Lecture 18: Multi-Armed Bandit

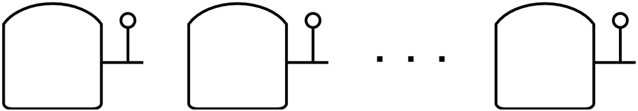
Course: Reinforcement Learning Theory
Instructor: Lei Ying
Department of EECS
University of Michigan, Ann Arbor

Exploration vs. Exploitation

- Exploration: Try actions to find the best one.
- Exploitation: Take the action "believed" to be the best.
- Example: ϵ-greedy

Multi-armed bandit

Stochastic bandits with k possible arms



- Reward: $X_i(t)$ when arm i is played at time t. $X_i(t)$ is i.i.d random variable (for given i).
- Goal: $\max \sum_{t=1}^{T} E[X_{i_t}(t)]$
- i_t: arm played at time t.

Multi-armed bandit problem

Model-based versus model-free

$\mu_i = E[X_i]$ and assume $\mu_1 > \mu_2 > \dots$.

Model-based approach

Since $V_k^* = \mu_1(T-k)$, we solve the following problem

$$\max_i E[X_i(k) + V_{k+1}^*]$$

$$\implies i^* = 1 \text{ (trivial)}.$$

Multi-armed bandit problem

What happens when μ_i are unknown?

Multi-armed bandit problem

Exploration vs. Exploitation

- Exploration: Try different arms to find the best one.
- Exploitation: Use the arm "believed" to be the best.
- Example: ϵ-greedy

Exploration vs. Exploitation

Exploitation: Greedy-algorithm

- At each step, estimate the return of each arm,

$$\mu_i(t) = \frac{\text{total reward from playing arm } i}{\text{number of times arm } i \text{ is played}}$$

- Choose i^* at time t such that

$$i^* \in \arg\max_i \mu_i(t)$$

- No exploration

Exploration vs. Exploitation

Example:

arm 1: $X_1 \in \{0, 1\}, \mu_1 = 0.9$ arm 2: $X_2 \in \{0, 1\}, \mu_2 = 0.5$

Exploration vs. Exploitation

Example:

arm 1: $X_1 \in \{0, 1\}, \mu_1 = 0.9$ arm 2: $X_2 \in \{0, 1\}, \mu_2 = 0.5$

- At time 1: arm 1, $X_1(0) = 0$
- At time 2: arm 2, $X_2(1) = 1$
- At time 3: $\mu_1(3) = \frac{0}{1} = 0$, $\mu_2(3) = \frac{1}{1} = 1$

Exploration vs. Exploitation

Example:

arm 1: $X_1 \in \{0,1\}, \mu_1 = 0.9$ arm 2: $X_2 \in \{0,1\}, \mu_2 = 0.5$

- At time 1: arm 1, $X_1(0) = 0$
- At time 2: arm 2, $X_2(1) = 1$
- At time 3: $\mu_1(3) = \frac{0}{1} = 0$, $\mu_2(3) = \frac{1}{1} = 1$

$\implies i^* = 2$

$\implies$ play arm 2 forever under the greedy policy.

Exploration vs. Exploitation

Example:

arm 1: $X_1 \in \{0, 1\}, \mu_1 = 0.9$ arm 2: $X_2 \in \{0, 1\}, \mu_2 = 0.5$

- At time 1: arm 1, $X_1(0) = 0$
- At time 2: arm 2, $X_2(1) = 1$
- At time 3: $\mu_1(3) = \frac{0}{1} = 0$, $\mu_2(3) = \frac{1}{1} = 1$

$\implies i^* = 2$

$\implies$ play arm 2 forever under the greedy policy.

- Simple remedy: ϵ-greedy

Exploration vs. Exploitation

ϵ-greedy:

- With probability $(1 - \epsilon)$,

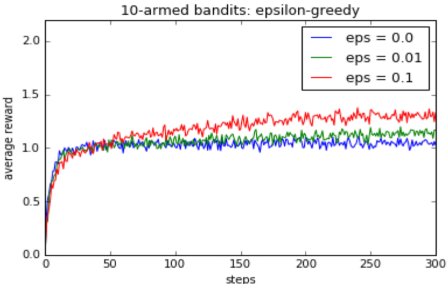
$$i^* \in \arg\max_i \mu_i(t)$$

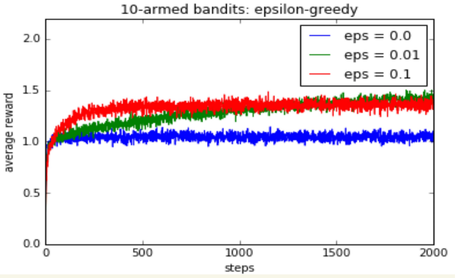
- With probability ϵ, randomly pick an arm.

Exploration vs. Exploitation

$\epsilon \uparrow$: fast exploration but lower long-term return

$\epsilon \downarrow$: slow exploration but higher long-term return





Upper Confidence Bound (UCB) Algorithm

- Play each arm once at the beginning
- At time $t > K$, choose arm i^* such that

$$i^* \in \arg\max_i \mu_i(t) + \sqrt{\frac{\alpha \log(t)}{N_i(t)}}$$

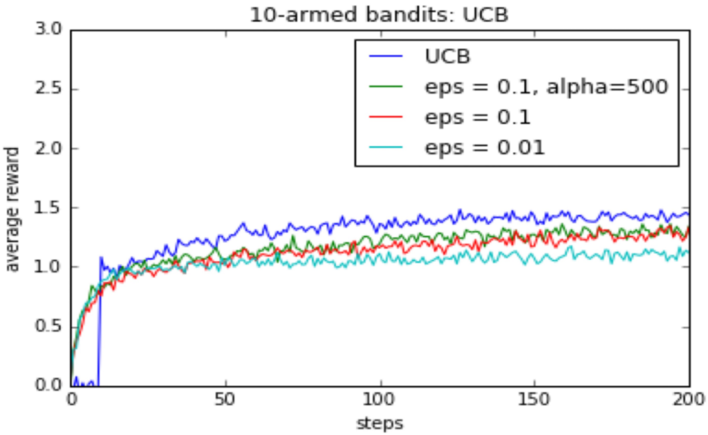
  - $N_i(t)$: number of times arm i played by time t
  - $\mu_i(t) = \frac{\sum_{s=1}^{t} \mathbb{I}_{i(s)=i} X_{i(s)}(s)}{N_i(t)}$
  - $\sqrt{\frac{\alpha \log(t)}{N_i(t)}}$: confidence about estimating μ_i with current observations
    large $N_i(t)$: more confident about arm i
    small $N_i(t)$: less confident

Upper Confidence Bound (UCB) Algorithm

- Explore arms with more uncertain. Optimism in Face of Uncertainty.
- UCB balances exploration and exploitation
- As $t \to \infty$, the probability of selecting the best arm goes to one because

$$\sqrt{\frac{\alpha \log(t)}{N_i(t)}} \to 0, \text{ as } t \to \infty$$

UCB Algorithm



Optimality of UCB

- Regret:

$$R_t = \mu_1 t - E\left[\sum_{s=1}^{t} X_{i_s}(s)\right] \qquad \text{(regret)}$$

- A bound on the regret of the UCB algorithm:[1]

$$R_t \leq \delta\alpha \left(\sum_{i:\Delta_i>0} \frac{1}{\Delta_i}\right) \log t$$

where

$$\Delta_i = \mu_1 - \mu_i.$$

[1] *Bandit Algorithm* by Lattimore and Szepesvari

Optimality of UCB

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

Consider $\alpha > 2$. Suppose at time t, arm i ($i \neq 1$) is played, then one of the following three events must occur:

(1) underestimate μ_1:

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

(2) overestimate μ_i:

$$\mu_i(t-1) > \mu_i + \sqrt{\frac{\alpha \log(t)}{N_i(t-1)}}$$

(3) haven't played arm i enough:

$$N_i(t-1) < \frac{4\alpha \log(t)}{\Delta_i^2} \quad (\text{or } \Delta_i \leq 2\sqrt{\frac{\alpha \log(t)}{N_i(t-1)}})$$

Optimality of UCB

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

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

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

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

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.
- Simulation figures are from the lecture notes available at `https://www.cs.princeton.edu/courses/archive/fall16/cos402/lectures/402-lec22.pdf`

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