

# Lecture 6 - Upper Confidence Bound Algorithms

Guiliang Liu

The Chinese University of Hong Kong, Shenzhen

DDA4230: Reinforcement Learning  
Course Page: [\[Click\]](#)

# DDA 4230 Resources

Join our Wechat discussion group.



Check our course page.



Course Page Link (all the course relevant materials will be posted here):

[https://guiliang.github.io/courses/cuhk-dda-4230/dda\\_4230.html](https://guiliang.github.io/courses/cuhk-dda-4230/dda_4230.html)



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# Motivation

Recall the previously introduced algorithms:

- Execute  $\varepsilon$ -greedy by choosing  $\varepsilon_t = \min\{1, Ct^{-1}\Delta_{\min}^{-2}m\}$ .
- Run ETC with  $k = \lceil \frac{2}{\Delta^2} W(\frac{T^2\Delta^4}{32\pi}) \rceil$

**Limitation** of  $\varepsilon$ -greedy and ETC.

1. Executing the algorithm **requires the knowledge of  $\Delta$** , which is usually not available in real applications.
2. The algorithm **uses  $T$** , but the horizon is unknown in real applications.
3. The theoretical result obtained by ETC is applied to **2-armed bandits only**.



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# The UCB Algorithms

---

**Algorithm 1:** The UCB algorithm.

---

**Input:**  $\delta$ : confidence level

**Output:**  $a_t, t \in \{0, 1, \dots, T\}$

**while**  $t \leq T - 1$  **do**

$$a_t = \arg \max_{i \in [m]} \text{UCB}_i(t - 1, \delta),$$

where ties break arbitrarily and for  $i \in [m]$ ,

$$\text{UCB}_i(t - 1, \delta) = \begin{cases} \infty, & N_{i,t-1} = 0, \\ \frac{1}{N_{i,t-1}} \sum_{t' \leq t-1} r_{t'} \mathbb{1}\{a_{t'} = i\} + \sqrt{\frac{2 \log(1/\delta)}{N_{i,t-1}}}, & N_{i,t-1} > 0; \end{cases}$$

# The Optimism Principle

The UCB algorithm is based on the principle of optimism in the face of uncertainty, which states that

*one should act as if the environment is as nice as plausibly possible.*

In fact, this principle is applicable to other bandit algorithms as well and is beyond the finite-armed stochastic bandit problem.



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# The Optimism Principle

For UCB, the optimism principle means using the data observed so far to assign to each arm a value, called the upper confidence bound. The first term,

$$\hat{\mu}_{i,t-1} = \frac{1}{N_{i,t-1}} \sum_{t' \leq t-1} r_{t'} \mathbb{1}\{a_{t'} = i\},$$

is the empirical mean of the rewards collected from arm  $i$ , where

$N_{i,t-1} = \sum_{t' \leq t-1} \mathbb{1}\{a_{t'} = i\}$  is the number of times arm  $i$  has been pulled up to time  $t - 1$ .



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# The Optimism Principle

Recall the Chernoff-Hoeffding bound for  $n$  independent 1-sub-Gaussian random variables

$$\mathbb{P}(\bar{X} - \mathbb{E}[\bar{X}] \geq z) \leq \exp(-nz^2/2).$$

The term  $\sqrt{\frac{2\log(1/\delta)}{N_{i,t-1}}}$ , is an at least  $(1 - \delta)$ -order statistics of  $\mu_i$ . With high probability the UCB term is an overestimate of the unknown mean, if  $N_{i,t-1}$  is a **constant**

$$\mathbb{P}(\mu_i \geq \hat{\mu}_{i,t-1} + \sqrt{\frac{2\log(1/\delta)}{N_{i,t-1}}}) \leq \delta.$$

While  $N_{i,t-1}$  is also a random variable that **is not independent of**  $\hat{\mu}_{i,t-1}$ , the claim holds up to constant factors (Exercise 7.1 on the book).



# The Regret of UCB Algorithms

The UCB Algorithm explores all arms exactly once and then estimates each arm using the (sample-mean based) **upper bound of its  $\delta$ -confidence interval**.

Intuitively, the arm chosen in round  $t$  either

- Has a large sample mean,
- Remain underexplored compared to other arms.



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen



# The Regret of UCB Algorithms

The key ingredient lies in **choosing a good confidence level  $\delta$** , which again balances the trade-off between exploration and exploitation.

## Theorem

*Assume the rewards of arms are 1-sub-Gaussian. Let  $\delta = T^{-2}$ . The regret under UCB is at most*

$$\bar{R}_T \leq 3 \sum_{i=1}^m \Delta_i + \sum_{i:\Delta_i>0} \frac{16 \log T}{\Delta_i}.$$

UCB does not require knowledge on the suboptimality gaps.



香港中文大學(深圳)  
The Chinese University of Hong Kong, Shenzhen

# The Regret of UCB Algorithms

The UCB Theorem may seem loose when  $\Delta_i$  are small. This can be fixed by separating the arms into two parts: those with a sub-optimality gap less than  $\sqrt{16m \log T / T}$  and greater than  $\sqrt{16m \log T / T}$ . Bounding  $\mathbb{E}[N_{i,T}]$  by  $T$  in the first part and by the UCB Theorem in the second part gives

$$\bar{R}_T \leq 3 \sum_{i \in [m]} \Delta_i + 8\sqrt{mT \log T}. \quad (1)$$



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# The Regret of UCB Algorithms

There are a few things we could consider for extension.

- The confidence level in UCB Theorem depends on horizon  $T$ . This can be removed by choosing  $\delta$  in a decreasing format, say,  $\delta_t = (1 + t \log^2 t)^{-1}$ .
- The Hoeffding inequality used in the algorithm can be rather loose sometimes. For example, consider the Bernoulli bandits whose means are close to 0 or 1. In such situations, one could apply the Chernoff bound instead, which gives a confidence interval based on relative entropy.



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen

# Question and Answering (Q&A)



香港中文大學(深圳)  
The Chinese University of Hong Kong, Shenzhen