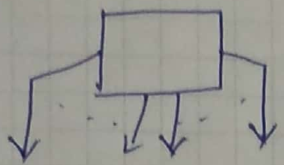
1|sept|2018

Multi-Arm Bandit



Actions, arms that we pull.

If we pull arm a, we get reward $r(a)$

$$r(a) = \begin{cases} r_1^a & w.p\ p_1^a \\ r_2^a & w.p\ p_2^a \\ \vdots & \\ r_m^a & w.p\ p_m^a \end{cases}$$

$$\sum_{i=1}^{m} p_i^a = 1$$

If action is A_t at time t, then reward is R_t

$$\boxed{Q_{n+1} = Q_n + \frac{1}{n+1}\left(R_{n+1} - Q_n\right)} \quad \text{--- } (1)$$

~~Stion~~ Stationary Setting: prob are time invariant.

Non-Stationary Setting: prob p_i^a, $i = 1, \cdots m$ are time dependent. They are $p_i^a = p_i^a(t)$

$$\hat{Q} = E[R]$$

In the non-stationary setting, we shall replace (1) by following setting

$$Q_{n+1} = Q_n + \alpha\left(R_{n+1} - Q_n\right) \quad \text{--- } (2)$$

Where $0 < \alpha < 1$ but constant.

(The effect of 'α' is more ~~from~~ effective for non-stationary conditions.)

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

but $Q_n = \alpha R_n + (1-\alpha) Q_{n-1}$

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

$$= \alpha R_{n+1} + \alpha(1-\alpha) R_n + (1-\alpha)^2 (Q_{n-1}) \ldots$$

w.r.t, $Q_{n-1} = (1-\alpha) Q_{n-2} + \alpha R_{n-1}$

$$= (1-\alpha)^n Q_1 + \alpha \sum_{i=1}^{n} (1-\alpha)^{n-i} (R_{i+1})$$

consider the weights $(1-\alpha)^n$ & $\alpha \sum_{i=1}^{n} (1-\alpha)^{n-i}$

$$(1-\alpha)^n + \alpha \left((1-\alpha)^{n-1} + (1-\alpha)^{n-2} \ldots + (1-\alpha) + 1 \right)$$

$$(1-\alpha)^n + \cancel{\alpha} \left(\frac{1-(1-\alpha)^{n-1}}{1-\cancel{(1-\alpha)}} \right)$$

$$= \boxed{1}$$

Thus (3) can be seen as weighted average of $Q_1, R_2, \ldots R_{n+1}$ where we assign $(1-\alpha)^n$ to Q_1 & $\alpha(1-\alpha)^{n-i}$ to R_{i+1}

Note: the past weights are lesser compared to latest ones. Thus these are called "faded memory systems"

As $n \to \infty \Rightarrow (1-\alpha)^n \longrightarrow 0$

Thus the effect of Q_1 vanishes asymptotically.

W P With probability.

Stability of recursion

$$Q_{n+1} = Q_n + \alpha(R_{n+1} - Q_n) \quad — ①$$

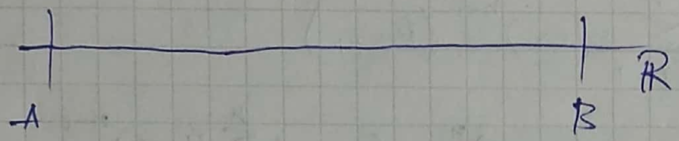
even $|Q_n|$ works

We need to ensure that $\sup_n \|Q_n\| < \infty$ w.p. 1

"sup" – Supremum – lowest upper bound.

Replace ① with

$$Q_{n+1} = \Pi\left(Q_n + \alpha(R_{n+1} - Q_n)\right)$$



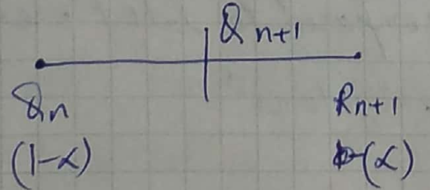
If $Q_{n+1} < A$, then set $Q_{n+1} = A$

If $Q_{n+1} > B$, then set $Q_{n+1} = B$

otherwise: $Q_{n+1} = Q_{n+1}$

And Π : projector operation.

Note: Q_{n+1} is convex combination of Q_n & R_{n+1}



$\therefore$ Q_{n+1} lies b/w Q_n & R_{n+1}

Supremum

$\{x_n\} \in R$

$\sup_n x_n$: least upper bound on $\{x_n\}$

$|x_n| \leq B \ \forall n$

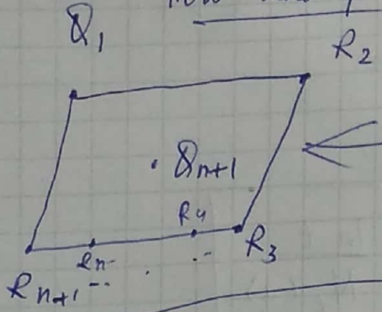
$|x_n| \leq A \ \forall n$

where $A < B$

then A is the tighter bound for x_n

for $(1-\alpha)^n Q_1 + \alpha \sum_{i=1}^{n} (1-\alpha)^{n-i} R_{i+1}$

How the plot looks:



Since the weighted avg sum to 1,

When new rewards after next iteration, Q_{n+1} has similar bound.

Thus stable.

$$\boxed{\begin{aligned} &\sup |R_n| < \infty \\ \Rightarrow &\sup |Q_n| < \infty \end{aligned}}$$

Q_n is weighted by $(1-\alpha)^n$

In the diminishing step size, i.e., $\alpha = \left(\frac{1}{n+1}\right)$, $n \geq 1$

Exercise 2 Simulate:-

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

has less oscillation in the end. Variance is low.

$$Q_{n+1} = Q_n + \alpha (R_{n+1} - Q_n) \quad \alpha = 0.01, 0.02, 0.001, 0.1$$

has oscillations. but the variance is high.

liner vs exp convergence ?

Action Selection methods.

→ ε - Greedy approach.

→ UCB - upper Confidence Bound:

Select actions according to

$$A_t \triangleq \underset{a}{\operatorname{argmax}} \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$

t = time instance

$N_t(a)$ = number of times action a is selected upto time 't'

If $N_t(a) = 0$, then a becomes maximizing action at time 't'

ie, intutively pick actions not ~~sele~~ selected in the past

$c\sqrt{\frac{\ln t}{N_t(a)}}$ ← correction / exploration term.

One, all actions have been explored sufficient no of times, from then $Q_t(a)$ starts to dominate.

$\boxed{\text{Ex 3:}}$ Repeat the ex 1 with UCB update. with $c = 0.5$, $c = 1$, ~~c = 0.5~~ $c = 2, 5$

~~Grad~~ Gradient Bandit Algorithms

Let $H_t(a)$ = a numerical preference for action a at time t

If large $H_t(a) \Rightarrow$ action a has a higher chance being picked.

Prob of picking action a at time t,

$$P(A_t = a) = \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}} \triangleq \pi_t(a)$$

This distribution is called Boltzman or Gibbs distribution

Initially set values of $H_0(a) = 0 \ \forall a = 1, \ldots k$

then $\pi_0(a) = \frac{1}{k} \ \forall a$ (Uniform distribution)

We update numerical preference $H_t(\cdot)$ as follows:

$$H_{t+1}(A_t) = H_t(A_t) + \alpha_t (R_t - \bar{R}_t)(1 - \pi_t(A_t)) \quad \text{--- (1)}$$

and $H_{t+1}(a) = H_t(a) - \alpha_t (R_t - \bar{R}_t)\pi_t(a) \ \forall a \neq A_t$ — (2)

A_t is picked by UCB or ϵ-Greedy. . .

$\bar{R}$ - running average

Note:

$0 \leq \pi_t(a), \pi_t(A_t) \leq 1 \ \forall a \neq A_t$

If $R_t > \bar{R}_t$, then,

$$H_{t+1}(A_t) \geqslant H_t(A_t)$$

$$H_{t+1}(a) \leqslant H_t(a) \quad \forall \quad a \neq A_t$$

From ① & ②,

$$\sum_a H_{t+1}(A) = \sum_a H_t(a) + \cancel{\alpha_t (R_t - \bar{R}_t)} - \cancel{\alpha_t (R_t - \bar{R}_t) \sum_a \pi(a)}$$

$$\Rightarrow \sum_a H_{t+1}(a) = \sum_a H_t(a)$$

Note that higher $H_{t+1}(A_t) \Rightarrow$ higher $\pi_{t+1}(A_t)$

$\rightarrow$ prob of picking action A_t at time $t+1$ $\uparrow$

Gradient Bandit Algo as gradient ascent:

$$H_{t+1}(a) \triangleq H_t(a) + \alpha \frac{\partial E[R_t]}{\partial H_t[a]} \quad \text{——③}$$

where $E[R_t] = \sum_x \pi_t(x)\, q_*(x)$

where $q_*(x) = E[R_t \mid A_t = x]$

(Law of Expectation) (← ! Revise!)

Performation measure : $E[R_t]$

Note that,

$$\frac{\partial E[R_t]}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)}\left(\sum_x \pi_t(x) \cdot q_*(x)\right)$$

$$= \sum_x q_*(x) \frac{\partial \pi(x)}{\partial H_t(a)}$$

$$= \sum_x (q_x(x) - B_t) \frac{\partial \pi(x)}{\partial H_t(a)}$$ where B_t does not depend on 'x'.

$$\text{or} \sum_x B_t \frac{\partial \pi(x)}{\partial H_t(a)} = 0$$

$$= B_t \sum_x \frac{\partial \pi(x)}{\partial H_t(a)} \Rightarrow B_t \frac{\partial \overbrace{\sum \pi(x)}^{=1}}{\partial H_t(a)} \Rightarrow 0$$

B_t = Baseline

Role of B_t: Through suitable choice of B_t, Variance of estimators can be reduced.

Thus $$\frac{\partial E[R_t]}{\partial H_t(a)} = \sum_x (q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)}$$

$$= \sum_x \left(\pi_t(x) \cdot (q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} \Big/ \pi_t(x)\right)$$

$$= E\left[(q_*(A_t) - B_t) \cdot \frac{\partial \pi_t(A_t)}{\partial H_t(a)} \Big/ \pi_t(A_t)\right] \quad \text{— (4)}$$

Now $E[q_*(A_t)] = E[E[R_t | A_t]] = E[R_t]$

$$\Rightarrow E\left[(R_t - B_t)\frac{\partial \pi_t(A_t)}{\partial H_t(a)} \Big/ \pi_t(A_t)\right]$$

& Suppose $B_t = \bar{R}_t$ where $\bar{R}_t = \frac{1}{t}\sum_{\tau=1}^{t} R_\tau$ t-~~th~~ step avg reward.

Thus (4) =

$$E\left[R_t - \bar{R}_t \quad \frac{\partial \pi_t(A_t)}{\partial H_t(a)} \Big/ \pi_t(A_t)\right] \quad \text{—} (5)$$

Consider $\frac{\partial \pi_t(A_t)}{\partial H_t(a)}$ in (5) where

$$\pi_t(x) = \frac{e^{H_t(x)}}{\sum_{y=1}^{k} e^{H_t(y)}}$$

$$\frac{\partial H_t(A_t)}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)}\left(\frac{e^{H_t(x)}}{\sum_{y=1}^{k} e^{H_t(y)}}\right)$$

This is of the form $\frac{\partial}{\partial x}\left(\frac{u(x)}{v(x)}\right) = \frac{v(x)\cdot\frac{\partial u(x)}{x} - u(x)\cdot\frac{\partial v(x)}{\partial x}}{v(x)^2}$

$$= \frac{\frac{\partial e^{H_t(x)}}{\partial H_t(x)}\cdot\sum_{y=1}^{k} e^{H_t(y)} - e^{H_t(x)}\cdot\frac{\partial}{\partial H_t(x)}\cdot\sum e^{H_t(y)}}{\left(\sum_{y=1}^{k} e^{H_t(y)}\right)^2}$$

$$= \frac{I_{(a=x)}\, e^{H_t(x)} \cdot \sum_{y=1}^{k} e^{H_t(y)} - e^{H_t(x)} \cdot e^{H_t(a)}}{\left(\sum_{y=1}^{k} e^{H_t(y)}\right)^2}$$

$$= \frac{I_{\{a=x\}}\, e^{H_t(x)}}{\left(\sum_{y=1}^{k} e^{H_t(y)}\right)} - \frac{e^{H_t(x)} \cdot e^{H_t(a)}}{\left(\sum_{y=1}^{k} e^{H_t(y)}\right) \cdot \left(\sum_{y=1}^{k} e^{H_t(y)}\right)}$$

$$I_{\{a=x\}}\, \pi_t(x) - \pi_t(a) \cdot \pi_t(x)$$

Now recall,

$$\frac{\partial E[R_t]}{\partial H_t(a)} = E\left[(R_t - \bar{R}_t)\, \frac{\partial H_t(A_t)}{\partial H_t(a)} \,\middle|\, \pi_t(A_t)\right]$$

$$\frac{\partial \pi_t(A_t)}{\partial H_t(a)} = \pi_t(A_t)\left(I_{\{a=x\}} - \pi_t(a)\right)$$

$$\frac{\partial E[R_t]}{\partial H_t(a)} = E\left[(R_t - \bar{R}_t) \cdot \cancel{\pi_t(A_t)} \cdot \left(I_{\{a=x\}} - \pi_t(a)\right) / \cancel{\pi_t(A_t)}\right]$$

$$= E\left[(R_t - \bar{R}_t) \cdot \left(I_{\{a=A_t\}} - \pi_t(a)\right)\right]$$

Recall the recursion,

$$H_{t+1}(a) = H_t(a) + \alpha_t \frac{\partial E[R_t]}{\partial H_t(a)}$$

$$= H_t(a) + \alpha_t \cdot E\left[(R_t - \bar{R}_t)\left(I_{\{a=A_t\}} - \pi_t(a)\right)\right] \quad (5)$$

Since we do not know $E[\cdot]$, lets drop the ~~exp~~ expectation. Then the stochastic algorithm that results would be

$$H_{t+1}(a) = H_t(a) + \alpha_t\left((R_t - \bar{R}_t)\left(\mathbb{I}_{\{a=A_t\}} - \pi_t(a)\right)\right) \quad \forall a$$

$$H_t(A_t) = H_t(A_t) + \alpha\left(R_t - \bar{R}_t\,(1 - \pi_t(A_t))\right) \quad — (7)$$

$$H_t(a) = H_t(a) - \alpha\left(R_t - \bar{R}_t\,(\pi_t(a))\right) \quad \forall a \neq A_t \quad — (8)$$

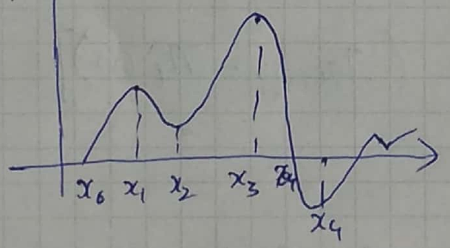
Note that (7)-(8) are same as (1) & (2)

$$H_t(a) \longrightarrow H_*(a) \quad \text{s.t}$$

$$\frac{\partial E[R_t]}{\partial H_*(a)} = 0$$

Then $H_*(a)$ is the local ~~lo~~ maxima for $E[R_t]$

Local min/ maxima



local ~~minima~~ maxima = x_1, x_3, x_5, x_7

~~local maxima~~

local minima = x

local min x_2, x_4 local max x_1, x_3,

We say that x^* is a local minimum of a function if

$\exists\ \epsilon > 0 \quad s.t\ \|x - x^*\| < \epsilon \Rightarrow f(x^*) \geq f(x)\ \forall x$

Necessary Condition for local maxima,

suppose f is differentiable $\qquad f : \mathbb{R}^d \to \mathbb{R}$

$$\nabla f(x) = 0$$

Sufficient condition for local max

suppose f is twice differentiable,

$$\nabla f(x) = 0$$

$\nabla^2 f(x)$ is negtive definite

$$y^T \nabla f(x) y < 0 \quad \forall y \neq 0 . \cancel{y\mathbb{R}}\ y \in \mathbb{R}^D$$

$$x_{n+1} = x_n + \alpha \frac{\partial f(x_n)}{\partial x_n}$$

Gradient Ascent:

$$f(x_{n+1}) = f\left(x_n + \alpha \frac{\partial f(x_n)}{\partial(x_n)}\right) \leftarrow \text{taylor expansion}$$

$$\simeq f(x_n) + \alpha \frac{\partial f(x_n)^2}{\partial(x_n)} + o(\alpha)$$

$$\frac{\partial f(x_n)^2}{\partial(x_n)} \geq 0$$

$$\Rightarrow f(x_{n+1}) \geq f(x_n)$$

The Algo with converge to local max