Note for Bandit Algorithms

Chenmi'en Tan
E-mail: chenmientan@outlook.com
Actively Updating (Last update: June 19, 2021)

Algorithm 1: Upper Confidence Bound (UCB)

Data: number of arms k and confidence parameter δ
for $t = 1, \dots, n$ **do**
  Choose the action $A_t \leftarrow \arg\max_{i \in \{1,\dots,k\}} \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}$
  Observe the reward and update the upper confidence bound
end

Theorem 1. *Suppose that the bandit $\nu \in \mathcal{E}_{\mathrm{SG}}^k(1)$, then with at least the probability $1-(n+k-1)\delta$, the pseudo-regret $\overline{R}_n = \sum_{t=1}^n \Delta_{A_t}$ for UCB depends on confidence parameter $\delta \in (0,1]$ is bounded by $(4k-4)\sqrt{2n\log\frac{1}{\delta}} + 2\sum_{i=1}^k \Delta_i$, i.e.,*

$$\mathbb{P}(\overline{R}_n \geq (4k-4)\sqrt{2n\log\frac{1}{\delta}} + 2\sum_{i=1}^k \Delta_i) \leq (n+k-1)\delta$$

Proof. Without loss of the gernerality, assume the first arm is optimal. For any $(u_2, \dots, u_k) \in \mathbb{N}_+^{k-1}$, define

$$G = \{\mu_1 < \min_{t \in \{1,\dots,n\}} \mathrm{UCB}_1(t,\delta)\} \cap \bigcap_{i=2}^k \{\hat{\mu}_{iu_i} + \sqrt{\frac{2}{u_i}\log\frac{1}{\delta}} < \mu_1\}$$

When G occurs, there is $T_i(n) \leq u_i, \forall i = 2, \dots, k$. Hence $T_i(n) \geq u_i + 1, \exists i = 2, \dots, k$ implies G_i^c. Assume that $u_i, i = 2, \dots, k$ are large sufficiently to satisfy

$$\Delta_i - \sqrt{\frac{2}{u_i}\log\frac{1}{\delta}} \geq c\Delta_i, \forall i = 2, \dots, k \tag{1}$$

for some $c \in (0,1)$. At this moment we have

$$\begin{aligned}
G^c &= \{\mu_1 \geq \min_{t\in\{1,\dots,n\}} \mathrm{UCB}_1(t,\delta)\} \cup \bigcup_{i=2}^{k} \{\hat{\mu}_{iu_i} + \sqrt{\frac{2}{u_i}\log\frac{1}{\delta}} \geq \mu_1\} \\
&\subset \{\mu_1 \geq \min_{s\in\{1,\dots,n\}} \hat{\mu}_1 + \sqrt{\frac{2}{s}\log\frac{1}{\delta}}\} \cup \bigcup_{i=2}^{k} \{\hat{\mu}_{iu_i} - \mu_i \geq \Delta_i - \sqrt{\frac{2}{u_i}\log\frac{1}{\delta}}\} \\
&\subset \bigcup_{s=1}^{n} \{\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2}{s}\log\frac{1}{\delta}}\} \cup \bigcup_{i=2}^{k} \{\hat{\mu}_{iu_i} - \mu_i \geq c\Delta_i\}
\end{aligned}$$

Hence we can obtain

$$\mathbb{P}(G^c) \leq n\delta + \sum_{i=2}^{k} \exp(-\frac{u_i c^2 \Delta_i^2}{2})$$

which holds under the restriction of (1). Assign $c = 1/2$ and $u_i, i = 2, \dots, k$ to be the minimal feasible value, i.e.,

$$u_i = \lceil \frac{8\log\frac{1}{\delta}}{\Delta_i^2} \rceil$$

we can obtain

$$\begin{aligned}
&\mathbb{P}(\exists i \in \{2,\dots,k\}, T_i \geq u_i + 1) \\
&\leq \mathbb{P}(G^c) \leq n\delta + (k-1)\exp(-\log\frac{1}{\delta}) = (n+k-1)\delta
\end{aligned} \tag{2}$$

Meanwhile, for any real number $\Delta > 0$

$$\begin{aligned}
&\mathbb{P}(\exists i \in \{2,\dots,k\}, T_i \geq u_i + 1) \\
&\geq \mathbb{P}(\overline{R}_n \geq \sum_{i:\Delta_i<\Delta} n\Delta_i + \sum_{i:\Delta_i\geq\Delta} (u_i+1)\Delta_i) \\
&\geq \mathbb{P}(\overline{R}_n \geq (k-1)n\Delta + \sum_{i:\Delta_i\geq\Delta} [(\frac{8\log\frac{1}{\delta}}{\Delta_i^2} + 2)\Delta_i]) \\
&\geq \mathbb{P}(\overline{R}_n \geq (k-1)(n\Delta + \frac{8\log\frac{1}{\delta}}{\Delta}) + 2\sum_{i=1}^{k}\Delta_i)
\end{aligned}$$

By letting $\Delta = \sqrt{8\log(1/\delta)/n}$, we have

$$\begin{aligned}
&\mathbb{P}(\exists i \in \{2,\dots,k\}, T_i \geq u_i + 1) \\
&\geq \mathbb{P}(\overline{R}_n \geq (4k-4)\sqrt{2n\log\frac{1}{\delta}} + 2\sum_{i=1}^{k}\Delta_i)
\end{aligned} \tag{3}$$

By combing equation (2) and (3) we obtain the desired conclusion. □

Theorem 2. *Suppose that RV X satisfies* $\mathrm{supp}(X) \subset [a, b]$ *and X is bounded by B with at least probability $1 - \beta$, i.e.,*

$$\mathbb{P}(X \geq B) \leq \beta$$

then for any $\alpha \in [\beta, 1)$, the conditional value at risk at level α is bounded by $\frac{\beta}{\alpha} b + (1 - \frac{\beta}{\alpha}) B$.

Theorem 3. *Suppose that the bandit $\nu \in \mathcal{E}^k_{\mathrm{SG}}(1)$ and the suboptimality gap $\Delta_i, i = 1, \ldots, k$ is bounded by U, then for any $\alpha \in [(n + k - 1)\delta, 1)$, the UCB depends on confidence parameter $\delta \in (0, 1]$ satisfies that the conditional value at risk for the pseudo-regret $\overline{R}_n = \sum_{t=1}^{n} \Delta_{A_t}$ at level α is bounded by*

$$\frac{(n + k - 1)\delta}{\alpha} nU + (1 - \frac{(n + k - 1)\delta}{\alpha})[(4k - 4)\sqrt{2n \log \frac{1}{\delta}} + 2kU]$$

References