

# Exploration vs. Exploitation <sup>1</sup>

## Exploration vs. Exploitation

Exploration: Try actions to find the best one.

## Exploration vs. Exploitation

Exploration: Try actions to find the best one.

Exploitation: Take the action “believed” to be the best.

## Exploration vs. Exploitation

Exploration: Try actions to find the best one.

Exploitation: Take the action “believed” to be the best.

A straightforward strategy:  $\epsilon$ -greedy.

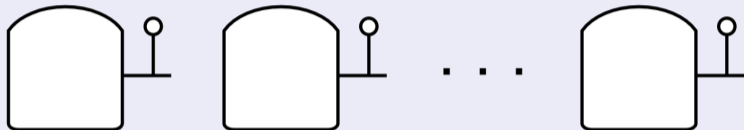
### Question

What is the optimal exploration and exploitation tradeoff?

# Multi-Armed Bandit

## Model

Stochastic bandits with  $I$  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:  $\max \sum_{k=1}^K E[X_{i_k}(k)]$
- $i_k$ : arm played at time step  $k$ .

## $\epsilon$ -Greedy Exploration

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

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

where

$$\mu_i(k) = \frac{\text{total reward received from playing arm } i}{\text{number of times arm } i \text{ has been played}}.$$

- With probability  $\epsilon$ , randomly pick an arm.

# Upper Confidence Bound (UCB) Algorithm

## UCB Exploration

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

$$i_k^* \in \arg \max_i \mu_i(k) + \sqrt{\frac{4 \log K}{N_i(k)}}$$

- ▶  $N_i(k)$ : number of times arm  $i$  played by time step  $k$ ;
- ▶  $\mu_i(k) = \frac{\sum_{\tau=1}^k \mathbb{I}_{i_\tau=i} X_{i_\tau}(\tau)}{N_i(k)}$
- ▶  $\sqrt{\frac{4 \log K}{N_i(k)}}$ : confidence interval about estimating  $\mu_i$  with current observations

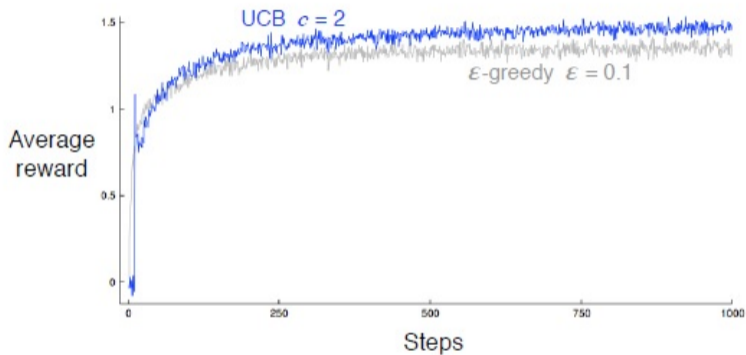
Large  $N_i(s)$  is more confident about arm  $i$  and small  $N_i(s)$  is less confident.

## Upper Confidence Bound (UCB) Algorithm

- Explore arms with more uncertain. Optimism in Face of Uncertainty.
- UCB balances exploration and exploitation
- For large  $K$ , as  $k \rightarrow K \rightarrow \infty$ , the probability of selecting the best arm goes to one because

$$\sqrt{\frac{4 \log K}{N_i(k)}} \rightarrow 0, \text{ as } k \rightarrow K \rightarrow \infty$$

# UCB Algorithm



**Figure:** UCB versus  $\epsilon$ -greedy (Figure 2.3 in Sutton and Barto)

# Regret Analysis of UCB

## Definition: Regret

Assume arm 1 is the best arm.

$$R_K = \mu_1 K - E \left[ \sum_{k=1}^K X_{i_k}(k) \right].$$

## Regret Bound

Under UCB,

$$R_K = O\left(\sqrt{K \log K}\right).$$

Under  $\epsilon$ -greedy,

$$R_K = O(K).$$

### Basic Definition: Martingale

A discrete-time stochastic process  $Y_1, Y_2, Y_3, \dots$ , that satisfies for any  $k$

$$\mathbb{E}[|Y_k|] < \infty$$

$$\mathbb{E}[Y_{k+1} | Y_k, Y_{k-1}, \dots, Y_1] = Y_k.$$

### Azuma-Hoeffding Inequality

Consider a martingale  $\{Y_k\}$  such that  $Y_0 = 0$  and  $|Y_k - Y_{k-1}| \leq c_k$ . For any  $m > 0$ ,

$$P(|Y_k| \geq m) \leq 2 \exp\left(-\frac{m^2}{2 \sum_{\tau=1}^k c_\tau^2}\right).$$

# Regret Analysis

## Proof

$$R_K = \mu_1 K - E \left[ \sum_{k=1}^K X_{i_k}(k) \right]$$

# Regret Analysis

## Proof

$$\begin{aligned} R_K &= \mu_1 K - E \left[ \sum_{k=1}^K X_{i_k}(k) \right] \\ &= \mu_1 K - \sum_{k=1}^K E[\mu_{i_k}] \end{aligned}$$

# Regret Analysis

## Proof

$$\begin{aligned}R_K &= \mu_1 K - E \left[ \sum_{k=1}^K X_{i_k}(k) \right] \\&= \mu_1 K - \sum_{k=1}^K E[\mu_{i_k}] \\&= \sum_{k=1}^K E \left[ \mu_1 - \mu_{i_k}(k) - \sqrt{\frac{4 \log K}{N_{i_k}(k)}} \right] \\&\quad + \sum_{k=1}^K E \left[ \mu_{i_k}(k) + \sqrt{\frac{4 \log K}{N_{i_k}(k)}} - \mu_{i_k} \right]\end{aligned}$$

# Regret Analysis

## Azuma-Hoeffding Inequality

Consider a martingale  $\{Y_k\}$  such that  $Y_0 = 0$  and  $|Y_k - Y_{k-1}| \leq c_k$ . For any  $m > 0$ ,

$$P(|Y_k| \geq m) \leq 2 \exp\left(-\frac{m^2}{2 \sum_{\tau=1}^k c_\tau^2}\right).$$

## Application of the Azuma-Hoeffding inequality

Define  $\mu_{i,n} = \frac{\sum_{\tau=1}^n X_{i,\tau}}{n}$  (the estimate after getting  $n$  samples). Then

$$Y_n = n\mu_{i,n} - n\mu_i = \sum_{\tau=1}^n X_{i,\tau} - n\mu_i$$

is a martingale.

# Regret Analysis

## Application of the Azuma-Hoeffding inequality

Define  $\mu_{i,n} = \frac{\sum_{\tau=1}^n X_{i,\tau}}{n}$  (the estimate after getting  $n$  samples). Then  $Y_n = n\mu_{i,n} - n\mu_i = \sum_{\tau=1}^n X_{i,\tau} - n\mu_i$  is a martingale.

$$E[Y_{n+1} | Y_1, \dots, Y_n] = E[Y_n + X_{i,n+1} - \mu_i | Y_n] = Y_n.$$

and

$$|Y_{n+1} - Y_n| = |X_{i,n+1} - \mu_i| \leq 1 = c_{n+1}.$$

# Regret Analysis

## Azuma-Hoeffding Inequality

Consider a martingale  $\{Y_k\}$  such that  $Y_0 = 0$  and  $|Y_k - Y_{k-1}| \leq c_k$ . For any  $m > 0$ ,

$$P(|Y_k| \geq m) \leq 2 \exp\left(-\frac{m^2}{2 \sum_{\tau=1}^k c_\tau^2}\right).$$

## Application of the Azuma-Hoeffding inequality

Define event  $\mathcal{A}$  such that

$$\mathcal{A} = \left\{ |\mu_{i,n} - \mu_i| \leq \sqrt{\frac{4 \log K}{n}} \quad \forall i, \forall n = 1, \dots, K \right\}.$$

Applying the Azuma-Hoeffding inequality and the union bound, we obtain

$$P(\mathcal{A}) \geq 1 - l \times K \times \frac{2}{K^2} = 1 - \frac{2l}{K}.$$

# Regret Analysis

$$F_1(k) = \mu_1 - \mu_{i_k}(k) - \sqrt{\frac{4 \log K}{N_{i_k}(k)}} \quad \text{and} \quad F_2(k) = \mu_{i_k}(k) + \sqrt{\frac{4 \log K}{N_{i_k}(k)}} - \mu_{i_k}.$$

## Fact 1

$$F_1(k) \leq 1 \quad \text{and} \quad F_2(k) \leq 1 + 2\sqrt{\log K}.$$

## Fact 2

When  $\mathcal{A}$  occurs,

$$F_1(k) \leq 0 \quad \text{and} \quad |F_2(k)| \leq 2 \sqrt{\frac{4 \log K}{N_{i_k}(k)}}.$$

# Regret Analysis

Proof (continued)

$$\begin{aligned} R_K &= \sum_{k=1}^K E[F_1(k) | \mathcal{A}] \Pr(\mathcal{A}) + \sum_{k=1}^K E[F_1(k) | \mathcal{A}^c] \Pr(\mathcal{A}^c) \\ &\quad + \sum_{k=1}^K E[F_2(k) | \mathcal{A}] \Pr(\mathcal{A}) + \sum_{k=1}^K E[F_2(k) | \mathcal{A}^c] \Pr(\mathcal{A}^c) \end{aligned}$$

# Regret Analysis

## Proof (continued)

$$\begin{aligned} R_K &= \sum_{k=1}^K E[F_1(k) | \mathcal{A}] \Pr(\mathcal{A}) + \sum_{k=1}^K E[F_1(k) | \mathcal{A}^c] \Pr(\mathcal{A}^c) \\ &+ \sum_{k=1}^K E[F_2(k) | \mathcal{A}] \Pr(\mathcal{A}) + \sum_{k=1}^K E[F_2(k) | \mathcal{A}^c] \Pr(\mathcal{A}^c) \\ &\leq 0 + K \Pr(\mathcal{A}^c) + \sum_{k=1}^K \mathbb{E} \left[ 2 \sqrt{\frac{4 \log K}{N_{i_k}(k)}} \middle| \mathcal{A} \right] \Pr(\mathcal{A}) + K \Pr(\mathcal{A}^c) (1 + 2\sqrt{\log K}) \end{aligned}$$

,

# Regret Analysis

## Proof (continued)

From the Azuma-Hoeffding inequality, we have  $\Pr(\mathcal{A}^c) \leq \frac{2l}{K}$ . Applying this bound and the bound  $\Pr(\mathcal{A}) \leq 1$ , we have

$$\begin{aligned} R_K &\leq 4\sqrt{\log K} \mathbb{E} \left[ \sum_{k=1}^K \sqrt{\frac{1}{N_{i_k}(k)}} \middle| \mathcal{A} \right] + 2l(1 + 2\sqrt{\log K}) \\ &\leq 4l\sqrt{\log K} \sum_{k=1}^K \sqrt{\frac{1}{k}} + 4l + 4l\sqrt{\log K} = O(\sqrt{K \log K}) \end{aligned}$$

# Optimality of UCB

## Instance Dependent $\log K$ Regret

With a more complicated analysis, we have

$$R_K \leq \left( \sum_{i=2}^I \frac{16}{\Delta_i} \right) \log K + o(1),$$

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

## Reference

- 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.
- Lattimore T, Szepesvári C. *Bandit algorithms*. Cambridge University Press; 2020 Jul 16.