2021-07-14 MABs

Monday, July 12, 2021 9:34

Multi
Armed
Bandit

Bandit 1

$X_{1,1}, X_{1,2}, \dots$ i.i.d
$\sim D_1$

Bandit 2

$X_{2,1}, X_{2,2}, \dots$ i.i.d
$\sim D_2$

⋮

Bandit n

$X_{n,1}, X_{n,2}, \dots$ i.i.d
$\sim D_3$

$X_{i,j} \in [0,1]$

Objective Maximize reward –

What do you do ?

Strategy: 1. Try out some arms
(exploration)

2. Pull the best performing arms
(exploitation)

Question - When to transition from
1. to 2. ?

- How good can we theoretically
perform ?

Notations K arms, T rounds

For each round $t \in \{1, \dots, T\}$

the algorithm chooses an arm a_t

giving a reward $r_t \sim D_{a_t}$

Distribution of arm a_t. Does NOT depend on t!

For simplicity,

$$r_t \in [0,1]$$

Write

$$\mu(a) = \mathbb{E}[r_t] \quad , \quad \mu^* = \max_a \mu(a)$$

$$\Delta(a) = \mu^* - \mu(a)$$

Regret at round T

$$R(T) = T\mu^* - \sum_t \mu(a_t)$$

NOT r_t

The regret deals whith expected (NOT actual) gains, and compares them to the best arm

Objective Minimize $\mathbb{E}[R(T)]$

Note $\mu(a), \mu^*, \Delta(a)$ are NOT known

↪ It seems impossible to calculate anything.

Regardless of the algorithm, we can classify the OBSERVED performance of the MAB into 2 categories

- Good performance

  observed mean

  $\mathcal{E} \; : \; \forall a \quad |\mu(a) - \bar{\mu}(a)|$ is small
- Bad performance $\neg \mathcal{E}$

AND we can estimate / bound

$P[\mathcal{E}]$, $\mathbb{E}[R(T) \mid \mathcal{E}]$

$P[\neg\mathcal{E}]$, $\mathbb{E}[R(T) \mid \neg\mathcal{E}]$

Theorem (Hoeffding's inequality)

$X_1, \dots, X_T \in [0,1]$ i.i.d ← arbitrary distribution!

$\mu = \mathbb{E}[X_i]$, $\bar{\mu} = \frac{1}{n}\sum_i X_i$

Then

$\equiv \mathbb{P}[\mathcal{E}]$ good performance!

$$\boxed{\mathbb{P}[|\mu - \bar{\mu}| \leq \epsilon] \geq 1 - 2\exp(-2T^2\epsilon^2)}$$

ϵ is called the confidence radius

$[\mu-\epsilon, \mu+\epsilon]$ is called the confidence interval

Let's tweak the formula

$$\epsilon = \sqrt{\alpha \log T / A}$$

$$\boxed{\mathbb{P}\left[|\mu - \bar{\mu}| \leq \sqrt{\alpha \frac{\log T}{A}}\right] \geq 1 - 2T^{-2\alpha T^2 / A}}$$

```
Algorithm 1 : Uniform exploration
   Fix $N < T/K$
1. (exploration) Try each arm $N$ times
2. Let $a^*$ be the best performing arm
      $a^* = \arg\max \bar{\mu}(a)$
3. (exploitation) Pull $a^*$ in all
      remaining rounds $(T-N)$
```

Recall

$$\mathbb{P}\left[|\mu-\bar{\mu}| \le \sqrt{\alpha \frac{\log T}{A}}\right] \ge 1-2T^{-2\alpha T^2/A}$$

setting $\alpha = 2$ and $A = N$ we get that for all arm $a \in K$

$$\mathbb{P}\left[|\mu(a)-\bar{\mu}(a)| \le e\right]$$

$$= \mathbb{P}\left[|\mu(a)-\bar{\mu}(a)| \le \sqrt{2\frac{\log T}{N}}\right]$$

$$\ge 1-2T^{-4T^2/N}$$

$$\ge 1-2T^{-4}$$

Few things we neet to understand

① Good performance

$$\mathcal{E} : \forall a \quad |\mu(a)-\bar{\mu}(a)| \le e$$

$$\Rightarrow \neg\mathcal{E} : \exists a \quad |\mu(a)-\bar{\mu}(a)| > e$$

$$\Rightarrow \boxed{P[\neg\mathcal{E}] \le 2T^{-4} \sim \mathcal{O}(T^{-4})}$$

② During the exploration phase, each step contributes to the end regret $R(T)$ by at most $\boxed{1}$

③ Let $a \in K$, a^+ be the best arm $\overset{\text{unknown!}}{}$

arm (after the exploration phase)

Assuming $\mathcal{E}$:

$$\mu(a^+) + \epsilon \geq \bar{\mu}(a^+) \geq \bar{\mu}(a) \geq \mu(a) - \epsilon$$

So if we choose the wrong arm for the exploitation phase $(a^* \neq a^+)$ then the contribution to $R(T)$ is at most

$$\mu(a^+) - \mu(a^*) \leq \boxed{2\epsilon}$$

② + ③ Assuming $\mathcal{E}$

$R(T) =$ regret during exploration

+ regret during exploitation

$$\leq KN \times 1 + (T - KN) \times 2\epsilon$$

$$\leq \boxed{KN + 2\epsilon T}$$

④ Assuming $\mathcal{E}$, and choosing

$$N = (T/K)^{2/3} (\log T)^{1/3}$$

We get

$$R(T) \leq KN + 2\overset{\sqrt{2 \frac{\log T}{N}}}{\epsilon} T$$

$$= K^{1/3} T^{2/3} (\log T)^{1/3} \, 2\sqrt{2}$$

$$\sim \boxed{O\left(T^{2/3} (\log T)^{1/3}\right)}$$

⑤ Finally

$$\mathbb{E}[R(T)] = \mathbb{E}[R(T) \mid \mathcal{E}] \, \mathbb{P}[\mathcal{E}] + \mathbb{E}[R(T) \mid \neg\mathcal{E}] \, \mathbb{P}[\neg\mathcal{E}]$$

For a generic N we would have

$$\mathbb{E}[R(T)] = \mathbb{E}[R(T) \mid \mathcal{E}] \, \mathbb{P}[\mathcal{E}] + \mathbb{E}[R(T) \mid \neg\mathcal{E}] \, \mathbb{P}[\neg\mathcal{E}]$$

$$\leq \overset{④}{\mathcal{O}\left(T^{2/3}(\log T)^{1/3}\right)} \times \overset{\mathbb{P}[\dots]\leq 1}{1}$$

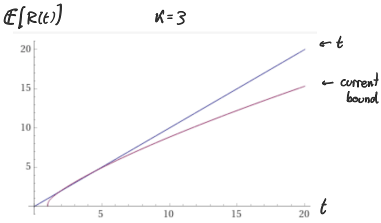
$$+ \underbrace{T}_{\text{Max possible regret}} \times \overset{①}{\mathcal{O}(T^{-4})}$$

$$= \boxed{\mathcal{O}\left(T^{2/3}(K \log T)^{1/3}\right)}$$

$$\leq KN + 2\varepsilon T + T\mathcal{O}(T^{-4})$$

$$= KN + 2T\sqrt{2\frac{\log T}{N}} + T\mathcal{O}(T^{-4})$$

$$\sim \boxed{\mathcal{O}\left(KN + T\sqrt{\frac{\log T}{N}}\right)}$$



```
Algorithm 2: ε-greedy
  Decide on a schedule
      $\varepsilon = \varepsilon_t : \{1, \dots, T\} \to [0,1]$
  for t in 1, ..., T
    - with probability $\varepsilon_t$
        (explore) pull a random arm
    - otherwise
        (exploit) pull the best performing arm
```

Theorem With $\varepsilon_t = t^{-1/3}(K \log t)^{1/3}$ we get the same bound as with uniform exploration

Let´s try an figure out a more clever algorithm

Notation During the run of a given algorithm, let $n_t(a)$ be the number of times arm a has been pulled up to round t

$$\sum_a n_t(a) = t$$

Confidence intervals

- $\rho_t(a) = \sqrt{2 \frac{\log T}{n_t(a)}}$ "confidence radius around the value of $\mu(a)$ at round t"
- $UCB_t(a) = \bar{\mu}_t(a) + \rho_t(a)$
  observed mean at round t
- $LCB_t(a) = \bar{\mu}_t(a) - \rho_t(a)$
- $[LCB_t(a), UCB_t(a)]$ is called the confidence interval

Idea An arm a is empirically worse than another arm b if

$$UCB_t(a) < LCB_t(b)$$

assuming some "good performance" condition $\mathcal{E}$

```
Algorithm 3: Successive elimination
1. Mark all arms as "active"
2. Until the polling budget is exhausted
    - (exploration) pull all active arms
    - (exploitation) deactivate all
      empirically suboptimal arms
```

Let the "good performance" condition be

$$\mathcal{E} : \forall a \ \forall t \quad |\mu(a) - \bar{\mu}_{-}(a)| \leq \rho_t(a)$$

Unfortunately, we cannot directly apply Hoeffding's inequality

$$\mathbb{P}\left[|\mu - \bar{\mu}| \leq \sqrt{\alpha \frac{\log T}{A}}\right] \geq 1 - 2T^{-2\alpha T^2/A}$$

because $A = n_t(a)$ is not known in advance... But using probabilistic technicalities, we still get the estimate:

① **Lemma** $\mathbb{P}[\mathcal{E}] \geq 1 - 2/T^2$

② Assuming $\mathcal{E}$, note that the best unknown! arm a^+ cannot be deactivated.

Thus, an arm a can only remain active if

$$UCB_t(a) \geq LCB_t(a^+)$$

If we are at the end of an iteration,

$$\Delta(a) = \mu(a^+) - \mu(a)$$

recall

$$\leq 2(\rho_t(a^+) + \rho_t(a))$$

$$= 4\, c_t(a)$$

since all active arms have been pulled the same amount of times

Assume now that a gets deactivated at the next iteration, i.e.

$$UCB_{t+1}(a) < LCB_{t+1}(a^+)$$

Then

$$c_{t+1}(a) = \sqrt{2\frac{\log T}{n_{t+1}(a)}} = \sqrt{2\frac{\log T}{n_T(a)}} = c_T(a)$$

$$\Rightarrow \quad \Delta(a) \leq 4\, c_T(a)$$

$$\sim \boxed{\mathcal{O}\left(\sqrt{\frac{\log T}{n_T(a)}}\right)}$$

for all suboptimal arm a

③ The regret accumulated by pulling the suboptimal arm a is

$$R_a(t) = n_t(a)\,\Delta(a)$$

Assuming $\mathcal{E}$,

$$R_a(t) = n_t(a)\,\Delta(a)$$

$$\overset{②}{\leq} n_t(a)\, \mathcal{O}\left(\sqrt{\frac{\log T}{n_t(a)}}\right)$$

$$= \mathcal{O}\left(\sqrt{\log T \; n_t(a)}\right)$$

Then,

$$R(t) = \sum_a R_a(t)$$

$$\leq \mathcal{O}\left(\sqrt{\log T}\sum_a \sqrt{n_t(a)}\right)$$

by Jensen's inequality,

$$\sum_a \sqrt{n_t(a)} \leq K\sqrt{\frac{1}{K}\sum_a n_t(a)}$$

$$\leq \mathcal{O}\left(\sqrt{\log T}\; K\sqrt{\frac{1}{K}\sum_a n_t(a)}\right)$$

$n_t(a) \leq t$

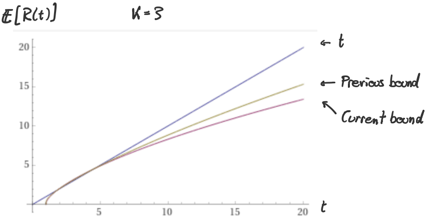
$$\le \boxed{\mathcal{O}\left(\sqrt{Kt \log T}\right)}$$

④ Finally,

$$\mathbb{E}[R(T)] = \mathbb{E}[R(T) \mid \mathcal{E}]\, \mathbb{P}[\mathcal{E}] + \mathbb{E}[R(T) \mid \neg\mathcal{E}]\, \mathbb{P}[\neg\mathcal{E}]$$

$$\le \overset{\textcircled{3}}{\mathcal{O}}\left(\sqrt{KT \log T}\right) \times 1 + T \times \mathcal{O}\left(2/T^2\right)^{\textcircled{1}}$$

$$= \boxed{\boxed{\mathcal{O}\left(\sqrt{KT \log T}\right)}}$$



Common variation :

```
Algorithm 5 : UCB1
1. try all arms
2. for $t = K+1, \dots, T$
     - pull some arm $a$ that maximizes
           $UCB_t(a)$
```

Theorem UCB1 gives the same bound on $\mathbb{E}[R(T)]$ as successive elimination.

Epilogue: exploration vs. exploitation

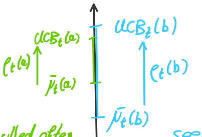
Recall

$$UCB_t(a) = \bar{\mu}_t(a) + e_t(a)$$

gets high if a is a good arm

$$e_t(a) = \sqrt{2 \log T / n_t(a)}$$

large on poorly explored arms



good arm, pulled often
$\rightsquigarrow$ $\bar{\mu}_t(a)$ high
$e_t(a)$ small

seemingly worse arm, neglected
$\rightsquigarrow$ $\bar{\mu}_t(b)$ small
$e_t(b)$ large