Lecture 19: Multi-Armed Bandit

Course: Reinforcement Learning Theory
Instructor: Lei Ying
Department of EECS
University of Michigan, Ann Arbor

Optimality of UCB

Suppose that none of the events occurs (i.e. all three equations are false). Then we have

$$\mu_1(t-1) + \sqrt{\frac{\alpha \log(t)}{N_1(t-1)}} > \mu_1 \qquad \text{(condition 1)}$$

$$\begin{aligned} \mu_1 &= \mu_i + \Delta_i \\ &\geq \mu_i + 2\sqrt{\frac{\alpha \log(t)}{N_i(t-1)}} && \text{(condition 3)} \\ &\geq \mu_1(t-1) + \sqrt{\frac{\alpha \log(t)}{N_i(t-1)}} && \text{(condition 2)} \end{aligned}$$

But then the algorithm should have picked arm 1 over arm i (contradiction).

Optimality of UCB

$\Longrightarrow$ {arm i played at t AND (3) false} $\subseteq$ {(1) true OR (2) true}

If arm i played at time t and (3) false, then (1) true or (2) true.

$$E[N_i(T)] = E\left[\sum_{t=1}^{T} \mathbb{I}\{\text{arm } i \text{ is played at time } t\}\right]$$

Define $u = \left\lceil \frac{4\alpha \log(T)}{\Delta_i^2} \right\rceil$, then

$$E[N_i(T)] \leq u + E\left[\sum_{t=u+1}^{T} \mathbb{I}\{\text{arm } i \text{ is played at time } t \text{ AND } N_i(t-1) \geq u\}\right]$$

Example

Consider each realization, 1 indicates arm i is played and 0 otherwise.

$$0\ 1\ 0\ 0\ 1\ 0\ 0\ 1\ 0\ 1\ 1\ 0$$

Example

Consider each realization, 1 indicates arm i is played and 0 otherwise.

$$0\ 1\ 0\ 0\ 1\ 0\ 0\ 1\ 0\ 1\ 1\ 0$$

Mark first u 1's to be blue ($u = 2$).

$$E[N_i(T)] \leq u + E[\sum_{t=u+1}^{T} \mathbb{I}\{\text{arm } i \text{ is played \& } N_i(t-1) \geq u\}]$$

Marked 1's

Unmarked 1 requires (3) false $\iff N_i(t-1) \geq u$

Remark: "$\leq$" because it is possible less than u times.

Optimality of UCB

$$E[N_i(T)] \leq u + E\left[\sum_{t=u+1}^{T} \mathbb{I}\{\text{arm } i \text{ is played } \&\ N_i(t-1) \geq u\}\right]$$

$$= u + \sum_{t=u+1}^{T} \Pr(\text{arm } i \text{ played } \&\ (3) \text{ false})$$

$$\leq u + \sum_{t=u+1}^{T} \Pr((1) \text{ true or } (2) \text{ true}),$$

where the last inequality holds because

$$N_i(t-1) \geq u = \left\lceil \frac{4\alpha \log(t)}{\Delta_i^2} \right\rceil \implies (3) \text{ false}$$

Thus,

$$E[N_i(T)] \leq u + \sum_{t=u+1}^{T} \Pr((1) \text{ true}) + \Pr((2) \text{ true})$$

Optimality of UCB

We now focus on the probability of (1) being true, i.e.

$$\Pr\left(\mu_1(t-1) + \sqrt{\frac{\alpha \log t}{N_1(t-1)}} \leq \mu_1\right).$$

The difficulty of analyzing the probability above, e.g. applying Azuma-Hoeffding's inequality is $N_1(t-1)$ is a random variable and is correlated with $\mu_1(t-1)$.

Difficulty

We need to somehow decouple the randomness of generating the rewards from the algorithm itself.

Azuma-Hoeffding Inequality

Consider Bernoulli random variables $X_i \in \{0, 1\}$. Define

$$\mu_i(t) = \frac{\sum_{s=1}^{t} X_i(s)}{t}$$

Imagine that we are estimating μ_i by pulling arm i t times. Define $E[X_i] = \mu_i$.
Azuma–Hoeffding inequality for Bernoulli random variables:

$$\Pr(\mu_i - \mu_i(t) > \epsilon) \leq e^{-\frac{t\epsilon^2}{2\mu_i}} \leq e^{-\frac{t\epsilon^2}{2}}.$$

Optimality of UCB: Example

Consider two arms and $t = 10$,

Generate 10 samples from Bernoulli (μ_1):

$$X_1(1), X_1(2), \ldots, X_1(10)$$

10 samples from Bernoulli (μ_2):

$$X_2(1), X_2(2), \ldots, X_2(10)$$

When arm i is played for the kth time, the reward is $X_i(k)$.

Note that $\{X_i(k)\}_{i=1,2,\, k=1,\ldots,10}$ are generated once at the beginning, so are not decoupled from the MAB algorithm.

Optimality of UCB: Example

When $\left\{\mu_1(t-1) + \sqrt{\frac{\alpha \log(t)}{N_1(t-1)}} \leq \mu_1\right\}$ occurs, it means that when we take the first $N_1(t-1)(\leq t-1)$ values of $X_1(k)$,

$$\frac{1}{N_1(t-1)} \sum_{k=1}^{N_1(t-1)} X_1(k) + \sqrt{\frac{\alpha \log(t)}{N_1(t-1)}} \leq \mu_1.$$

A Necessary Condition

There exists $s \leq t-1$ such that

$$\frac{1}{s} \sum_{k=1}^{s} X_1(k) + \sqrt{\frac{\alpha \log(t)}{s}} \leq \mu_1. \tag{1}$$

Optimality of UCB

Note that for a fixed s, we can apply concentration inequality such as the Azuma-Hoeffding inequality to understand the probability of (1).

$$
\begin{aligned}
&\Pr\left(\mu_1(t-1)+\sqrt{\frac{\alpha\log(t)}{N_1(t-1)}}\le\mu_1\right)\\
\le&\Pr\left(\exists s,\ \frac{1}{s}\sum_{k=1}^{s}X_1(k)+\sqrt{\frac{\alpha\log(t)}{s}}\le\mu_1\right)\\
\le&\sum_{s=1}^{t-1}\Pr\left(\frac{1}{s}\sum_{k=1}^{s}X_1(k)+\sqrt{\frac{\alpha\log(t)}{s}}\le\mu_1\right)\\
<&\sum_{s=1}^{t-1}e^{-s(\frac{\alpha\log(t)}{s})\times\frac{1}{2}}=\sum_{s=1}^{t-1}e^{-\frac{1}{2}\alpha\log(t)}\\
=&\sum_{s=1}^{t-1}\frac{1}{t^{\alpha/2}}=(t-1)\times\frac{1}{t^{\alpha/2}}\le\frac{1}{t^{\frac{\alpha}{2}-1}}.
\end{aligned}
$$

Optimality of UCB

We analyze $P((1) \text{ true})$ for given t, and a similar analysis can be done for $P((2) \text{ true})$ for given t. We then obtain for some $b > 0$ and $\alpha > 4$,

$$
\begin{aligned}
E[N_i(T)] &\leq u + \sum_{t=u+1}^{T} \Pr((1) \text{ true}) + \Pr((2) \text{ true}) \\
&\leq \left\lceil \frac{4\alpha \log(T)}{\Delta_i^2} \right\rceil + \sum_{t=u+1}^{T} \frac{b}{t^{\frac{\alpha}{2}-1}} \\
&\leq \left\lceil \frac{4\alpha \log(T)}{\Delta_i^2} \right\rceil + \int_{u+1}^{\infty} \frac{b}{\tau^{\frac{\alpha}{2}-1}} \, d\tau \\
&= \left\lceil \frac{4\alpha \log(T)}{\Delta_i^2} \right\rceil + \frac{b}{\frac{\alpha}{2}-2} \frac{1}{(u+1)^{\frac{\alpha}{2}-2}} \\
&\leq \left\lceil \frac{4\alpha \log(T)}{\Delta_i^2} \right\rceil + o(1).
\end{aligned}
$$

Thus we have derived a lower bound for the regret,

$$
\begin{aligned}
R_T \leq & \sum_{i=2}^{K} \Delta_i \left\lceil \frac{4\alpha \log(T)}{\Delta_i^2} \right\rceil + o(1) \\
\leq & 4\alpha \left(\sum_{i=2}^{K} \frac{1}{\Delta_i} \right) \log(T) + O(1).
\end{aligned}
$$

Thompson Sampling (Thompson '33)

- Start with prior over parameters

  Example: Bernoulli(μ_i) with prior $\mu_i \sim \text{Beta}(\alpha_i(0), \beta_i(0))$
- Note:

$$X \sim \text{Beta}(\alpha, \beta)$$

$$E[X] = \frac{\alpha}{\alpha + \beta}$$

  X concentrates as $\alpha + \beta$ increases.

Thompson Sampling (Thompson '33)

- Sample $\mu_i(t)$ from distribution Beta$(\alpha_i(t), \beta_i(t))$ for each arm.
- Select arm $i(t)$ such that $i(t) \in \arg\max_i \mu_i(t)$
- Observe the reward, and update distribution of $\mu_i(t)$

  If the reward is 1, $\alpha_i(t+1) = \alpha_i(t) + 1$
  $\beta_i(t+1) = \beta_i(t)$

  If the reward is 0, $\alpha_i(t+1) = \alpha_i(t)$
  $\beta_i(t+1) = \beta_i(t) + 1$

Regret under Thompson sampling: $O(\log T)$.

Reference

- Chapter 2.2 of Bubeck, Sébastien, and Nicolo Cesa-Bianchi. "Regret analysis of stochastic and nonstochastic multi-armed bandit problems." Foundations and Trends® in Machine Learning 5, no. 1 (2012): 1-122.

Acknowledgements: I would like to thank Alex Zhao for helping prepare the slides, and Honghao Wei and Zixian Yang for correcting typos/mistakes.