Chap. 7 The Upper Confidence Bound Algorithm

- Version 0.1
- The contents are adapted from UIUC IE 498, Chap. 7 in Bandit Algorithm book, and Chap. 6 in reinforcement learning book.

§7.0 Setting

- FSG bandit.

§7.1 Algorithm

Define

$$UCB_i(t) = \begin{cases} \hat{\mu}_{t-1,i} + \sqrt{\dfrac{2 \cdot \ln n \cdot k \cdot \delta^{-1}}{T_i(t-1)}}, & T_i(t-1) \geq 1 \\ \infty, & T_i(t-1) = 0. \end{cases}$$

Algorithm (UCB Algorithm)

- for $t = 1, \dots, n$ do
  - Choose action $A_t = \underset{i \in [k]}{\operatorname{argmax}}\ UCB_i(t)$.
  - Observe reward X_t and update upper confidence bounds.

§7.2 Analysis

Theorem 7.1 [Gap-independent Regret Bound]

It holds that the pseudo-regret of UCB algorithm is upper bounded by

$$O\left(\sqrt{kn \cdot \ln(nk/\delta)} + K\right),$$

w.p. at least $1-\delta$.

Proof.

① We start by proving the optimism. By Hoeffding inequality, it holds that

$$\mathbb{P}\left(\hat{\mu}_{t-1,i} + \sqrt{\frac{2 \cdot \ln(kn/\delta)}{T_i(t-1)}} \geq \mu_i\right) \geq 1 - \frac{\delta}{TK},$$

for some fixed $t \in [n]$. Let

$$E_{t,i} = \left\{ \hat{\mu}_{t-1,i} + \sqrt{\frac{2 \cdot \ln(kn/\delta)}{T_i(t-1)}} \geq \mu_i \right\}.$$

Then

$$\mathbb{P}\left(\exists t \in [n], \exists i \in [k] : \bar{E}_{t,i} \text{ holds}\right) \leq \frac{\delta}{kT} \cdot (KT)$$

$$= \delta,$$

by union bound. Let $\mathcal{E} := \bigcap_{t \in [n]} \bigcap_{i \in [k]} E_{t,i}$ be the "good event".

② Now we take a look at the per-step regret. W.l.o.g., we assume arm 1 is the optimal arm. Condition on $\mathcal{E}$, it holds that

$$
\begin{aligned}
\mu^* - \mu_{A_t} &\leq UCB_1(t) - \mu_{A_t} \\
&\leq UCB_{A_t}(t) - \mu_{A_t}. \\
&= CI + \hat{\mu}_{t-1,i} - \mu_i. \\
&\leq 2CI. \\
&= 2\sqrt{\frac{2\ln(nk/\delta)}{T_i(t-1)}}.
\end{aligned}
$$

③ Summing over all steps leads to

$$
\begin{aligned}
\sum_{t=1}^{n} \mu^* - \sum_{t=1}^{n} \mu_{A_t} &\lesssim \sum_{t=1}^{n} \sqrt{\frac{\ln(nk/\delta)}{T_{A_t}(t-1)}} \\
&= \sqrt{\ln(nk/\delta)} \cdot \sum_{i=1}^{k} \sum_{t=1}^{T_i(n)} \frac{1}{\sqrt{t}} \\
&\lesssim \sqrt{\ln(nk/\delta)} \cdot \sum_{i=1}^{k} \sqrt{T_i(n)} \\
&\leq \sqrt{\ln(nk/\delta)} \cdot \sqrt{k \sum_{i=1}^{k} T_i(n)} \\
&= \sqrt{kn \cdot \ln(nk/\delta)},
\end{aligned}
$$

w.p. $1-\delta$. ∎

Theorem 7.2 [Gap-dependent Regret Bound]

It holds that the pseudo-regret of UCB algorithm is upper bounded by

$$O\left(\sum_{i:\Delta_i>0} \frac{\ln(nk/\delta)}{\Delta_i} + \sum_{i:\Delta_i>0} \Delta_i\right).$$

w.p. at least $1-\delta$.

Proof. Note that $T_i(n) = \max\{t \in [n] : \mathbb{1}\{A_t = i\} = 1\}$. Let $\tau_i \in [n]$ s.t. $T_i(\tau_i) = T_i(n)$ and $A_{\tau_i} = i$. It is clearly that

$$\sqrt{\frac{2 \cdot \ln(nk/\delta)}{T_i(n)}} \geq \Delta_i, \quad (1)$$

which is due to that $UCB_i(\tau_i) \geq \mu^*$ is the necessary condition that arm i is pulled at time step τ_i. Eq.(1) implies that

$$T_i(n) \lesssim \frac{\ln(nk/\delta)}{\Delta_i^2}.$$

Thus it holds that w.p. $1-\delta$

$$R(n) = \sum_{i:\Delta_i>0} \Delta_i \cdot T_i(n)$$

$$\lesssim \sum_{i:\Delta_i>0} \frac{\ln(nk/\delta)}{\Delta_i},$$

which together with the fact that the algorithm pull each arm once in the first k steps concludes the proof.

$\blacksquare$