - Z_n\|^2\right]\right] < \infty$,

then, $\{Z_n\}$ converges.

(Quadratic Variation process)

$$(Z_{n+1} - Z_n)^2$$
$$= \left(\frac{1}{n+1}\right)^2 \cdot \left(\xi_{n+1} - p(x_n)\right)^2$$
$$= \sum_n \left(\frac{1}{n+1}\right)^2 \left(\xi_{n+} - p(x_n)\right)^2$$
$$\leq \frac{1}{n} \cdot \sum_n \frac{1}{(n+1)^2} < \infty$$

$\therefore$ ~~Converges~~ converges.

$$x_{n+1} = x_n + \frac{1}{n+1}(p(x_n) - x_n) + \frac{1}{n+1}(\xi_{n+1} - p(x_n))$$
$$= x_0 + \sum_{m=0}^{n} \left(\frac{1}{m+1}\right)\left(p(x_m) - x_m\right) + \underbrace{\sum_{m=0}^{n} \left(\frac{1}{m+1}\right)\left(\xi_{m+1} - p(x_m)\right)}_{\lim_{n\to\infty} \sum_{m=0}^{n} \frac{1}{m+1}(\xi_{m+1} - p(x_m)) < \infty}$$

Thus, the algorithm can be considered to be a noisy Euler discretization of the ODE

$$\dot{x}(t) = p(x(t)) - x(t)$$ with non uniform step size

$$x(0) \in [0,1]$$

A4 College Book Kavitha

Scanned by CamScanner

to the bounds of $\dot{x}(t)$

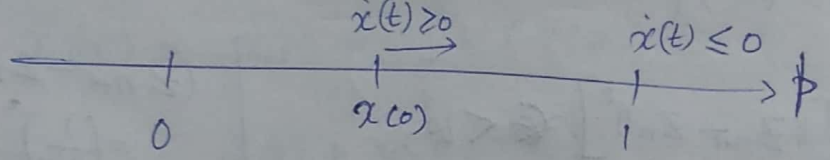
(I) Suppose $x(t) = 0 \Rightarrow P(x(t)) \geq 0$

$\therefore \quad p(x(t)) - x(t) \geq 0$

Case II $\quad x(t) = 1 \Rightarrow P(x(t)) \leq 1$

$p(x(t)) - x(t) \leq 0$

Suppose, p is (Lipschitz Continious)



$\ddot{x}(t) > 0 \qquad$ At some point in $[x(0), 1]$, $\dot{x}(t) = 0$

(Equillibrium pt of the ODE.)

Suppose

$$H = \{ x \mid p(x) = x \}$$

Then,

$$x(t) \longrightarrow H \text{ as } t \longrightarrow \infty$$

Then the original recursion with noise $\{x_n\}$ satisfies

$$x_n \longrightarrow H \text{ as } n \longrightarrow \infty \quad (\text{With prob} = 1)$$

Scanned by CamScanner

(This is not related to prev page derivation)

Estimate $q_*(a)$

Incremental Implementation

$\cancel{\frac{Q_n + Q_n - Q_n}{n+1 \qquad Q_{n+1}}}$

Let $A_t = a \quad \forall t$

Let R_i = reward received after i^{th} selection action a

Then $Q_n = \left(\dfrac{R_1 + R_2 + R_2 \dots + R_n}{n}\right)$

$$Q_{n+1} = \left(\frac{1}{n+1}\right) \sum_{i=1}^{n+1} R_i$$

$$\frac{(R_1 + R_n)}{n+1} + \frac{R_{n+1}}{n+1}$$

$$\frac{n}{n+1} \cdot \left(\frac{R_1 \dots R_n}{n}\right) + \frac{R_{n+1}}{n+1}$$

$$\left(\frac{n\, Q_n}{n+1}\right) + \frac{R_{n+1}}{}$$

$$= Q_n + \frac{1}{n+1}\left(R_{n+1} - Q_n\right) \quad \text{[Incremental Implementation]}$$

Need to store (Q_n, R_{n+1}) at any time.

this is space efficient.!

$Q_n \;,\; R_n$

$Q_{n+1} \quad R_{n+1}$

$Q_{n+2} \quad R_{n+2}$

Exercise: Simulate 15-arm bandit problem.

set $Q_0(a) = 5 \;\; \forall a$ consider 2 cases

(a) rewards for the 15 arms →
5, 6, 7, ... 19

(b) rewards for 15 arms
$N(5,1), N(6,1) \dots N(19,1)$

A4 College Book

Kavitha

Scanned by CamScanner

Incorporate Greedy exploration with different values of ϵ

$\epsilon = 0, 0.01, 0.1, 0.2$

plot for 5000 time steps.

II Multiple copies of this simulation 100copies & then averages.

Consider the Increment algo again,

$$Q_{n+1} = Q_n + \frac{1}{n+1}(R_{n+1} - Q_n)$$

$$= Q_n + \frac{1}{n+1}\left(E[R_{n+1}] - Q_n\right) + \frac{1}{n+1}\left(R_n - E[R_{n+1}]\right)$$

$$M_{n+1} = \left(R_{n+1} - E[R_{n+1}]\right)$$

$$E[M_{n+1}] = 0.$$

Now the question is

$$\sum_{m=0}^{n} \left(\frac{1}{m+1}\right)\left(R_{m+1} - E[R_{m+1}]\right) < \infty \; ?$$

Martingale

suppose $|R_m| < \infty$

Then $\sum_{m=0}^{n} \left(\frac{1}{m+1}\left(R_{m+1} - E[R_{m+1}]\right)\right) < \infty$

$$\dot{Q}(t) = E[R_t] - Q(t)$$

$$\{Q^* \mid Q^* = E[R]\}$$

$$\frac{Q_{n+1} - Q_n}{(1/n+1)} \approx [R_{n+1} - Q_n] \quad (E[R] - Q_n)$$

$Q_n \to E[R]$ with prob $=1$

Scanned by CamScanner