Multi-armed Bandit Problems with Regret Analysis

Jungbin Jun

Department of Statistics, Inha University

March 23, 2026

1 Pseudo-Regret

Online Player가 s번 round까지 i-th arm을 선택한 횟수를 $T_i(s) := \sum_{i=1}^{s} \mathbb{I}(I_t = i)$로 정의하자. 또한 $\Delta_i = \mu^* - \mu_i$를 i-th arm의 suboptimality parameter라고 하자. 그러면 pseudo-regret은 아래와 같이 표현된다.

$$\begin{aligned}
\bar{R}_n &= \max_{i\in[K]} \mathbb{E}\left[\sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t}\right] = \max_{i\in[K]}\left[n\mu_i - \sum_{t=1}^{n}\mathbb{E}[\mu_{I_t}]\right] \\
&= n\mu^* - \sum_{i=1}^{t}\mathbb{E}[\mu_{I_t}] \\
&= \left(\sum_{j\in[K]}\mathbb{E}T_j(n)\right)\mu^* - \mathbb{E}\sum_{j\in[K]} T_j(n)\mu_j \\
&= \left(\sum_{j\in[K]}\mathbb{E}T_j(n)\right)(\mu^* - \mu_j) \\
&= \sum_{j\in[K]} \Delta_j \mathbb{E}T_j(n).
\end{aligned}$$

2 Upper Confidence Bound (UCB) Strategies

Hoeffding's lemma를 이용하면 유계인 확률변수 X에 대해 다음과 같은 부등식을 이끌어낼 수 있다.

$$\log \mathbb{E}[\lambda \exp(X - E(X))] \leq \phi(\lambda)$$

Reward $X_{i,j}$가 유계로 주어진다고 가정하면 다음을 이끌어낼 수 있다.

$$\begin{aligned}
\mathbb{P}(\mu_i - \hat{\mu}_{i,s} > \epsilon) &= \mathbb{P}\left(\sum_{t=1}^{s}(\mu_i - X_{i,t}) > s\epsilon\right) \\
&= \mathbb{P}\left[\exp\left(\lambda\sum_{t=1}^{s}(\mu_i - X_{i,t})\right) > \exp(\lambda s\epsilon)\right] \\
&\leq \frac{\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{s}(\mu_i - X_{i,t})\right)\right]}{\exp(\lambda s\epsilon)} = \prod_{t=1}^{s}\frac{\mathbb{E}\left[\exp\left(\lambda(\mu_i - X_{i,t})\right)\right]}{\exp(\lambda s\epsilon)} \\
&\leq \frac{\exp(s\phi(\lambda))}{\exp(\lambda s\epsilon)} = \exp(-s(\lambda\epsilon - \phi(\lambda))).
\end{aligned}$$

위 식에서 첫번째 부등식은 Markov 부등식에 의한 것이고, 두번째 부등식은 Hoeffding's lemma에 근거한다. Bound를 최대한 tight하게 잡기 위해서, bound에 다음을 집어넣어서 사용한다.

$$\psi^*(\epsilon) = \sup_{\lambda\in\mathbb{R}} \lambda\epsilon - \psi(\lambda).$$

그러면, 최종적으로 얻어진 bound는 다음과 같다.

$$\mathbb{P}(\mu_i - \hat{\mu}_{i,s} > \epsilon) \leq \exp(-s\psi^*(\epsilon)).$$

이제 이 확률이 δ 이하가 되게 하는 ϵ을 골라보자. 이는 어렵지 않다. Upper bound가 δ 이하가 되도록 통제하면 된다.

$$\exp(-s\psi^*(\epsilon)) \leq \delta \iff \epsilon \geq (\psi^*)^{-1}\left(\frac{1}{s}\log\frac{1}{s}\right)$$

subgaussian과 subexponential과 같은 경우에 대해서는 ψ^*가 monotone increasing한다는 사실이 알려져 있으므로 위의 논증이 가능할 것이다. 그럼 이제, 최종적으로 다음의 사실을 알 수 있다. With probability $1-\delta$,

$$\mathsf{UCB}_i(s) = \hat{\mu}_{i,s} + (\psi^*)^{-1}\left(\frac{1}{s}\log\frac{1}{s}\right) > \mu_i.$$

한편, $\delta > 0$을 상수로 고정하게 되면, 다음의 문제가 발생한다.

$$\mathbb{P}(\cup_{j=1}^{p}\{\mu_i \geq \mathsf{UCB}_i(j)\}) \leq t\delta.$$

따라서, $\delta = t^{-\alpha}$, $\alpha > 0$로 두게 되면 위와 같은 문제를 해결할 수 있고, 그러한 전략을 (α,ψ)-UCB라고 한다.

$$I_t \in \arg\max_{i\in[K]}\left[\hat{\mu}_{i.T_i(t-1)} + (\psi^*)^{-1}\left(\frac{1}{T_i(t-1)}\alpha\log t\right)\right]$$

2.1 Regret Analysis for UCB strategies

Theorem 1 (Pseudo-regret of (α, ψ)-UCB)**.** *Assume that the reward distributions are bounded. Then* (α, ψ)*-UCB with* $\alpha > 2$ *satisfies*

$$\bar{R}_n \leq \sum_{i:\Delta_i>0} \left(\frac{\alpha \Delta_i}{\psi^*(\Delta_i/2)} \log n + \frac{\alpha}{\alpha - 2} \right)$$

위 bound는 UCB가 suboptimal arm을 얼마나 적게 뽑는지, 따라서 총 regret이 얼마나 작게 유지되는지를 보여준다. i번째 arm의 suboptimality gap인 Δ_i가 클 수록 나쁜 arm인데, 위 bound는 직관적으로 각 suboptimal arm i는 대략적으로 $\log n$번만 뽑힘을 보여준다.