A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

$$A \leftarrow \begin{cases} \operatorname{argmax}_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

$R \leftarrow bandit(A)$

$N(A) \leftarrow N(A) + 1$

$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\big[R - Q(A)\big]$

Lecture -02

Immediate RL / Bandits

$q_*(1)$ $q_*(2)$ ... $q_*(k)$

[1] [2] ... [K]

$Q_t(1)$ $Q_t(2)$ ... $q_t(k)$

$Q_t(a)$ = Sum of rewards when a taken prior to t

# of times 'a' taken prior to t.

$$= \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

- Greedy $A_t = \underset{a}{\text{argmax}}\, Q_t(a)$
- exploratory action for 'ϵ' times. You pick a random action.

Any action a

$$Q_n = \frac{R_1 + R_2 + \dots + R_{n-1}}{n-1}$$

Naive implementation: Store all rewards & compute average every time.

Incremental implementation:-

$$Q_{n+1} = \frac{1}{n} \sum_{i=1}^{n} R_i$$

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

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

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

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

$$= \frac{1}{n}\left(R_n + nQ_n - Q_n\right)$$

$$Q_{n+1} = Q_n + \frac{1}{n}\left[R_n - Q_n\right]$$

New estimate — old estimate — Current reward

for $n=1$, $Q_2 = R_1$

New estimate = old estimate + stepsize [target − old estimate]

Non-stationary problems: -

Stationary bandits : reward prob. do not change over time.

$$Q_n = \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}$$

$$Q_{n+1} = Q_n + \alpha \left[R_n - Q_n \right]$$

$\alpha \in [0, 1]$

$$Q_{n+1} = Q_n + \alpha \left[R_n - Q_n \right]$$

$$= \alpha R_n + (1-\alpha) Q_n$$

$$= \alpha R_n + (1-\alpha) \left[\alpha R_{n-1} + (1-\alpha) Q_{n-1} \right]$$

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

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

$$\cdots + (1-\alpha)^{n-1} \alpha R_1 +$$

$$(1-\alpha)^n Q_1$$

$$(1-\alpha)^n + \sum_{j=1}^{n} \alpha (1-\alpha)^{n-j} = 1$$

$\alpha_n(a)$ – step size param. used to process the reward received after the n^{th} selection of a.

$$\alpha_n(a) = \frac{1}{n}$$

$$\sum_{n=1}^{\infty} a_n(a) = \infty$$

$$\sum_{n=1}^{\infty} \alpha_n^2(a) < \infty.$$

$1/n$

α

Initial values :-

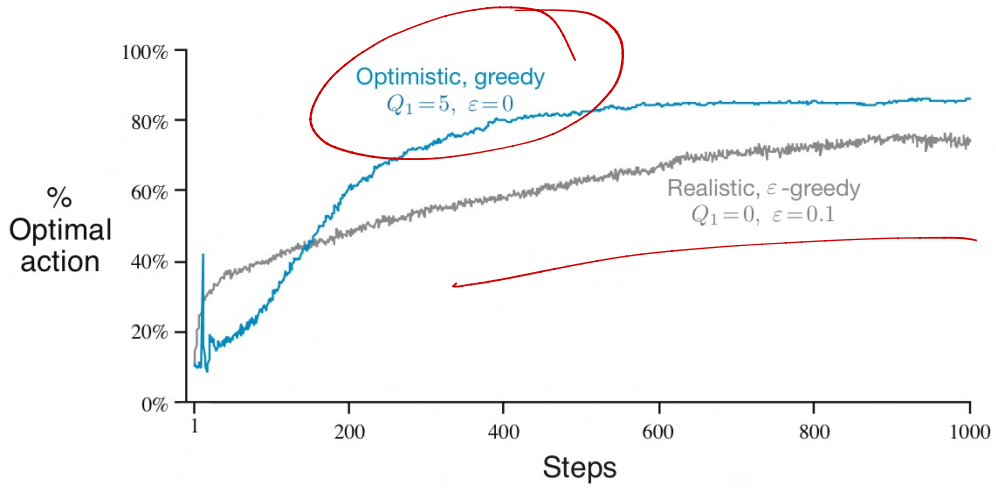
bias

$Q_1(1) = 0$ $Q_1(2) = 0$ ... $Q_1(k) = 0$

$Q_2 = R_1$

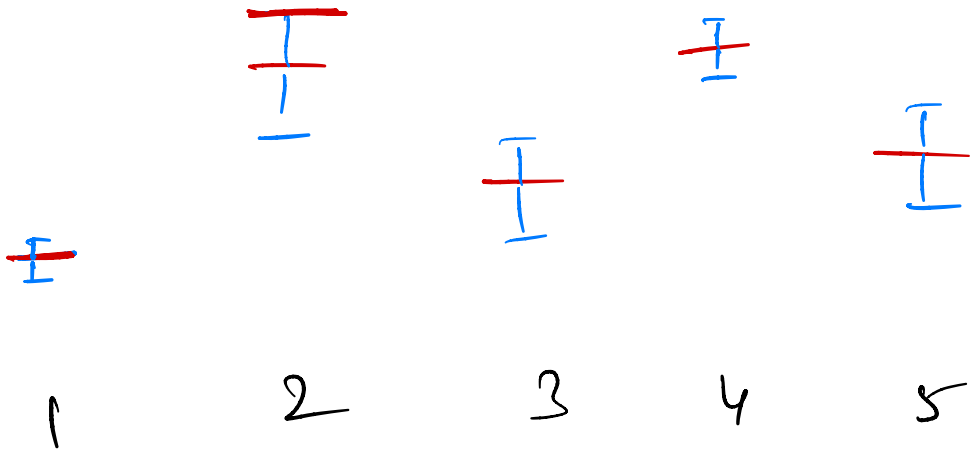
$Q_1(1) = 5$ $Q_1(2) = 5$ ... $Q_1(k) = 5$

1 2 1,5


100%
80%
60%
40%
20%
0%
% Optimal action
Optimistic, greedy
$Q_1 = 5,\ \varepsilon = 0$
Realistic, ε-greedy
$Q_1 = 0,\ \varepsilon = 0.1$
1
200
400
600
800
1000
Steps

Upper-Confidence bound (UCB) action Selection

- There are other promising actions.
- there are other actions with high uncertainty.



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

$c > 0$ Controls the degree of exploration.

- ε-Greedy
- UCB

} ~~Each~~ action-value estimation.

Gradient bandit algorithm :—

$H_t(a)$ — numerical preference for choosing action a

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

Gibbs or Boltzmann distribution.

Softmax.

$\pi_t(a)$ = prob. of taking action a.

$H_1(a) = 0$ for all a.

How to learn the preferences?

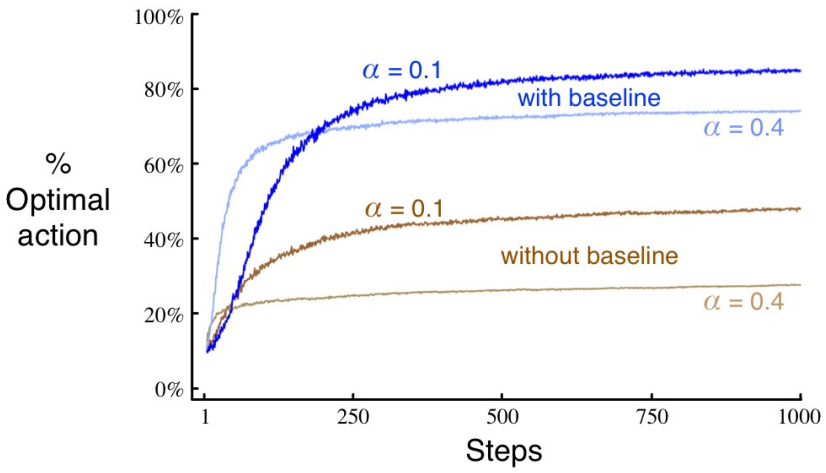
$$H_{t+1}(A_t) = H_t(A_t) + \alpha \left(R_t - \bar{R}_t\right)\left(1 - \pi_t(A_t)\right)$$

step size $\alpha > 0$

$\bar{R}_t \in \mathbb{R}$ is the average of all rewards up to t.

$$H_{t+1}(a) = H_t(a) - \alpha \left(R_t - \bar{R}_t\right)\pi_t(a) \quad \forall\, a \neq A_t$$

mean = +4



1 2 3 4 5

$\bar{R}_t$ baseline.

The bandit gradient algorithm as Stochastic gradient ascent :

Exact gradient ascent :

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

Performance $\Phi = \mathbb{E}[R_t]$

$$= \sum_x \pi_t(x)\, q_*(x)$$

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

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

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

$\sum_x \frac{\partial \pi_t(x)}{\partial H_t(a)} = 0$

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

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

$$= \mathbb{E}\left[\left(q_*(x) - B_t\right) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} \Big/ \pi_t(A_t) \right]$$

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

Note: $\mathbb{E}[R_t \mid A_t] = q_*(A_t)$

$$\frac{\partial \pi_t(x)}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \pi_t(x)$$

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

$\partial \left(\frac{u}{v}\right)$

$\frac{v\,du - u\,dv}{v^2}$

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

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

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

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

$$= \mathbb{1}_{a=x} \pi_t(x) - \pi_t(x) \cdot \pi_t(a)$$

$$\frac{\partial \pi_t(x)}{\partial H_t(a)} = \pi_t(x) \left(\mathbb{1}_{a=x} - \pi_t(a) \right)$$

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

$$= \mathbb{E}\left[(R_t - \bar{R}_t) \pi_t(A_t) \left(\mathbb{1}_{a=A_t} - \pi_t(a) \right) \Big/ \pi_t(A_t) \right]$$

$$= \mathbb{E}\left[(R_t - \bar{R}_t) \left(\mathbb{1}_{a=A_t} - \pi_t(a) \right) \right]$$

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

Contextual bandits :-

Bandits → Non-associative tasks.

Full RL — many situations.

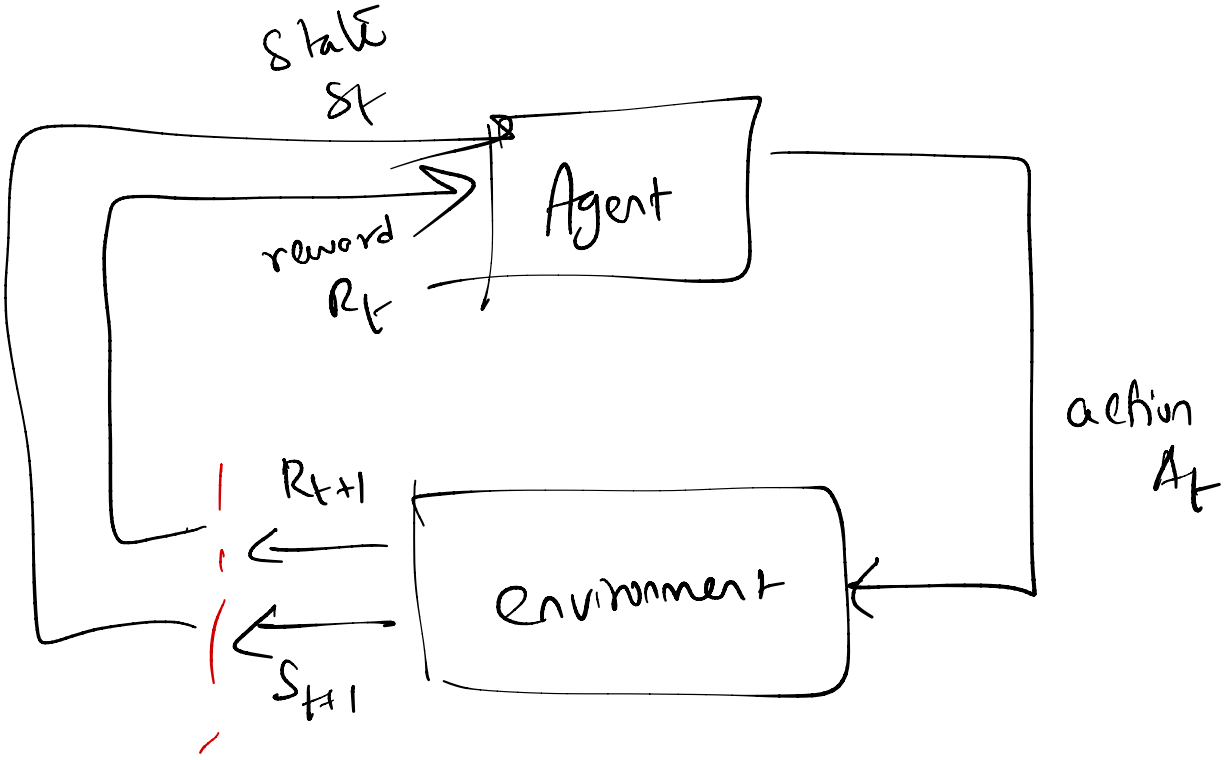
Contextual bandits

Context or id

Ad-placement;

Context: User history.

Finite Markov Decision Processes (MDP)



Sequence / trajectory: $S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3 \ldots$

finite MDP (S, A, R)

R_t, S_t — well defined discrete prob. distrib. that depend only on prev. state & action

$$p(s', r \mid s, a) = \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}$$

Dynamics fn of the MDP.

$$\forall\, s', s \in \mathcal{S},\ r \in \mathcal{R},\ a \in \mathcal{A}(s)$$

Note: $$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \mid s, a) = 1$$

$$\forall\, s \in \mathcal{S},\ a \in \mathcal{A}(s)$$

$$p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \longrightarrow [0, 1]$$

State-transition prob:

$$p(s' \mid s, a) = \sum_{r \in \mathcal{R}} p(s', r \mid s, a)$$

Expected reward for s, a pair:

$$r(s, a) = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r \mid s, a)$$

Markov property:

$$p(S_t, R_t \mid S_{t-1}\ A_{t-1})$$

$$= p(S_t, R_t \mid S_{t-1}\ A_{t-1}, R_{t-1}, S_{t-2}\ A_{t-2} \ldots)$$

boundary between Agent & env:

The boundary represents the limit of agent's absolute control

Not its knowledge.

MDP — ~~one step~~ the actions ✓

— the state ✓

— the reward. ✓

Goals & rewards: —

↓

(what to achieve

Not how to achieve it.

Returns

Cumulative reward = return

$$G_t = R_{t+1} + R_{t+2} + \cdots + R_T$$

final time step

episodes

Tasks

- episodic tasks
- Continuing tasks.

discount rate

$0 \leq \gamma \leq 1$

$\gamma = 0$ "myopic"

$\gamma = 1$

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \cdots$$

$$= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$