UCB

1 Introduction

In the Multi-Armed Bandit problem, a recurring idea in the analysis of algorithms such as UCB and ε-greedy is the following:

If a given action (arm) is selected with a sufficiently large probability at each time step, then it will be selected sufficiently often in total.

The following lemma provides a probabilistic and quantitative formulation of this idea. Importantly, the selection process does not need to be independent; the result also holds for adaptive selection rules that depend on past observations.

Lemma 1 (Lower bound on the number of selections)**.** *Let $(B_t)_{t=1}^T$ be a sequence of $\{0,1\}$-valued random variables such that*

$$(B_t)_{t=1}^T \text{ is adapted to a filtration } (\mathcal{F}_t)_{t\geq 0},$$

and

$$\mathbb{E}[B_t \mid \mathcal{F}_{t-1}] = p_t, \qquad p_t \in (0,1].$$

Define

$$p_{\min} := \min_{t\leq T} p_t > 0.$$

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\forall t \geq \frac{8}{p_{\min}} \log \frac{T}{\delta}, \qquad \sum_{\tau=1}^{t} B_\tau \;\geq\; \frac{1}{2}\sum_{\tau=1}^{t} p_\tau.$$

Proof. Define the cumulative number of selections and its conditional expectation by

$$S_t := \sum_{\tau=1}^{t} B_\tau, \qquad \mu_t := \sum_{\tau=1}^{t} p_\tau.$$

We define the bad event

$$\mathcal{E}_t := \left\{ S_t < \tfrac{1}{2}\mu_t \right\}.$$

Although the variables (B_t) are not independent, the process

$$M_t := S_t - \mu_t$$

is a martingale with respect to $(\mathcal{F}_t)$, with bounded increments $|M_t - M_{t-1}| \leq 1$.

By standard martingale concentration inequalities (e.g. Azuma–Hoeffding or Freedman's inequality), for any $0 < \alpha < 1$,

$$\mathbb{P}(S_t \leq (1-\alpha)\mu_t) \leq \exp\left(-\frac{\alpha^2}{2}\mu_t\right).$$

Setting $\alpha = \frac{1}{2}$ yields

$$\mathbb{P}(\mathcal{E}_t) \leq \exp\left(-\tfrac{1}{8}\mu_t\right).$$

Since $p_\tau \geq p_{\min}$ for all τ,

$$\mu_t = \sum_{\tau=1}^{t} p_\tau \geq t\, p_{\min}.$$

Therefore, for each $t \geq 1$,

$$\mathbb{P}(\mathcal{E}_t) \leq \exp\left(-\tfrac{1}{8}\, t p_{\min}\right).$$

We aim to ensure that

$$\mathbb{P}(\exists t \in [t_0, T] : \mathcal{E}_t) \leq \delta.$$

By a union bound, it suffices to require

$$\mathbb{P}(\mathcal{E}_t) \leq \frac{\delta}{T} \quad \text{for all } t \geq t_0.$$

Using the above tail bound, this condition is satisfied whenever

$$\exp\left(-\tfrac{1}{8}\, t_0 p_{\min}\right) \leq \frac{\delta}{T}.$$

Taking logarithms on both sides yields

$$t_0 \geq \frac{8}{p_{\min}} \log \frac{T}{\delta}.$$

We therefore choose

$$t_0 = \left\lceil \frac{8}{p_{\min}} \log \frac{T}{\delta} \right\rceil.$$

Applying a union bound, we obtain

$$\mathbb{P}(\exists t \in [t_0, T] : \mathcal{E}_t) \leq \sum_{t=t_0}^{T} \mathbb{P}(\mathcal{E}_t) \leq (T - t_0 + 1)\frac{\delta}{T} \leq \delta.$$

Hence, with probability at least $1 - \delta$,

$$\forall t \geq t_0, \qquad S_t \geq \tfrac{1}{2}\mu_t.$$

□

Corollary 1. *We now apply the lemma to the bandit setting. Let*

$$B_\tau = \mathbf{1}(a_\tau = i)$$

be the indicator variable that action i is selected at time τ.

Assume that for all τ,

$$\mathbb{P}(a_\tau = i \mid \mathcal{F}_{\tau-1}) \;\geq\; \frac{\varepsilon_\tau}{N}, \qquad \varepsilon_\tau \in (0,1].$$

That is, action i is explored with probability at least ε_τ/N at each time step.

Define

$$p_{\min}^{(i)} := \min_{\tau \leq T} \frac{\varepsilon_\tau}{N}.$$

Then, with probability at least $1-\delta$,

$$\sum_{\tau=1}^{t} \mathbf{1}(a_\tau = i) \;\geq\; \frac{1}{2N} \sum_{\tau=1}^{t} \varepsilon_\tau, \qquad \forall t \geq \frac{8}{p_{\min}^{(i)}} \log \frac{T}{\delta}.$$

2 Proof of the Chernoff Bound

This section provides a self-contained proof of the standard Chernoff bound for the independent case. While the lemma above relies on martingale concentration inequalities, the independent Chernoff bound is included here for completeness and intuition.

Let

$$X = \sum_{i=1}^{n} X_i,$$

where

- $X_i \in \{0,1\}$,
- $X_1, \ldots, X_n$ are independent,
- $\mathbb{E}[X_i] = p_i$.

Define

$$\mu := \mathbb{E}[X] = \sum_{i=1}^{n} p_i.$$

Our goal is to bound

$$\mathbb{P}(X \leq (1-\delta)\mu)\,, \qquad 0 < \delta < 1.$$

Step 1: Exponential Transformation

For any $\lambda > 0$,

$$\mathbb{P}(X \leq (1-\delta)\mu) = \mathbb{P}\Big(e^{-\lambda X} \geq e^{-\lambda(1-\delta)\mu}\Big)\,.$$

By Markov's inequality,

$$\mathbb{P}(X \leq (1-\delta)\mu) \leq \frac{\mathbb{E}[e^{-\lambda X}]}{e^{-\lambda(1-\delta)\mu}} = e^{\lambda(1-\delta)\mu}\,\mathbb{E}[e^{-\lambda X}].$$

Step 2: Factorization via Independence

Using independence,

$$\mathbb{E}[e^{-\lambda X}] = \mathbb{E}\Big[e^{-\lambda \sum_{i=1}^{n} X_i}\Big] = \prod_{i=1}^{n} \mathbb{E}[e^{-\lambda X_i}].$$

For each Bernoulli random variable X_i,

$$\mathbb{E}[e^{-\lambda X_i}] = (1 - p_i) + p_i e^{-\lambda} = 1 - p_i(1 - e^{-\lambda}).$$

Step 3: Chernoff Inequality

Using the inequality $1 - x \leq e^{-x}$ for all $x \geq 0$,

$$1 - p_i(1 - e^{-\lambda}) \leq \exp\big(-p_i(1 - e^{-\lambda})\big).$$

Thus,

$$\mathbb{E}[e^{-\lambda X}] \leq \exp\left(-(1 - e^{-\lambda}) \sum_{i=1}^{n} p_i\right) = \exp\big(-\mu(1 - e^{-\lambda})\big).$$

Step 4: Combining the Bounds

Substituting into the Markov bound,

$$\mathbb{P}(X \leq (1 - \delta)\mu) \leq \exp\Big(\lambda(1 - \delta)\mu - \mu(1 - e^{-\lambda})\Big).$$

Equivalently,

$$\mathbb{P}(X \leq (1 - \delta)\mu) \leq \exp\Big(\mu\big[\lambda(1 - \delta) - (1 - e^{-\lambda})\big]\Big).$$

Step 5: Optimal Choice of λ

Choose

$$\lambda = -\ln(1 - \delta), \qquad 0 < \delta < 1.$$

Then

$$e^{-\lambda} = 1 - \delta, \qquad \lambda = \ln \frac{1}{1 - \delta}.$$

Substituting,

$$\lambda(1 - \delta) - (1 - e^{-\lambda}) = (1 - \delta) \ln \frac{1}{1 - \delta} - \delta.$$

Step 6: Analytic Inequality

For all $0 < \delta < 1$,

$$(1 - \delta) \ln \frac{1}{1 - \delta} - \delta \leq -\frac{\delta^2}{2}.$$

Conclusion

Therefore,

$$\mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(-\frac{\mu\delta^2}{2}\right).$$

$$\boxed{\mathbb{P}(X \leq (1-\delta)\mu) \leq e^{-\mu\delta^2/2}}$$

Problem Setup

Fix an arm i. Let

$$T_i(t) := \sum_{\tau=1}^{t} \mathbf{1}(a_\tau = i)$$

be the number of times arm i is selected up to time t.

The empirical mean is defined as

$$\hat{\mu}_{i,t} = \frac{1}{T_i(t)} \sum_{\tau=1}^{t} \mathbf{1}(a_\tau = i)\, X_{\tau,i},$$

where $X_{\tau,i} \in [0,1]$ and $\mathbb{E}[X_{\tau,i}] = \mu_i$.

Event A: Selection Count Lower Bound

Assume that with high probability,

$$\mathbb{P}\left(\bigcap_{t \geq \frac{8}{p_{\min}^{(i)}} \log T} \left\{ T_i(t) \geq \frac{1}{2N} \sum_{\tau=1}^{t} \varepsilon_\tau \right\} \right) \;\geq\; 1 - \frac{1}{T}. \tag{A}$$

This event guarantees a deterministic lower bound on the (random) sample size $T_i(t)$ for all sufficiently large t. The threshold follows directly from Lemma 1 by choosing $\delta = 1/T$.

Hoeffding Inequality with Fixed Sample Size

For m conditionally independent samples taking values in $[0,1]$ with common mean μ_i, Hoeffding's inequality states that for any $\delta \in (0,1)$,

$$\mathbb{P}\left(\left|\hat{\mu}_{i,t} - \mu_i\right| \geq \sqrt{\frac{2\log(1/\delta)}{m}} \right) \leq \delta. \tag{H}$$

Random Sample Size Issue

In the bandit setting, the number of samples used to form $\hat{\mu}_{i,t}$ is random:

$$m = T_i(t).$$

Therefore, Hoeffding's inequality cannot be applied directly without additional control on $T_i(t)$.

Conditioning on Event A

On Event (A), for all $t \geq \frac{8}{p_{\min}^{(i)}} \log T$, we have the deterministic lower bound

$$T_i(t) \geq \frac{1}{2N} \sum_{\tau=1}^{t} \varepsilon_\tau. \tag{C}$$

Substituting this bound into (H), we obtain that on Event (A),

$$\left|\hat{\mu}_{i,t} - \mu_i\right| \leq \sqrt{\frac{2\log(1/\delta)}{\frac{1}{2N}\sum_{\tau=1}^{t} \varepsilon_\tau}} = \sqrt{\frac{4N\log(1/\delta)}{\sum_{\tau=1}^{t} \varepsilon_\tau}}.$$

Uniform Control over Time

To ensure that the deviation bound holds simultaneously for all $t \leq T$, we choose $\delta = 1/T^2$ and apply a union bound over time.

Then

$$\log(1/\delta) = 2\log T,$$

and the bound becomes

$$\left|\hat{\mu}_{i,t} - \mu_i\right| \leq \sqrt{\frac{4N\log T}{\sum_{\tau=1}^{t} \varepsilon_\tau}}, \qquad \forall t \geq \frac{8}{p_{\min}^{(i)}} \log T.$$

Event B: Mean Estimation Error Bound

Thus,

$$\mathbb{P}\left(\bigcap_{t \geq \frac{8}{p_{\min}^{(i)}} \log T} \left\{ \left|\hat{\mu}_{i,t} - \mu_i\right| \leq \sqrt{\frac{4N\log T}{\sum_{\tau=1}^{t} \varepsilon_\tau}} \right\} \right) \geq 1 - \frac{1}{T}. \tag{B}$$

Conclusion

Event (B) is obtained by

- first establishing a lower bound on $T_i(t)$ via Event (A),
- conditioning on this event to apply Hoeffding's inequality,
- and applying a union bound over time.

Combining the failure probabilities of Events (A) and (B) via a union bound yields an overall failure probability of order $O(1/T)$.