Exploration vs. Exploitation [1]

[1] Chapter 8: Yang&Ying

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: ϵ-greedy.

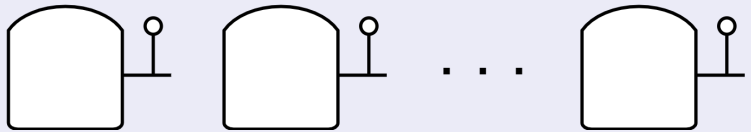
Question

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. $X_i(t) \in [0,1]$ 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.

ϵ-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 ϵ, 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 internal about estimating μ_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 $k \to K \to \infty$, the probability of selecting the best arm goes to one because

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

UCB Algorithm

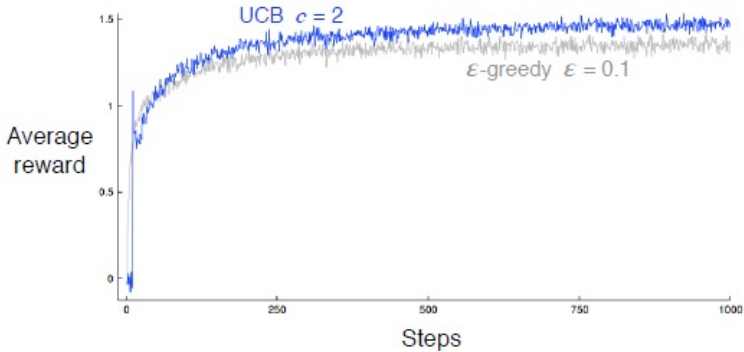


Figure: UCB versus ϵ-greedy (Figure 2.3 in Sutton and Barto)

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 ϵ-greedy,

$$R_K = O(K).$$

Basic Definition: Martingale

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

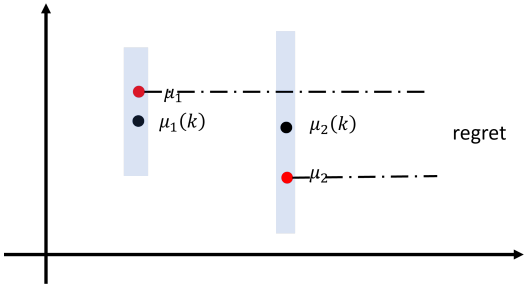
$$\mathbb{E}[|Y_k|] < \infty$$
$$\mathbb{E}[Y_{k+1}|Y_k, Y_{k-1}, \cdots, Y_1] = Y_k.$$

Azuma-Hoeffding Inequality

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

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

UCB Algorithm: Regret bound



Intuition: Regret at time k is bounded by $2\sqrt{\frac{4 \log K}{N_2(k)}}$. Since $N_2(k) \leq k$, the total regret is bounded by

$$\sum_{k=1}^{K} 2\sqrt{\frac{4 \log K}{N_2(k)}} \leq \sum_{k=1}^{K} 2\sqrt{\frac{4 \log K}{k}} = O(\sqrt{K \log K})$$

Regret Analysis

Proof

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

Regret Analysis

Proof

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

Regret Analysis

Proof

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

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.

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

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 $\{Y_n\}$ such that $Y_0 = 0$ and $|Y_n - Y_{n-1}| \leq c_n$. For any $m > 0$,

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

Application of the Azuma-Hoeffding inequality

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

$$P\left(|\mu_{i,n} - \mu_i| \geq \sqrt{\frac{4\log K}{n}}\right) = P\left(|Y_n| \geq \sqrt{4n\log K}\right) \leq 2\exp\left(-\frac{4n\log K}{2n}\right) = \frac{2}{K^2}.$$

Regret Analysis

Azuma-Hoeffding Inequality

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

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

Application of the Azuma-Hoeffding inequality

Define event $\mathcal{A}$ such that

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

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

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

Regret Analysis

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

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,

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

Regret Analysis

Fact 2 holds because arm i_k was selected at step k.

Proof (continued)

$$R_K = \sum_{k=1}^{K} E\left[F_1(k) | \mathcal{A}\right] \Pr(\mathcal{A}) + \sum_{k=1}^{K} E\left[F_1(k) | \mathcal{A}^c\right] \Pr(\mathcal{A}^c)$$
$$+ \sum_{k=1}^{K} E\left[F_2(k) | \mathcal{A}\right] \Pr(\mathcal{A}) + \sum_{k=1}^{K} E\left[F_2(k) | \mathcal{A}^c\right] \Pr(\mathcal{A}^c)$$

Regret Analysis

Fact 2 holds because arm i_k was selected at step k.

Proof (continued)

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

Regret Analysis

Proof (continued)

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

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

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 $\Delta_i = \mu_1 - \mu_i$.

Thompson Sampling (1933)

Thompson Sampling

Assume a uniform prior on $\mu_i \in [0,1]$ and Bernoulli reward $X_i \in \{0,1\}$. Let $\pi_{i,k}$ be the posterior distribution for μ_i at the kth step.

- Sample $\mu_{i,k}$ according to distribution $\pi_{i,k}$
- Pull arm $i_k \in \arg\max_i \mu_{i,k}$
- Update $\pi_{i_k,k+1}$ based on $X_{i_k}(k)$.

Thompson Sampling (1933)

Thompson Sampling

Assume a uniform prior on $\mu_i \in [0,1]$ and Bernoulli reward $X_i \in \{0,1\}$. Let $\pi_{i,k}$ be the posterior distribution for μ_i at the kth step.

- Sample $\mu_{i,k}$ according to distribution $\pi_{i,k}$
- Pull arm $i_k \in \arg\max_i \mu_{i,k}$
- Update $\pi_{i_k,k+1}$ based on $X_{i_k}(k)$.

$\pi_{i,k}$ is a Beta distribution Beta(α, β) with mean $\frac{\alpha}{\alpha+\beta}$. If arm k is pulled, α increases by one if $X_i(k) = 1$ and β increases by one if $X_i(k) = 0$.

Thompson Sampling (1933)

Thompson Sampling

Assume a uniform prior on $\mu_i \in [0,1]$ and Bernoulli reward $X_i \in \{0,1\}$. Let $\pi_{i,k}$ be the posterior distribution for μ_i at the kth step.

- Sample $\mu_{i,k}$ according to distribution $\pi_{i,k}$
- Pull arm $i_k \in \arg\max_i \mu_{i,k}$
- Update $\pi_{i_k,k+1}$ based on $X_{i_k}(k)$.

$\pi_{i,k}$ is a Beta distribution Beta(α, β) with mean $\frac{\alpha}{\alpha+\beta}$. If arm k is pulled, α increases by one if $X_i(k) = 1$ and β increases by one if $X_i(k) = 0$.

Same (order) Regret bounds but better empirical performance.

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.
- W. Thompson. *On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.* Bulletin of the American Mathematics Society, 25:285–294, 1933.