## 6.4 Summary

In this chapter, we mainly studied two algorithms about Bandits.

A bandit problem is about the balance of exploit & explore, the goal of such a problem is to find the optimal action at each time. We care about the estimate of the regret.

First we learned the $K$-Armed Bandit Problem.

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

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

4: $\quad r_t := r_{I_t}$

5: end for

The regret of the algorithm above can be concluded as:

[Th 6.1] with prob. $\geq 1-\delta$,

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

Then to deal with a large/infinite $K$ arms, we learned the Linear Bandits.

Pseudo code: The Linear UCB algorithm

Input: $\lambda$, $\beta_t$

1: for $t = 0, 1, \cdots$ do

2: $\quad$ Execute $x_t = \underset{x \in D}{\text{argmax}} \max_{\mu \in BALL_t} \mu \cdot x$

and observe $r_t$

3: Update $\text{BALL}_{t+1}$

4: End for.

The estimate of the regret can be concluded as:

[Th. 6.3] Suppose $|\mu^* \cdot x| \le 1$ for all $x \in D$, $\|\mu^*\| \le W$, $\|x\| \le B$ and $\eta_t$ is $\sigma^2$ sub-Gaussian. Set

$$\lambda = \sigma^2/W^2, \quad \beta_t := \sigma^2\left(2 + 4d\log\left(1 + \frac{tB^2W^2}{d}\right) + 8\log(4/\delta)\right)$$

We have that with prob. $\ge 1-\delta$, for all $T \ge 0$

$$R_T \le c\sigma\sqrt{T}\left(d\log\left(1 + \frac{TB^2W^2}{d\sigma^2}\right) + \log(4/\delta)\right)$$

where the prob. comes from the prob. that $\mu^* \in \text{BALL}_t$.

Addition:

The idea of upper confidence bounds comes from the estimate of $|\hat{\mu}(a) - \mu(a)|$ by the Hoeffding's inequlity, which says

$$|\hat{\mu}(a) - \mu(a)| \le \sqrt{\ln(Tk/\delta)/N(a)} \quad \text{with prop} \ge 1-\delta.$$

There are two reasons for trying the arm $I_t$ with highest UCB:

(i): $I_t$ has large confidence interval, which means it has not been tried many times yet. [high uncertainty]

i.e. We do exploration in this case!

(ii): $I_t$ has small confidence interval, then it is simply a good arm.

i.e. We do exploitation in this case!