Notes to Multi-armed bandit paper

btypoon*

Aug-2023

Contents

*eduhk

1 Game Rule and Symbols definition

- Let there be in total K slot machines. Subscript k define the k-th machine from the set of K machines.

$$k \in \{1, 2, ..., K\}$$

- Let the game to be round based. In each round, the player can only choose to play one slot machines. Define the concept of time t to be the decrete round number in the set of T total rounds

$$t \in \{1, 2, ..., T\}$$

- Define the chosen k-th slot machine at time t to be

$$a^t \in \{1, 2, ..., K\} \; at \; time \; t$$

- Let the variable x be the reward of slot machine. Define $P_k(x)$ to be the probability density function of the rewards x of k-th machine.

$$\text{Probability of getting } x_k \text{ from k-th machine } = P_k(x_k)dx$$

- Define expected value of reward of k-th machine as

$$\langle x \rangle_k = \int_{-\infty}^{\infty} x P_k(x) dx$$

- The true values of $\langle x \rangle$ are unknown for all machines
- Base on the expected value of rewards of all machine, in every round, the player should always choose to play the slot machine with the highest EV.

$$chosen \; a_t = Best \; k = max\langle x \rangle_k$$

- In other words, player want to maximize their cumulative reward. Define cumulative rewards up to T rounds as R:

$$R = \sum_{t=0}^{T} x_{a^t}$$

- In theory, the cumulative regret is more useful for analysis. Define cumulative regret as small letter r:

$$\begin{aligned} r &= \sum_{t=0}^{T} (max\langle x \rangle_k - x_{a^t}) \\ &= (T \times max\langle x \rangle_k) - (cumulative \; reward) \end{aligned}$$

- Define n_k^t as the number of times k-th slot machines has been pulled up to time t.

$$n_k^t = \sum_{\tau=1}^{t} \delta(a^\tau, k) + n_k^0$$

- In some strategies, all slots machine was pulled a fixed amount of times for initilization. Therefore, n_k^0 was seperated from n_k^t as the initial pull number.

$$n_k^0 = \text{initial pull number for k-th machines at time t=0}$$

- Define the cumulative reward contributed by k-th slot machine up to time t as s_k^t.

$$s_k^t = \sum_{\tau=1}^{t} x_{a^\tau} \delta(a^\tau, k) + s_k^0$$

- Similarly, due to initilization, s_k^0 was seperated from s_k^t.

$$s_k^0 = \text{initial reward contributed by k-th machines at time t=0}$$

- Connection between cumulative reward R, x and s

$$R = \sum_{t=0}^{T} x_{a^t} = \sum_{k=1}^{K} s_k^T$$

- Connection between cumulative regret r, x and s if there are m initial pull for all K slot machines

$$\begin{aligned} r &= (\text{max reward}) \ - \ (\text{cumulative reward}) \\ &= [(T + mK) \times max\langle x \rangle_k] - \sum_{k=1}^{K} s_k^T \end{aligned}$$

- Since the true values of $\langle x \rangle_k$ are unknown for all machines, we use sample mean as the best estimator for the population mean. Define the μ_k^t as the sample mean of k-th machine at time t

$$\text{sample mean } \mu_k^t = \frac{s_k^t}{n_k^t}$$

2 Balance between exploration and exploitation modes

2.1 Method 1: Pull all slot machines once intially

At time $t = 0$, roll all K machine once. (*homogeneous initial exploration*). Afterwards ($t \in 1, 2, ..., T$), choose only the best machine. i.e. the machine with the heighest expected rewarded $max\langle x\rangle_k$ (*exploitation*)

2.2 Method 2: Pull all slot machines m times intially

At time $t = 0$, roll all K machine m times. (*homogeneous initial exploration*). Afterwards ($t \in 1, 2, ..., T$), choose only the best machine. i.e. the machine with the heighest expected rewared $max\langle x\rangle_k$ (*exploitation*)

2.3 Method 3: Simulated annealling

Base on a predefined *acceptance probability function*, the player choose to either

1. randomly pull a slot machines to gain new information
2. pull the slot machine with highest $\langle x\rangle_k$

2.3.1 ϵ-greedy algorithm

In this method, the *acceptance probability function* is a constant of ϵ.

$$\text{Acceptance Probability Function} = \epsilon$$

2.3.2 Softmax algorithm

In this method,the *acceptance probability function* is defined to be a *Boltzmann distribution*

$$\begin{aligned}\text{Acceptance Probability Function} &= \text{Boltzmann distribution} \\ &= \frac{e^{\beta\mu_k^t}}{\sum_{j=1}^{K} e^{\beta\mu_j^t}} \end{aligned} \tag{1}$$

where

$$\begin{aligned}\beta &= \frac{1}{k_B T'} \\ k_B &= \text{Boltzmann constant} = 1.380649 \times 10^{-23} J/K \\ T' &= \text{temperature}\end{aligned}$$

2.4 Method 4: Upper confidence bound algorithm

This method choose the best slot machine by the upper bound of estimated expected value instead of the expected value itself. It is quoted as the principle of *"optimism in the face of uncertainty"*

Define B_k^t as the upper confidence bound of the sample mean estimation for k-th machine up to time *table*.

$$\begin{aligned} B_k^t &= \mu_k^t + \text{constant} \times \sigma_{\mu_k^t} \\ &= \frac{s_k^t}{n_k^t} + \text{constant} \times \frac{\sigma}{\sqrt{n_k^t}} \\ &= \frac{s_k^t}{n_k^t} + b^t \frac{1}{\sqrt{n_k^t}} \\ b^t &= \text{parameter on controlling the confidence level} \\ \sigma_{\mu_k^t} &= \text{standard deviation of sample mean} \\ &= \frac{\sigma}{\sqrt{n_k^t}} \end{aligned}$$

For optimal setting, use

$$b^t = c\sqrt{\log(mK+t)}$$

where, c is a parameter for turning the level of exploration

$$\therefore B_k^t = \frac{s_k^t}{n_k^t} + c\sqrt{\frac{\log(mK+t)}{n_k^t}} \tag{2}$$

We can see that the definition of the upper confidence bound itself implied balanced between the exploration and exploitation modes. When the number of pull is low, the upper bound was large, which encourage the machine to be chosen and gain more information on it. When the number of pull is high, the upper bound tends to be $\langle x \rangle_k$, which return to our original stategy of choosing the slot machine with highest expected value.

3 Combination of simulated annealling and Upper confidence bound algorithm

Another complex approach is to use both (1) and (2). By substituting the sample mean μ in (1) with the upper confidence bound B in (2), we have a new *acceptance probability function*

$$\begin{aligned} &\text{Acceptance Probability Function for } k\text{-th machine} \\ &= \frac{e^{\beta B_k^t}}{\sum_{j=1}^{K} e^{\beta B_j^t}} \end{aligned}$$