6.1 The K-Armed Bandit Problem

Bandit problem : a problem to deal with exploration and exploitation. For the clear low risk machine, one should exploit it more; On the other hand, for more profit, one should also try to explore more unknow risk machines, despite high risk we might face with.

Settings: let $\gamma = 0$ in unknow MDPs.

we have K decisions ("arm"). When we play arm $a \in \{1, 2, \cdots K\}$, we obtain a random $r_a \in [-1, 1]$ from $R(a) \in \Delta[-1, 1]$, which has mean reward.

$$E_{r_a \sim R(a)}[r_a] = \mu_a \quad \in [-1, 1]$$

Every iter t, the learner will pick an arm $I_t \in [1, 2, \cdots K]$.

Define cumulative regret as:

$$R_T = T \cdot \max_i \mu_i - \sum_{t=0}^{T-1} \mu_{I_t}$$

we denote ① $a^* = \arg\max_i \mu_i$,

② $\Delta_a = \mu_{a^*} - \mu_a$.

Theorem 6.1 There exist an algorithm s.t. with prob $\geq 1-\delta$,

$$R_T = O\left(\min\left\{\sqrt{KT \cdot \ln(TK/\delta)},\ \sum_{a \neq a^*} \frac{\ln(TK/\delta)}{\Delta_a}\right\} + K\right)$$

1. The upper confidence bound (UCB) algorithm

Pseudo code of UCB algorithm:

1: Play each arm once, denote as $\{r_a \mid a = 1, \cdots, K\}$

2: for $t = 1 \to T-K$ do

for $t = 1 \to T-K$ do

$\quad I_t = \text{argmax}_{i \in [K]} \left(\hat{\mu}_i^t + \sqrt{\frac{\log(TK/\delta)}{N_i^t}} \right)$

$\quad r_t := r_{I_t}$

end for

where we maintain counts of each arm:

$$N_a^t = 1 + \sum_{i=1}^{t-1} \mathbb{1}\{I_i = a\}$$

and we compute empirical mean:

$$\hat{\mu}_a^t = \frac{1}{N_a^t} \left(r_a + \sum_{i=1}^{t-1} \mathbb{1}\{I_i = a\} r_i \right)$$

and we maintain the upper confidence upper for each arm:

$$\hat{\mu}_a^t + 2\sqrt{\frac{\ln(TK/\delta)}{N_a^t}}$$

Lemma 6.2 [Upper Confidence Bound]

For all $t \in [1, 2, \cdots, k]$ and $a \in [1, 2, \cdots, K]$. We have the prob. $\geq 1-\delta$ that $|\hat{\mu}_a^t - \mu_a^t| \leq 2\sqrt{\frac{\ln(TK/\delta)}{N_a^t}}$

Pf: We use Hoeffding-Azuma inequation: suppose $X_1, \cdots, X_T$ is a martingale difference sequence where each X_t is a σ_t sub-Gaussian, Then for all $\varepsilon > 0$, $N > 0$

$$P\left(\sum_{i=1}^{N} X_i \geq \varepsilon \right) \leq \exp\left(\frac{-\varepsilon^2}{2\sum_{i=1}^{N} \sigma_i^2} \right)$$

$$\hat{\mu}_a^t - \mu_a^t = \frac{1}{N_a^t} \left[r_a + \sum_{i=1}^{t-1} \mathbb{1}_{\{I_i = a\}} r_i \right] - \mu_a^t$$

$$= \frac{1}{N_a^t} \left[r_a - \mu_a^t + \sum_{i=1}^{t-1} \left(\mathbb{1}_{\{I_i = a\}} r_i - \mu_a^t \right) \right]$$

if we assume that $r_a \sim N(\mu_a^t, 1)$

assume $\varepsilon = 2\sqrt{\frac{\ln(TK/\delta)}{N_a^t}} \Rightarrow P(\hat{\mu}_a^t - \mu_a^t \geq \varepsilon) \leq e^{-2} \frac{\delta}{TK}$

$\Rightarrow P(|\hat{\mu}_a^t - \mu_a^t| \leq \varepsilon) \geq 1 - 2 \cdot e^{-2} \frac{\delta}{TK} \geq 1 - \frac{\delta}{TK}$. $\square$

Pf from book.

We consider a fixed arm a, define $X_0 = r_a - \mu_a$,

$X_i = \mathbb{1}\{I_i = a\}(r_i - \mu_a)$, $i = 1, 2, \dots, T$.

if $\mathbb{1}\{I_i = a\} = 1$, $|X_i| = |r_i - \mu_a| \leq |r_i| - |\mu_a| \leq 2$.

$E[X_i | \mathcal{H}_{<i}] = 0$ sin I_t is determined when $\mathcal{H}_{<t}$ is known.

$\Rightarrow \{X_t\}$ is martigale difference sequence.

Via Azuma-Hoeffding's ineq.

$$\left|\sum_{i=1}^{t-1} X_i\right| = \left|N_a^t \hat{\mu}_a^t - N_a^t \mu_a\right| \leq 2\sqrt{\ln(1/\delta) N_a^t}$$

$$\Rightarrow \quad |\hat{\mu}_a^t - \mu_a^t| \leq 2\sqrt{\ln(1/\delta)/N_a^t}$$

Apply union bound over $[T]$ and $[K]$ we prove the lemma.

We now prove Th 6.1

> **Theorem 6.1** There exist an algorithm s.t. with prob $\geq 1-\delta$,
> $$R_T = O\left(\min\left\{\sqrt{KT \cdot \ln(TK/\delta)}, \sum_{a \neq a^*} \frac{\ln(TK/\delta)}{\Delta_a}\right\} + K\right)$$

Pf $\quad \mu_a \leq \hat{\mu}_a^t + 2\sqrt{\dfrac{\ln(TK/\delta)}{N_a^t}} \quad \forall a, t.$

(Assume Lemma 6.2 ineq. holds)

$$\mu_{a^*} - \mu_{I_t} \leq \hat{\mu}_{I_t} + 2\sqrt{\frac{\ln(TK/\delta)}{N_{I_t}^t}} - \mu_{I_t}$$

$$\leq 4\sqrt{\frac{\ln(TK/\delta)}{N_{I_t}^t}}$$

$$\Rightarrow \sum_{t=0}^{T-1}(\mu_{a^*} - \mu_{I_t}) \leq 4\sqrt{\ln(TK/\delta)} \sum_{t=0}^{T-1}\sqrt{\frac{1}{N_{I_t}^t}}$$

$$\leq 4\sqrt{\ln(TK/\delta)} \sum_a \sum_{i=1}^{N_a^T} \frac{1}{\sqrt{i}}$$

$$\leq 8 \sqrt{\ln(TK/\delta)} \sum_a \sqrt{N_a}$$
$$\leq 8 \sqrt{\ln(TK/\delta)} \sqrt{K \sum_a N_a^T} \quad \text{(Cauchy ineq.)}$$
$$\leq 8 \sqrt{\ln(TK/\delta)} \sqrt{KT}$$

On the other hand, if $\Delta_a > 0$ for each $a \neq a^*$,

$$N_a^T \leq \frac{4 \ln(TK/\delta)}{\Delta_a^2}$$

because if upper confidence boundary of $a < \mu_{a^*}$, then ~~ar~~ a will never be selected again (by the algorithm).

$$\Rightarrow \sum_{t=0}^{T-1} \mu_{a^*} - \mu_{I_t} \leq \sum_{a \neq a^*} N_a^T \Delta_a \leq \sum_{a \neq a^*} \frac{4 \ln(TK/\delta)}{\Delta_a}$$

Since first K steps have max K regret. the proof is completed $\square$